Auditory Learning in Biological and Artificial Systems

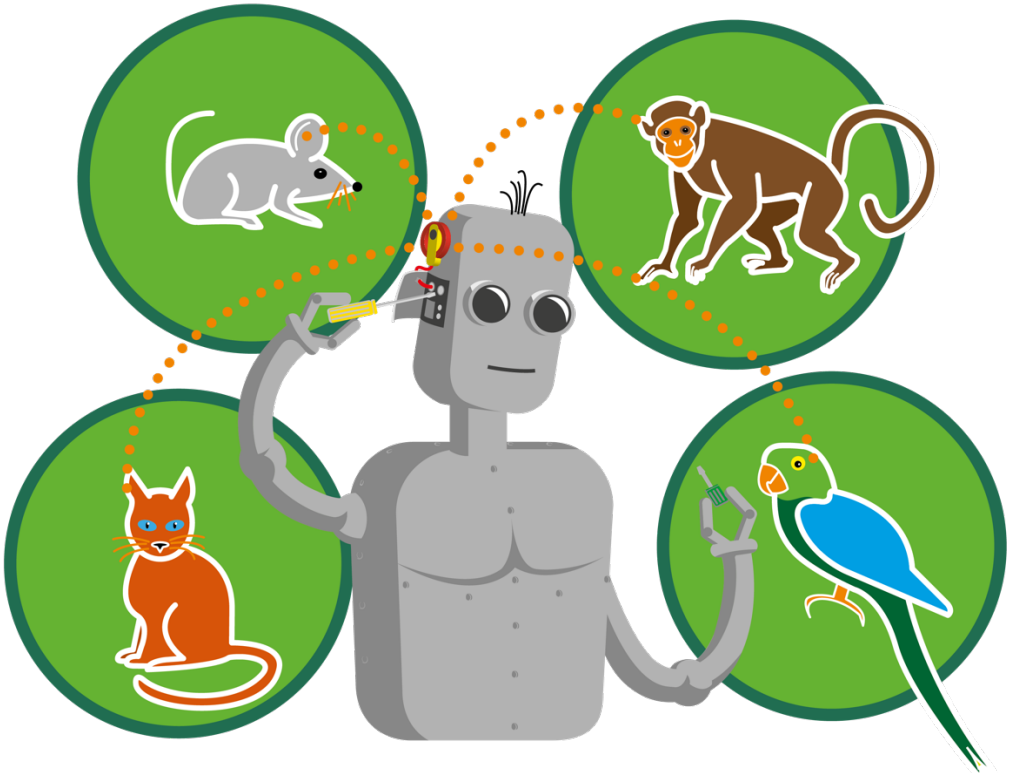


Illustration by Wet Designer Dog (www.wetdesignerdog.dk)

www.isaar.eu

ISAAR is organized by the Danavox Jubilee Foundation
and supported by ReSound GN

Members of the board of the Danavox Jubilee Foundation (2019)

Torsten Dau, Technical University of Denmark

Lisbeth Tranebjærg, University of Copenhagen

Jakob Christensen-Dalsgaard, University of Southern Denmark

Caroline van Oosterhout, Technical University of Denmark

Katrine Bang Termansen

The Danavox Jubilee Foundation
c/o GN ReSound
Lautrupbjerg 7
DK-2750 Ballerup
Denmark

Preface

The 7th International Symposium on Audiological and Auditory Research (ISAAR) was held at Hotel Nyborg Strand in Nyborg, Denmark, from August 21 to 23, 2019. Approximately 170 colleagues from all over the world participated; 32 talks and 80 posters were presented. Many of these contributions can be found as written articles in the present proceedings book.

The focus of this ISAAR was on auditory learning in biological and artificial systems and also included contributions covering a wide range of other topics within auditory and audiological research. Different perspectives were presented and discussed, including current physiological concepts, perceptual measures and models, as well as implications for new technical applications.

The goal of the symposium was to gain insight from current research in different areas and disciplines within hearing science and to relate the findings across these disciplines. The programme was comprised of the following sections: auditory precision medicine; learning from natural sounds; machine listening and intelligent auditory signal processing; and novel directions in hearing-instrument technology. The various presentations reviewed current knowledge in the respective areas and shared new developments, hot topics, and future challenges. In addition to the presentation of the scientific topics, one of the major aims of ISAAR is to promote networking and dialogue between researchers from the various institutions and research centres. ISAAR enables young scientists to approach more experienced researchers and vice-versa and supports links across disciplines. At the symposium, there was a very lively discussion between the researchers spanning a large variety of academic backgrounds.

The organising committee would like to thank GN Hearing for the financial support that made this symposium possible. A special thank goes to Nikolai Bisgaard for his help and support in various matters during the planning and implementation of the symposium. Thank you also to GN Hearing for preparing and providing the symposium material. Last, but not least, the committee thanks all of the authors for their excellent presentations and all of the participants for the lively discussions.

On behalf of the organizing committee,

Torsten Dau

Organizing committee, ISAAR 2019

Scientific

Torsten Dau, Technical University of Denmark

Jakob Christensen-Dalsgaard, University of Southern Denmark

Lisbeth Tranebjærg, Rigshospitalet & University of Copenhagen

Sébastien Santurette, Technical University of Denmark (aff.)

Abigail Anne Kressner, Technical University of Denmark & Rigshospitalet

Administrative

Caroline van Oosterhout, Technical University of Denmark

Katrine Bang Termansen

Abstract, programme, and manuscript coordinator

Abigail Anne Kressner, Technical University of Denmark & Rigshospitalet

Webmaster

Sébastien Santurette, Technical University of Denmark (aff.)

About ISAAR

The "International Symposium on Auditory and Audiological Research" is formerly known as the "Danavox Symposium". The 2019 edition was the 28th symposium in the series and the 7th symposium under the ISAAR name, adopted in 2007. The Danavox Jubilee Foundation was established in 1968 on the occasion of the 25th anniversary of GN Danavox. The aim of the foundation is to support and encourage audiological research and development.

Funds are donated by GN Hearing (formerly GN Danavox and later GN ReSound) and are managed by a board consisting of hearing science specialists who are entirely independent of GN Hearing. Since its establishment in 1968, the resources of the foundation have been used to support a series of symposia, at which a large number of outstanding scientists from all over the world have given lectures, presented posters, and participated in discussions on various audiological topics.

Proceedings from previous symposia are openly accessible in electronic form at the ISAAR proceedings website: http://proceedings.isaar.eu

Contents

The implementation of efficient hearing tests using machine learning

JOSEF SCHLITTENLACHER[1,*] RICHARD E. TURNER[2] AND BRIAN C. J. MOORE[1]

[1] *Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge, CB2 3EB, UK*

[2] *Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK*

Time-efficient hearing tests are important in both clinical practice and research studies. Bayesian active learning (BAL) methods were first proposed in the 1990s. We developed BAL methods for measuring the audiogram, conducting notched-noise tests, determination of the edge frequency of a dead region (f_e), and estimating equal-loudness contours. The methods all use a probabilistic model of the outcome, which can be classification (audible/inaudible), regression (loudness) or model parameters (f_e, outer hair cell loss at f_e). The stimulus parameters for the next trial (e.g. frequency, level) are chosen to yield maximum reduction in the uncertainty of the parameters of the probabilistic model. The approach reduced testing time by a factor of about 5 and, for some tests, yielded results on a continuous frequency scale. For example, auditory filter shapes can be estimated for centre frequencies from 500 to 4000 Hz in 20-30 minutes. The probabilistic modelling allows quantitative comparison of different methods. For audiogram determination, asking subjects to count the number of audible tones in a sequence with decreasing level was slightly more efficient than requiring Yes/No responses. Counting tones yielded higher variance for a single response, but this was offset by the higher information per trial.

INTRODUCTION

Time efficiency is an important attribute of any test. Making a test time efficient is important if it is to be used in clinical practice, and it also reduces costs and allows bigger sample sizes with higher accuracy in research studies.

Most traditional psychophysical methods, like the method of adjustment, magnitude estimation (e.g., Stevens, 1956) or transformed up-down methods (Levitt, 1971), sample at discrete points only. For example, one frequency is tested at a time when measuring an audiogram or the percentage correct is determined at one level at a time when measuring a psychometric function. Von Békésy (1947) circumvented this limitation for the audiogram by slowly sweeping the signal frequency over time and decreasing the level when the subject indicated that the tone was heard and increasing it otherwise. Although this procedure is time efficient and samples at informative points around the threshold, it is problematic because subjects may be slow to respond

*Corresponding author: js2251@cam.ac.uk. Currently at the Department of Neurosciences, University of Cambridge

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

when they stop/start hearing the signal, there may be lapses of attention that affect the measurements even after attention is restored, and the subject may "loose what to listen for", since only near-threshold stimuli are presented.

An ideal procedure would sample at informative points and on continuous scales but also clearly separate stimuli between trials. An early Bayesian procedure, QUEST (Watson and Pelli, 1983), estimated the detection threshold given the data obtained already. The level used in the next trial was the current estimate of threshold. Similar maximum-likelihood methods were developed (e.g., Brand and Kollmeier, 2002).

To our knowledge, the first Bayesian active-learning (BAL) method in psychophysics that used Bayesian principles for both modelling the response and choosing the parameters for the next trial was introduced by Cobo-Lewis (1997). His method was designed to classify a subject into one of nine audiometric groups, e.g., "normal hearing" or "mild to severe sloping loss". The stimulus for the next trial was chosen to maximise the mutual information between the current estimate and that after obtaining one more response. To do this, the posterior probabilities for all candidates that were considered for the next trial were calculated and the one with the least expected entropy (Shannon, 1948) was chosen. Cobo-Lewis validated the method with numerical simulations.

Kontsevich and Tyler (1999) presented a BAL method for estimating the threshold and the slope of a psychometric function, and, like Cobo-Lewis, maximised mutual information when choosing the stimulus for the next trial. They evaluated the procedure with simulations and with real subjects. At that time, computational limits restricted BAL methods to one independent variable only, which in this case was sound pressure level.

Houlsby *et al.* (2011) presented general BAL methods for classification and preference tasks that used Gaussian Processes (GPs; Rasmussen and Williams, 2006) for modelling a subject's response probabilistically. GPs can be multidimensional, i.e., model several independent variables, and incorporate prior beliefs about the mean, the smoothness of the boundaries between classes and the covariance between data points. The latter allows the experimenter to determine how the threshold changes along a given dimension. Houlsby *et al.* (2011) also presented a formula for calculating mutual information without the costly computation of the expected posterior entropy. This was done by exploiting the commutativity of mutual information. The mutual information between the outcome and the model parameters does not require computation of the posterior entropy across the whole space for each candidate data point and outcome ($H(X|Y)$); evaluating the conditional entropy for each data point given the current GP ($H(Y|X)$) is considerably faster.

This approach worked well for determining the similarity between images (Houlsby *et al.,* 2013) and has also been used in auditory applications. For example, GPs have been used to search for the optimal setting of a hearing aid (Nielsen *et al.*, 2014; Jensen *et al.*, 2019) and for determining audiograms (Song *et al.*, 2015; Cox and de Vries, 2015; Schlittenlacher *et al.*, 2018a), equal-loudness contours (Schlittenlacher and Moore, 2019), and psychometric functions (Song *et al.*, 2017). Other BAL

approaches, often using parametric models but also maximising mutual information or something similar, have been used to determine auditory filter shapes (Shen and Richards, 2013; Shen *et al.*, 2014), equal-loudness contours (Shen *et al.*, 2018) and the edge frequency of a dead region (Schlittenlacher *et al.*, 2018b).

The remainder of the paper is organised as follows: First, we briefly present the basics of GPs and BAL hearing tests using the example of determination of an audiogram using Yes/No responses and a "Counting" method (Schlittenlacher *et al.*, 2018a). Second, we present a new BAL test for determining auditory filter shapes and its evaluation using eleven hearing-impaired subjects. In contrast to the procedure of Shen and Richards (2013), our procedure estimates the auditory filter shape not just at a single frequency but over the whole range from 500 to 4000 Hz. The results suggest that this can be done with good accuracy within 20-30 minutes.

PREVIOUS WORK AND MATHEMATICAL BACKGROUND

Binary classification for a Yes/No audiogram

An audiogram is an estimate of the detection threshold of tones as a function of frequency. A GP yields a probabilistic estimate (a Gaussian distribution with a mean and variance) of signal detectability for each point in the two-dimensional frequency-level space:

$$f(x_*, \boldsymbol{x}, \boldsymbol{y}) = GP(m(x_*, \boldsymbol{x}, \boldsymbol{y}), k(x_*, \boldsymbol{x})) \quad \text{(Eq. 1)}$$

with x_* a point in frequency-level space, f the GP function at x_* given already obtained responses $\boldsymbol{y}$ at frequencies and levels $\boldsymbol{x}$, m the mean and k the kernel, which determines the covariance between two data points. We chose a mean based on the data already obtained, a linear covariance in level, which represents the fact that detectability increases with level, and a squared-exponential kernel in frequency with a length scale of 0.5 octaves, which represents the fact that the threshold varies smoothly with frequency.

Equation 1 gives the GP function in latent variable space, which spans (-∞,∞). In order to yield detection probabilities, it was squashed through a likelihood function

$$p_h(x_*, \boldsymbol{x}, \boldsymbol{y}) = 0.01 + 0.98\Phi(f(x_*, \boldsymbol{x}, \boldsymbol{y})) \quad \text{(Eq. 2)}$$

with Φ denoting the Gaussian cumulative density function (CDF) and p_h the probability of x_* (a tone) being reported. Equation 2 produces values between 0.01 and 0.99, accounting for potential lapses in attention that lead to pressing the wrong button independent of x_*. The linear covariance was scaled so that the Gaussian CDF had a standard deviation of 3 dB, thus yielding a common shape for psychometric functions.

Equation 1 requires approximate inference when used for classification. We did this using expectation propagation (EP; Minka, 2001), with Laplace approximation (Williams and Barber, 1998) as a fall back when EP did not converge. Except for the mean, the hyperparameters were not optimized during the BAL process in order to provide stability, especially when early responses were wrong.

The procedure presented here is also applicable to regression tasks such as magnitude estimation and preference tasks such as paired comparisons. For regression, equation 2 is not necessary, and for preference tasks equation 2 needs to be replaced by an appropriate alternative (Chu and Ghahramani, 2005). For further details of GPs, see Rasmussen and Williams (2006). MATLAB code for the Yes/No audiogram is available on github.com/cambridge-mlg/BALaudiogram. The code requires the GPML toolbox (Rasmussen and Nickisch, 2010).

Policy for choosing the next trial

Intuitively one would place the level of the stimulus for the next trial close to threshold. However, the outputs of Equations 1 and 2 also give a variance, allowing us to choose regions where the current model is not confident. There are two major sources for a lack of confidence: no or inadequate sampling of a certain frequency range; and inconsistent responses by the subject.

Ideally, the stimulus for the next trial should minimise the expected entropy in the model after the response for that trial. Houlsby *et al.* (2011) showed that this gain in information can be expressed as the mutual information between the expected response y_* and the model f given the obtained data D ($\boldsymbol{x}$ and $\boldsymbol{y}$) and next data point x_*

$$I(f, y_*|x_*, D) = H(y_*|x_*, D) - \mathbb{E}_{f\sim p(f|D)}[H(y_*|x_*, D)] \qquad \text{(Eq. 3)}$$

In contrast to evaluating the expected entropy of the posterior directly, which requires evaluating one GP for each possible outcome and candidate data point, evaluating the expected entropy of the response (last term in equation 3) only requires a single GP, using the data obtained already. Equation 3 provides an efficient way of looking one step ahead. Less myopic policies that look several steps ahead may further speed up BAL procedures, but this is usually computationally intractable.

Increasing the information per trial

In a binary task like responding "Yes" or "No", the maximum information per trial is 1 bit. It is possible to increase the information per trial by increasing the number of possible responses. Schlittenlacher *et al.* (2018a) presented a variant of the audiogram task where the subject was asked to count the number of pulses heard, with possible counts ranging from 0 to 6. The maximum information per trial in this task is 2.8 bit:

$$H = -\sum_{i=1}^{N} p(x_i) log_2 p(x_i) \qquad \text{(Eq. 4)}$$

where N is the number of different response possibilities and $p(x_i)$ is the probability of the i-th response. This upper limit is reached when all responses have equal probability and no data have been obtained so far. The additional information can be offset by bigger variance in the responses; it is probably more difficult for a subject to count than to select between two alternatives. Nonetheless, the counting procedure converged more quickly towards the ground truth (which was assumed to be the final estimate after 100 or 120 trials) than the Yes/No procedure, with a root-mean-square difference (RMSD) less than 5 dB after only 20 trials.

Another popular task in psychophysics is the two-interval two-alternative forced-choice (2I-2AFC) task. For an audiogram, a tone would be presented in one of two intervals and the subject would have to indicate the interval in which the tone was presented. This procedure reduces the effects of the response criterion of the subject. However, correct responses may result from lucky guesses, which reduces the information gained per trial. The response can be modelled as a binary channel where one crossover probability is 0 (there is no wrong response when a tone is heard) and the other crossover probability is half the probability that a tone is not heard (a correct guess). The information gained per trial without any prior knowledge is

$$I = H_b\left(\frac{1}{2} + \frac{1}{2}p_h\right) - [(1 - p_h)H_b\left(\frac{1}{2}\right) + p_h H_b(1)] \qquad \text{(Eq. 5)}$$

where p_h is the probability that the tone is heard and H_b is the binary entropy. The first term is the entropy of the output and the second term is the entropy of the output given the input and collapses to 1- p_h. I has a maximum (also known as the channel capacity) of 0.32 bit for p_h = 0.6.

The 2I-2AFC task requires about three times as many trials to get the same amount of information as the Yes/No task, which is why it is rarely used in BAL applications. Furthermore, the response criterion effects in a Yes/No task can sometimes be taken into account by model parameters. When estimating auditory filter shapes, for example, the response criterion is incorporated in the "efficiency" parameter K (Patterson, 1976), leaving the shape parameters of interest unaffected.

METHOD FOR ESTIMATING AUDITORY FILTER SHAPES

A BAL method was developed for estimating the thresholds of sinusoidal signals in notched noise as a function of notch width for signal frequencies between 500 and 4000 Hz, on a continuous scale. After the test, auditory filter shapes were estimated from the data. The new method was assessed with hearing-impaired subjects, who were also tested using a conventional method for comparison.

Subjects

Eleven hearing-impaired subjects participated, three female and eight male, aged 55 to 82 years (mean: 70 years). None reported any ear disease or trauma, except for S6 who reported having had a ruptured ear drum. They were paid to participate. They were tested using their better-hearing ear, based on the mean audiometric threshold across 500 to 4000 Hz. Audiograms were obtained using the counting method (Schlittenlacher *et al.*, 2018a) described above. Audiograms are depicted by dashed lines in Figure 2.

Stimuli and apparatus

The experiments took place in a double-walled sound-attenuating chamber. The stimuli were generated digitally with a sampling rate of 48000 Hz and a resolution of 24 bits, converted from digital to analog form by an M-Audio Delta 44 audio interface

(Cumberland, RI), and attenuated by 15 dB with a manual attenuator. They were presented monaurally via a Sennheiser HDA200 headset (Wedemark, Germany).

The task was to detect a pure-tone signal in a notched-noise masker. The signal consisted of three pulses with a duration of 150 ms each and an interval of 100 ms between them. The duration of the noise was 850 ms. It started 100 ms before the first signal pulse and finished 100 ms after the last pulse. The signal pulses and the noise had 20-ms raised-cosine rise/fall times. The signal level (L_s) was 15 dB SL and the signal frequency (f_s) varied from 500 to 4000 Hz or the frequency at which the audiogram reached 40 dB HL for S1 to S6 or 50 dB HL for S7 to S11. The higher signal levels for S7 to S11 were allowed after estimating the loudness of the stimuli for S1 to S6, using the model of Moore and Glasberg (2004). Only 0.5% of the stimuli had a loudness level above 80 phon. For S7 to S11, 0.6% of the stimuli had a loudness level above 80 phon and none had a loudness level above 90 phon. The masker consisted of two noise bands, one centred below the signal frequency and one above, each with a bandwidth of $0.4f_s$. The frequency differences between the signal frequency and the upper edge of the lower noise band or the lower edge of the upper band were chosen to give five symmetric and four asymmetric notch configurations. These frequency differences, expressed as a proportion of f_s, were (0|0), (0.1|0.1), (0.2|0.2), (0.3|0.3), (0.4|0.4), (0.1|0.3), (0.3|0.1), (0.2|0.4) and (0.4|0.2). The level of the noise (L_m) was an independent variable but was bounded so that at most 0.05% of the samples of the entire stimulus were clipped and the level was at most 95 dB SPL. L_m was defined as the sound pressure level in a 1-Hz wide bin, i.e. the spectrum level.

Procedure

After the audiogram was obtained, the subjects did the notched-noise BAL test. Then, they repeated the notched-noise BAL test but using only the (0.2|0.2) notch, to check the consistency of the estimates. After this, notched-noise thresholds were determined using a 2-up/1-down procedure (Levitt, 1971) for the symmetric notches at f_s = 1400 Hz, with the (0.2|0.2) notch in the second and last run. The total test time was about 2 hours including breaks and all tests were conducted in one session.

Notched-noise Bayesian active-learning test

There were three intervals in each trial, separated by 100 ms, containing the signal only, the noise only, and the signal plus noise. This was done to allow the subject to know what to listen for, since the signal varied in frequency from trial to trial. The task was to indicate whether or not the signal was present in the third interval (Yes/No). 10% of the trials did not contain the signal in the third interval to give an estimate of false positives. While sounds were played, a blue rectangle appeared on the screen in the first and second intervals, and a green rectangle in the third interval.

Before the BAL procedure commenced, f_s and L_m were chosen by simple rules for a few trials. The following procedure was repeated for each notch condition: (i) f_s was 1000 Hz and L_m was −20 dB SPL. L_m was increased by 20 dB or decreased by 10 dB, depending on the response, and this was continued (but with the lower limit of L_m set to −30 dB SPL) until a Yes and No response were obtained for f_s = 1000 Hz; (ii) f_s

was set to 2000 Hz and L_m to the mean level used for the two previous trials; (iii) f_s was set to the highest frequency used with that subject and L_m was set either 10 dB below or above the level used for f_s = 2000 Hz, depending on the response for that frequency; (iv) L_m was decreased or increased by 10 dB until both a Yes and No response were obtained; (v) f_s was set to 500 Hz and a procedure similar to that for the highest frequency was used, except that L_m was first set to the same value as used for f_s = 2000 Hz. This typically required 10 trials or less per notch condition.

After the initial grid was completed for each of the nine notch conditions, a GP was calculated for each notch condition. The hyperparameters of the GP, namely the mean, covariance and shape of the CDF, were the same as for the Yes/No audiogram (Schlittenlacher *et al.*, 2018a, see above), namely a linear covariance in L_m, a squared-exponential covariance in f_s with a length-scale of 0.5 octaves and a likelihood function that allowed for lapses.

The parameters for the next trial, namely the notch condition, f_s and L_m, were chosen to yield the highest mutual information about the threshold as a function of notch condition and f_s. This was the same as in Schlittenlacher *et al.* (2018a), except that the maximum was chosen out of nine GPs instead of one (see also Houlsby *et al.*, 2011). The procedure terminated after 594 trials (540 signal trials + 54 catch trials, an average of 60 per notch condition).

2-up/1-down tests

Thresholds were also estimated using a 2I-2AFC 2-up/1-down adaptive procedure (Levitt, 1971) for the symmetric notches, i.e. (0|0), (0.1|0.1), (0.2|0.2), (0.3|0.3) and (0.4|0.4). The (0.2|0.2) notch condition was tested twice, as the second and last runs. The other notch conditions were run in random order. L_s was 15 dB SL and f_s was 1400 Hz. L_m was changed by 5 dB until the second reversal, then by 3 dB until the fourth reversal and by 1 dB for the remainder. The procedure terminated after the 10th reversal. The average of L_m at the last four reversals was taken as the threshold.

RESULTS

For the BAL test, the 50% detection probability of the GP for each notch condition was taken as the threshold for that condition. This provided nine thresholds at each signal frequency, sampled in steps of 0.1 octaves. These were used to estimate auditory filter shapes using a model with three parameters, p_l and p_u, which define the steepness of the lower and upper skirts, respectively, and K, which characterises detection efficiency (Glasberg and Moore, 1990). This simple model does not allow for the flatter "tail" of the auditory filter, so the results for the (0.4|0.4) notch were not used in the analysis. The individual values of p_l and p_u are shown in Figure 1. Lower values indicate less sharp filters.

As expected, the p_l and p_u values (black lines) are generally smaller than expected for normal-hearing subjects (grey lines), especially for the higher signal frequencies, for which the hearing losses were often greater. For S10 and S11, the value of p_u increased markedly for the highest frequency tested, which is unrealistic. This reflects the fact

that the upper slope of the auditory filter is not well defined using the notched-noise method when the lower slope is very shallow (Glasberg and Moore, 1990).

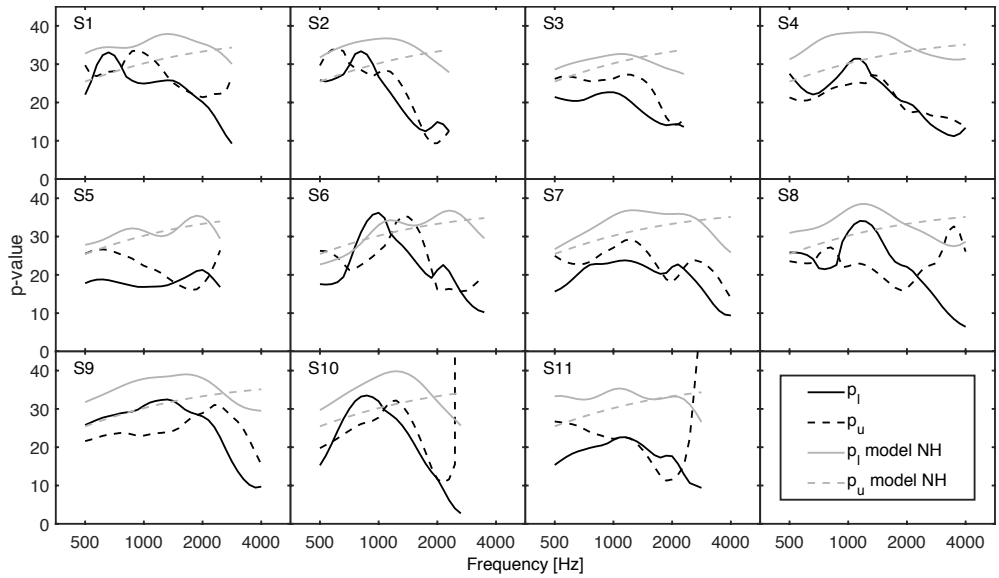


Fig. 1: Black lines show estimated values of p_l (solid lines) and p_u (dashed lines). Grey lines show model predictions for normal-hearing subjects.

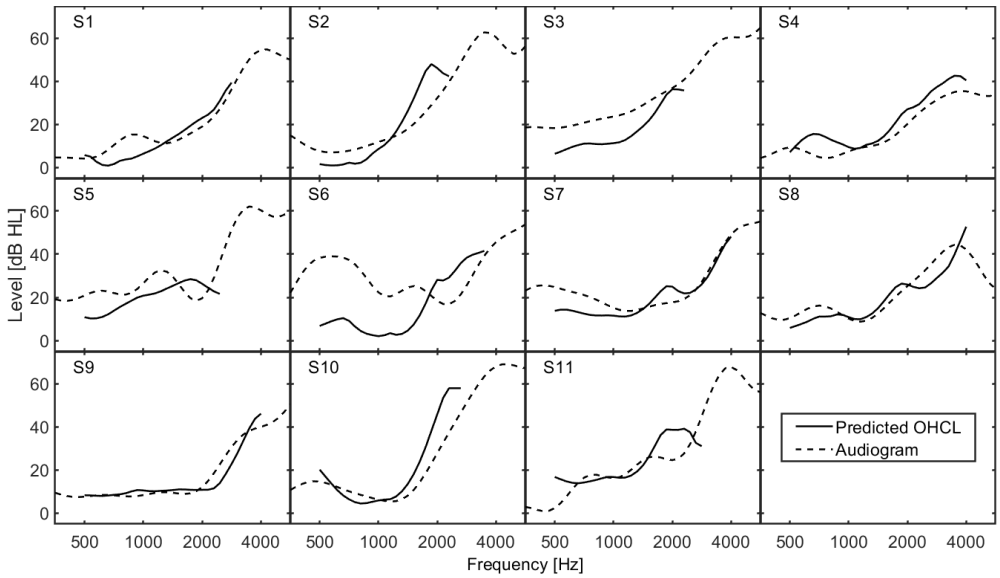


Fig. 2: Solid lines show OHCL values derived from p_l and p_u using the model of Moore and Glasberg (2004). Dashed lines show the audiometric thresholds.

The p_l and p_u values can be related to the amount of hearing loss due to outer hair cell dysfunction (OHCL), using the model of Moore and Glasberg (2004); smaller values of p_l and p_u indicate greater OHCL. Figure 2 shows these relations. For a typical cochlear hearing loss, OHCL is about 90% of the audiometric threshold for hearing

losses up to about 55 dB. Consistent with this, the estimated values of OHCL were usually close to the audiometric thresholds except for S6, who probably had a conductive component to her hearing loss.

The experiment was terminated after an average of 60 trials per notch condition. The estimated auditory filter width was calculated after each trial and divided by the final estimate. The inverse was taken if the ratio was smaller than 1. Figure 3 shows the geometric mean ratio across subjects. The ratio drops below 1.12, representing a small error, after 30 trials per notch condition, which could be obtained in about 20-30 minutes given that the whole test with nine notches took 48-61 minutes.

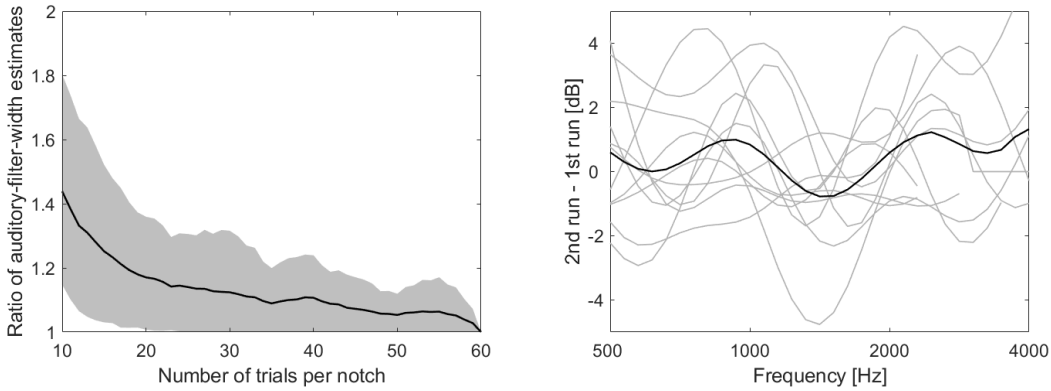


Fig. 3 (left): Ratio between estimated auditory-filter width after *n* trials per notch and the final estimate, plotted as a function of *n*. The inverse was taken if the ratio was smaller than 1. The solid line shows the geometric mean across subjects and the grey area shows the geometric standard deviation.

Fig. 4 (right): Difference between the threshold for the second BAL for the (0.2|0.2) notch only and the threshold for that notch obtained in the main test. The black and grey lines show the mean and individual results, respectively.

The BAL was re-run using the (0.2|0.2) notch width to assess consistency and repeatability. The differences between main test and re-test are shown in Figure 4. The average difference was 0.4 dB and the root mean square difference (RMSD) was 1.8 dB. The slightly higher mean noise level at threshold for the second run may indicate a small learning effect.

Thresholds for the five symmetric notch conditions were estimated at 1.4 kHz using a 2I-2AFC 2-up/1-down procedure. The differences between thresholds obtained with this procedure and with the BAL method are shown in Figure 5. The overall difference was 2.1 dB and the RMSD was 4.0 dB. A small difference would be expected since the 2-up/1-down procedure tracked the 71% correct point in a 2AFC task while the BAL method estimated the 50% point on the psychometric function for the Yes/No procedure. The difference did not vary significantly across notch conditions, as confirmed by a within-subjects analysis of variance, $F(4,40) = 1.25$, $p = 0.31$, $\eta_p^2 = 0.11$. The mean difference between the first and second runs for the (0.2|0.2) notch with the 2-up/1-down procedure was 0.2 dB and the RMSD was 1.2 dB.

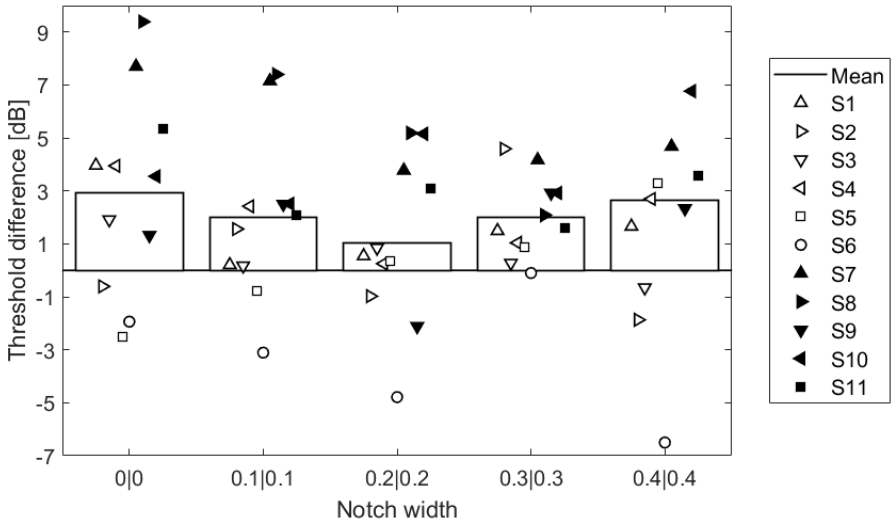


Fig. 5: Difference between the thresholds at 1.4 kHz obtained using the 2-up/1-down procedure and the BAL method for the five symmetric notches. Bars show the mean across subjects and symbols show individual results.

DISCUSSION

The proposed BAL notched-noise method proved to be consistent; thresholds for the (0.2|0.2) notch were similar when estimated in isolation or as part of the main procedure including all notch conditions. Furthermore, differences between the BAL method and the 2-up/1-down procedure were small and similar across notch conditions. Systematic differences across conditions do not affect estimates of the auditory-filter shape, but only affect the "efficiency" parameter, *K*.

The BAL method proved to be fast, yielding reliable estimates of the auditory-filter shape across three octaves in less than 30 minutes. For comparison, it would take approximately the same amount of time to estimate the auditory filter shape at a single frequency using a conventional 2I-2AFC, 2-up/1-down procedure.

Figure 2 shows that, for the subjects with presumed cochlear hearing loss, the derived values of OHCL were close to the audiometric thresholds, as expected. They were sometimes higher than the audiometric threshold, perhaps because the Counting method was used for the audiogram, and this typically gives slightly lower thresholds than the Yes/No method.

Instead of using nine independent two-dimensional GPs, one could use a single three-dimensional GP, exploiting covariance between thresholds for the different notch conditions and possibly making the test even faster. However, more low-dimensional GPs have the advantage of being computationally less expensive, an important aspect given the extensive computation that is required between trials. Furthermore, only one of the nine GPs needed to be updated after each trial.

CONCLUSIONS

BAL methods have the potential to introduce tests into clinical practice that previously took too much time. In addition, they increase the information provided since they are not limited to a grid. The tests described here have been shown to be reliable and valid, making them useful for scientific research, allowing more information to be collected in a given amount of experimental time.

The auditory-filter test described here gives information that may be useful for more personalised initial fitting of a hearing aid. For example, the frequency-dependent gains can be chosen based on the shapes of the auditory filters so as to reduce across-channel masking for speech-like sounds (Fletcher, 1953). Together with other BAL tests for the audiogram, dead regions, or fine-tuning an initial fitting (see introduction), this provides a potential tool for personalised precision medicine.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council (UK, grant number RG78536).

REFERENCES

Békésy, G. von (**1947**). "A new audiometer," Acta Otolaryngol. **35**, 411-422.

Brand, T., and Kollmeier, B. **(2002)**. "Efficient adaptive procedures for threshold and concurrent slope estimation for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am., **111**, 1857-1868.

Chu, W., and Ghahramani, Z. (**2005**). "Preference learning with Gaussian processes," Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 137-144.

Cobo-Lewis, A. B. (**1997**). "An adaptive psychophysical method for subject classification," Percept. Psychophys., **59**, 989-1003.

Cox, M., and de Vries, B. (**2015**). "A Bayesian binary classification approach to pure tone audiometry," arXiv:1511.08670.

Fletcher, H. (**1953**). *Speech and Hearing in Communication* (Van Nostrand, New York), pp. 1-461.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103-138.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (**2011**). "Bayesian active learning for classification and preference learning," arXiv:1112.5745.

Houlsby, N. M., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M., and Lengyel, M. (**2013**). "Cognitive tomography reveals complex, task-independent mental representations," Current Biol., **23**, 2169-2175.

Jensen, N. S., Hau, O., Nielsen, J. B. B., Nielsen, T. B., and Legarth, S. V. (**2019**). "Perceptual effects of adjusting hearing-aid gain by means of a machine-learning approach based on individual user preference," Trends Hear., **23**, 1-23.

Kontsevich, L. L., and Tyler, C. W. (**1999**). "Bayesian adaptive estimation of psychometric slope and threshold," Vision Res., **39**, 2729-2737.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am., **49**, 467-477.

Minka, T. P. (**2001**). "Expectation propagation for approximate Bayesian inference," Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Seattle, Washington, USA, 362-369.

Moore, B. C. J, and Glasberg, B. R. (**2004**). "A revised model of loudness perception applied to cochlear hearing loss," Hear. Res. **188**, 70-88.

Nielsen, J. B. B., Nielsen, J., and Larsen, J. (**2014**). "Perception-based personalization of hearing aids using Gaussian processes and active learning," IEEE/ACM Trans. Audio, Speech, Language Process., **23**, 162-173.

Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli," J. Acoust. Soc. Am., **59**, 640-654.

Rasmussen, C. E., and Williams, C. K. I. (**2006**). *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, USA.

Rasmussen, C. E., and Nickisch, H. (**2010**). "Gaussian processes for machine learning (GPML) toolbox," J. Mach. Learn. Res. **11**, 3011-3015.

Schlittenlacher, J., Turner, R. E., and Moore, B. C. J. (**2018a**). "Audiogram estimation using Bayesian active learning," J. Acoust. Soc. Am. **144**, 421-430.

Schlittenlacher, J., Turner, R. E., and Moore, B. C. J. (**2018b**). "A hearing-model-based active-learning test for the determination of dead regions," Trends Hear. **22**, 1-13.

Schlittenlacher, J., and Moore, B. C. J. (**2019**). "Fast estimation of equal-loudness contours using Bayesian active learning and direct scaling," Acoust. Sci. Tech. (in press).

Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423, 623–656.

Shen, Y., and Richards, V. M. (**2013**). "Bayesian adaptive estimation of the auditory filter," J. Acoust. Soc. Am., **134**, 1134-1145.

Shen, Y., Sivakumar, R., and Richards, V. M. (**2014**). "Rapid estimation of high-parameter auditory-filter shapes," J. Acoust. Soc. Am., **136**, 1857-1868.

Shen, Y., Zhang, C., and Zhang, Z. (**2018**). "Feasibility of interleaved Bayesian adaptive procedures in estimating the equal-loudness contour," J. Acoust. Soc. Am., **144**, 2363-2374.

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., and Barbour, D. L. (**2015**). "Fast, continuous audiogram estimation using machine learning," Ear Hearing, **36**, e326–e335.

Song, X. D., Garnett, R., and Barbour, D. L. (**2017**). "Psychometric function estimation by probabilistic classification," J. Acoust. Soc. Am., **141**, 2513-2525.

Stevens, S. S. (**1956**). "The direct estimation of sensory magnitudes: Loudness," Am. J. Psychol., **69**, 1-25.

Watson, A. B., and Pelli, D. G. (**1983**). "QUEST: A Bayesian adaptive psychometric method," Percept. Psychophys., **33**, 113-120.

Williams, C. K. I. and Barber, D. (**1998**). "Bayesian classification with Gaussian Processes," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1342–1351.

From derived-band envelope-following responses to individualized models of near- and supra-threshold hearing deficits

SARINEH KESHISHZADEH* AND SARAH VERHULST

Hearing Technology @WAVES, Department of Information Technology, Ghent University, Belgium

Auditory models which include frequency-dependent profiles of near and supra-threshold hearing deficits can aid the design of individualized hearing-aid algorithms. However, determining individual auditory-nerve (AN) fiber loss parameters is controversial as diagnostic metrics are presently based on auditory brainstem responses (ABRs) or envelope following responses (EFRs). These measures do not necessarily yield a frequency-specific quantification and might be affected by both outer-hair-cell and AN damage. We developed a derived-band EFR (DBEFR) metric to offer a frequency-specific assessment and complemented these with click-evoked otoacoustic emissions and audiometry. Cochlear-gain-loss profiles were derived from the latter measurements and inserted into individualized models, in which different synaptopathy profiles were introduced and DBEFRs simulated. Using a clustering technique, the best match between experimental and simulated synaptopathy profiles was determined and validated using the ABR data collected from the same listener. Results showed promise in offering a method to determine individualized sensorineural hearing-loss profile given a limited number of objective metrics.

INTRODUCTION

The number of auditory nerve (AN) fibers which synapse onto inner-hair-cells is an important factor for auditory processing of supra-threshold sound. A reduction of synapses as a consequence of noise exposure or ageing, i.e. cochlear synaptopathy (CS), degrades temporal encoding fidelity of supra-threshold sound, while leaving the audiometric thresholds unaffected (Kujawa and Liberman, 2009; Bharadwaj *et al.*, 2014). Evidence from animal studies have shown that the amplitude of auditory evoked potentials (AEPs), such as auditory brainstem responses (ABR) and envelope following responses (EFR) are sensitive markers of histologically verified CS (Kujawa and Liberman, 2009; Furman *et al.*, 2013; Shaheen *et al.*, 2015). CS compromises the number of AN fibers which can fire synchronously to a stimulus and hence reduces the supra-threshold ABR wave-I amplitude (Kujawa and Liberman, 2009) and its growth-slope as a function of increasing stimulus level (Furman *et al.*, 2013; Möhrle *et al.*, 2016). Moreover, noise-induced CS degrades the phase-locking strength to the envelope of a modulated tone (Shaheen *et al.*, 2015; Bharadwaj *et al.*,

*Corresponding author: sarineh.keshishzadeh@ugent.be

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

2014, 2015; Parthasarathy and Kujawa, 2018). However, direct assessment of AN synapses requires invasive procedures, therefore quantification of CS in humans is challenging, as scalp-recorded AEPs represent summed activity of large populations of neurons which can be influenced by multiple sources, such as outer-hair-cell (OHC) loss and CS (Gorga *et al.*, 1985; Verhulst *et al.*, 2016). Furthermore, the spread of basilar membrane (BM) excitation caused by the stimulation paradigm causes off-frequency channels to contribute and complicate frequency-specific AEP-based CS diagnostics (Bharadwaj *et al.*, 2014; Encina-Llamas *et al.*, 2019). To address these issues and make a precise quantification of CS in humans possible, this study adopts a combined experimental and modelling approach. A comparison between simulated and experimental frequency-specific EFRs are used to find the best matching AN-damage profile among different simulated CS profiles, which can explain the experimental observations. Cochlear gain model parameters (OHC damage) were set based on experimental audiometric thresholds and click-evoked otoacoustic emissions (CEOAE), while individual CF-specific CS profiles were the outcome parameter, given the EFR recordings.

METHOD

A. Experimental approach

Participants

Two groups were recruited for the experiment, (i) a normal-hearing (NH: 24.21 $\pm$ 4.10 years, N = 16) group and (ii) a group of listeners with normal audiometric thresholds ($<$ 20 dB HL at 4 kHz and $<$ 25 dB HL at frequencies above 4 Hz), but self-reported hearing difficulties in noisy environments (NHSR: 33.78 $\pm$ 8.57 years, N = 9). The latter participant group was recruited using flyers asking *"Do you experience problems when communicating in noisy environments?"*. The underlying assumption was that the NHSR group could suffer from supra-threshold hearing deficits and yield EFR metrics representative of those recorded in synaptopathy animals. Audiometric thresholds at half-octave frequencies between 250 and 8000 Hz were assessed and the best audiometric ear at 4 kHz was chosen for the experiment. Measurements were performed in an acoustically and electrically shielded booth, while subjects were watching a silent movie with subtitles. Participants were informed about the experiment details and an informed consent was received according to the ethical commission at Ghent University.

Derived-band envelope following responses (DBEFRs)

EFRs were recorded to 120-Hz amplitude-modulated white-noise carriers with a bandwidth of [2-22] or [4-22] kHz and a modulation depth of 100%. Stimuli were presented monaurally with an equal spectral level of 70 dB SPL, yielding a lower loudness percept for the narrower band stimulus. Alternate polarity 1.25 s-length epochs were presented 370 times (185 of each polarity) using the same experimental setup as described in (Keshishzadeh *et al.*, 2019a). 1s-long epochs were extracted

from Cz-channel EFRs, starting from 0.25 s after the trigger onset. Employing the same steps adopted in Keshishzadeh *et al.* (2019b), peak to noise-floor spectral absolute values at the fundamental frequency ($f_0 = 120$Hz) and the two following harmonics were extracted (EFR_{PtN}). Afterwards, the DBEFR magnitude was defined by subtracting the EFR_{PtN} to different bandwidths using:

$$DBEFR_{[2-4]} = \begin{cases} (EFR_{PtN})_{[2-22]} - (EFR_{PtN})_{[4-22]}, & (EFR_{PtN})_{[2-22]} > (EFR_{PtN})_{[4-22]} \\ 0, & \text{else} \end{cases} \quad \text{(Eq. 1)}$$

Auditory brainstem responses (ABRs)

Alternate polarity 80 μs-length clicks were presented with a 11.38 Hz rate using the setup described in Keshishzadeh *et al.* (2019a). ABRs were recorded to 90 and 100 dB peSPL clicks and 3000 repetitions with a 10% jitter around the inter-click interval (ICI). The waveforms from nine central channels, i.e. F1, Fz, F2, FC1, FCz, FC2, C1, Cz and C2 were bandpass-filtered using an 800-th order FIR-filter (zero-phase filtering with *filtfilt* function of MATLAB) in [100-1500]-Hz bandwidth and were epoched between -5 to 20-ms relative to the onset. A baseline drift correction was applied to each epoch by subtracting corresponding means. Afterwards, every positive epoch was averaged with the following negative one and the subtraction between the peak and trough of each averaged pair was calculated as the criterion for rejection of noisy trials. Ninety paired-averages (i.e., 180 trials) with the highest peak-to-trough values were assumed as artifact-contaminated pairs and removed. The peak-to-trough subtracted values of the remaining trials did not exceed 25 μV. This approach was adopted to avoid the possible unequal averaging of epochs of each polarity after artifact rejection and was the same for all listeners. Finally, ABRs were calculated by averaging the remaining paired-averages across the nine channels. Wave-V amplitudes were identified manually from the wave-V peak to the next trough in the waveform.

Click-evoked otoacoustic emissions (CEOAEs)

CEOAEs were recorded in response to 83.33 μs duration clicks and presented at 70 d -peSPL with a rate of 25 Hz with ICI of 39.92 ms and 2000 repetitions. Clicks were generated in MATLAB and sent via a Fireface UCX external sound card (RME) and headphone driver to an ER10X Extended-Bandwidth Etymotic Research Probe System. The ER10X recorded ear-canal pressure and responses were digitized via the external sound card (RME). First, time-domain raw recordings were converted to pressure by using the microphone sensitivity (50 $\frac{mV}{Pa}$) and amplifier gain (40dB), and were then bandpass-filtered between 250 and 6000 Hz using a 32^{nd} order zero-phase FIR-filter. Ten percent of the trials with amplitudes larger than twice the standard deviation of the mean, were rejected. The noise-floor was calculated by subtracting the odd and even trials, assuming that the residual noise was not correlated to the stimulus. Afterwards, a linear windowing method was adopted to extract the OAE

using a tenth-order recursive exponential window (Kalluri and Shera, 2001; Shera and Zweig, 1993).

B. Individualized auditory periphery model

A computational model of the auditory periphery (Verhulst *et al.*, 2018) was employed to simulate the experimental data and derive individualized frequency-specific CS profiles to simulate individual sensorineural hearing-loss (SNHL) profiles. Two main steps were necessary to simulate individualized SNHL profiles: (i) Introduce frequency-specific OHC damage profiles using the CEOAE and audiometric data and (ii) simulate EFRs for a variety of CS profiles. In the first step, we first simulated individual cochlear BM impedance discontinuities (Shera and Guinan Jr, 2007), which were derived from the 70 dB peSPL CEOAE recordings. The randomized discontinuities across CF in the model, i.e. the CF-dependent roughness parameters, were individualized by matching the spectral peaks and troughs in the [250-6000] Hz frequency range of the recorded CEOAEs. Spectral bins without peaks or troughs were set to zero and the generated CF-dependent vector was normalized between -1 and +1. Next, individual audiometric thresholds were used to adjust the cochlear-gain-loss parameters by determining the pole values of the BM admittance function across CF (Verhulst *et al.*, 2016). This operation translates a dB-HL loss into a corresponding cochlear filter with lower gain and wider bandwidth. In the second step, the individualized model for OHC-damage-related parameters was used to simulate EFR/ABRs for a range of loss profiles.

To model the spread of total AN fibers per CF, the CF-dependence of counted synapses for rhesus monkey reported in Valero *et al.* (2017) were adopted and mapped to the human cochlea using the Greenwood function (Greenwood, 1990). This resulted in a non-uniform distribution of AN fibers across the CF channels, where a specific CF encompasses $N_{HSR} = 68\%, N_{MSR} = 16\%$ and $N_{LSR} = 16\%$ of the total population of AN fibers in that channel. The population of AN fibers were reduced in all CF channels to simulate different degrees of CS (Table 1). Lastly, models with different CS profiles were used to simulate EFRs. DBEFRs were extracted from the steady-state part of the response, using the same method as for the experimental data. The simulations were run for five different individualized CS profiles introduced in Table 1. To determine the CS profile which best matched the simulated and recorded $DBEFR_{[2-4]}$, a Fuzzy C-Means (FCM)-based clustering technique was adopted to

AN Type	**Total number of AN fibers per CF for no CS *(N)***	**Simulated CS Profile**				
		N	**A**	**B**	**C**	**D**
HSR	13	100%	100%	54%	31%	8%
MSR	3	100%	–	–	–	–
LSR	3	100%	–	–	–	–

Table 1: Simulated CS profiles

explain the uncertainty with which a recorded DBEFR pair belonged to a certain CS profile (using a Fuzzy membership function). Different from original FCM clustering, where initial clusters centers are unknown and updated in each iteration, the cluster centers were defined as the simulated $\text{DBEFR}_{[2-4]}$ for each CS profile. Therefore, a single iteration was run to calculate the partition matrix. Algorithm 1 presents a pseudocode to cluster the experimental $\text{DBEFR}_{[2-4]}$ using model simulations.

RESULTS AND VALIDATION

Experimental EFRs (Fig. 1a) showed individual variabilities and overall lower group-means for the NHSR group ($\text{EFR}_{[2-22]}$: $t(19) = 3.36$, $p \approx 0.003$ and $\text{EFR}_{[4-22]}$: $t(19) = 2.76$, $p \approx 0.012$). Extracting the $\text{DBEFRs}_{[2-4]}$ (Fig. 1b) reduced the group-mean differences ($\text{DBEFR}_{[2-4]}$: $t(19) =$ -0.90, $p \approx 0.338$). Individual DBEFRs were employed to find the best match between simulated/recorded personalized CS profiles. Table 2 shows ranked clustering results based on the individual Fuzzy membership degrees (u_k). The first column refers to the best-matched CS profile and the last column corresponds to the lowest ranked CS profile. To validate the predicted DBEFR-based CS-profiles, an independent measure was used which was also recorded (the ABR), but not adopted in the model fit procedure. Although evidence from animal studies point to decreased ABR wave-I growth-slopes as a consequence of CS, robust wave-I peak-picking is controversial in humans. Given that the ABR growth-slope is a relative metric, we assumed that any hearing deficit reflecting on ABR wave-I, would travel through the auditory pathway to inferior colliculus (IC) and reflect on ABR wave-V, as well. The ABR wave-V has similar generator sources in the vicinity of the IC as the EFR and can be recorded with higher signal-to-noise ratios. ABR wave-V amplitudes to 90 and 100 dB peSPL clicks were simulated for the same five CS profiles in Table 1, and the corresponding wave-V amplitudes growth-slope was calculated. Note that subjects without experimentally growing ABR slopes (N = 7) were dropped, since validation was impossible for these individuals. Model-predicted slopes for each CS profile were compared to the experimental ABR slopes and best matching CS profiles were determined based

Algorithm 1 FCM-based Clustering

Fix the number of clusters c to 5 and fuzzifier m to 1.2

$x = x_1, x_2, ..., x_n$, experimental $\text{DBEFR}_{[2-4]}$ for subject n

$v_n = v_{n1}, v_{n2}, ..., v_{nc}$, simulated $\text{DBEFR}_{[2-4]}$ as cluster centers for subject n

1: **for** $k = 1$ to n **do**
2: **for** $j = 1$ to c **do**
3: $d[v_{kj}, x_k] = abs(x_k - v_{kj})$
4: Calculate membership matrix $u_{kj} = \frac{1}{\sum_{i=1}^{c} (\frac{d[v_{kj}, x_k]}{d[v_{ki}, x_k]})^{\frac{2}{m-1}}}$
5: **end for**
6: **end for**

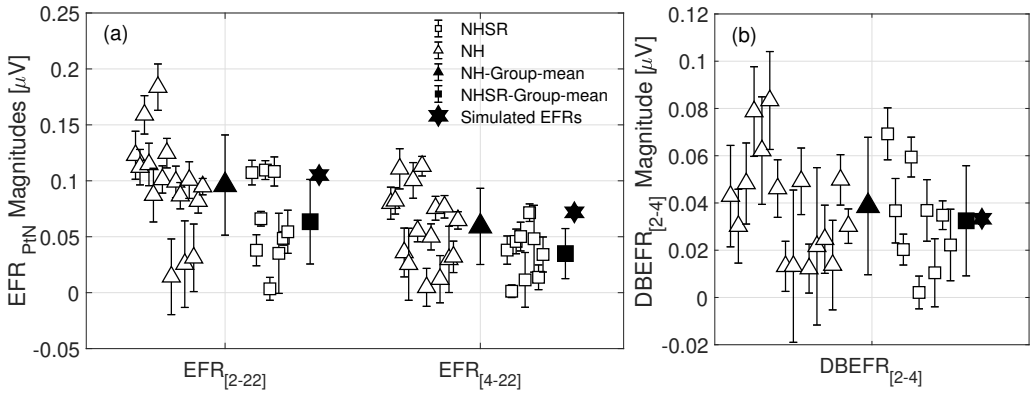


Fig. 1: Experimental and simulated (a) EFRs and (b) $\text{DBEFR}_{[2-4]}$. Error bars indicate the standard deviation of the mean-values.

on the lowest absolute difference between experimental and simulated growth-slopes (indicated with * in Table 2).

DISCUSSION

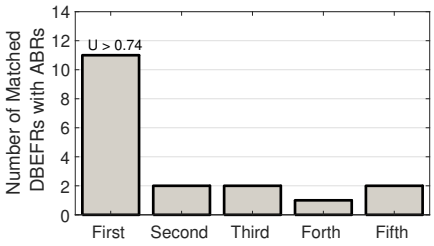
The employed relative derived-band metric was designed to yield a frequency-specific EFR marker of supra-threshold temporal envelope coding which would suppress individual variability (Fig.1a) stemming from subject-specific factors, such as head-size and gender. In a previous study, we used the same DBEFR method (Keshishzadeh *et al.*, 2019a), which revealed statistically significant difference between young normal-hearing (yNH) and old normal-hearing (oNH) DBEFRs, and assigned the degraded DBEFRs in the oNH group to age-induced CS. By extension, one could

NH	**Ranked Predicted CS Profile**					**NHSR**	**Ranked Predicted CS Profile**				
	First	Second	Third	forth	Fifth		First	Second	Third	Forth	Fifth
1	N^*	A	B	C	D	**1**	N^*	A	B	C	D
2	N^*	A	B	C	D	**2**	N	A^*	B	C	D
3	N^*	A	B	C	D	**3**	A	B	C	N^*	D
4	N^*	A	B	C	D	**4**	N^*	A	B	C	D
5	N^*	A	B	C	D	**5**	N^*	A	B	C	D
6	B	C	A	D	N^*	**6**	B	C	D	A	N^*
7	N^*	A	B	C	D	**7**	N^*	A	B	C	D
8	A^*	N	B	C	D	**8**	A	B	N^*	C	D
9	D	C^*	B	A	N						
10	N	A	B^*	C	D						

* *ABR-based predicted CS profiles*

Table 2: DBEFR-based predictions of individual CS profiles validated using experimental ABR growth-slopes derived from the same listeners. *(N: Normal)*

assume that the group mean differences between NH and NHSR group could be explained by noise-induced CS. However differently, we did not observe significant DBEFR group-mean differences and obtained *Normal* CS profiles for most participants of both groups together. Taking into account our previous findings, and the insensitivity of the DBEFR metric to head-size and DPOAE threshold differences (Keshishzadeh *et al.*, 2019a, see Fig.7), we believe that the present results reflect that not all listeners with self-reported hearing difficulties suffer from CS. Additionaly, the unquantifiable separation criterion between the two tested groups, i.e. self-reported hearing difficulties in noisy environments (Coughlin, 1990) and insufficient number of participants can explain the absense of significant differences between the groups. To evaluate the quality of our CS profile predictions, Fig. 2 depicts the performance of our method in correctly predicting either the ABR or DBEFR experimental results using the individual CS profile extracted from the DBEFR clustering method. A coincidence of 61.11% validated profiles with the first-ranked $\text{DBEFR}_{[2-4]}$ based predictions shows promising results. In the future our method can be improved by including other AEP-derived metrics and OHC-deficit related measures in the clustering to yield more robust predictions of individual CS degrees.



Rank	Coincidence (%)
1	61.11
2	11.11
3	11.11
4	5.56
5	11.11

Fig. 2: Evaluation of the method based on the coincidence between first-ranked $\text{DBEFR}_{[2-4]}$ based predictions and ABR wave-V growth-slopes

ACKNOWLEDGEMENT

Work supported by European Research Council grant ERC-StG-678120 (RobSpear).

REFERENCES

Bharadwaj, H. M., Masud, S., Mehraei, G., Verhulst, S., and Shinn-Cunningham, B. G. (**2015**), "Individual differences reveal correlates of hidden hearing deficits," J. Neurosci., **35**(5), 2161–2172.

Bharadwaj, H. M., Verhulst, S., Shaheen, L., Liberman, M. C., and Shinn-Cunningham, B. G. (**2014**), "Cochlear neuropathy and the coding of supra-threshold sound," Front. Syst. Neurosci., **8**, 26.

Encina-Llamas, G., Harte, J. M., Dau, T., Shinn-Cunningham, B., and Epp, B. (**2019**), "Investigating the Effect of Cochlear Synaptopathy on Envelope Following Responses Using a Model of the Auditory Nerve," J. Assoc. Res. Oto., 1–20.

Furman, A. C., Kujawa, S. G., and Liberman, M. C. (**2013**), "Noise-induced cochlear

neuropathy is selective for fibers with low spontaneous rates," J. Neurophysiol., **110**(3), 577–586.

Gorga, M. P., Worthington, D. W., Reiland, J. K., Beauchaine, K. A., and Goldgar, D. E. (**1985**), "Some comparisons between auditory brain stem response thresholds, latencies, and the pure-tone audiogram." Ear Hearing, **6**(2), 105–112.

Greenwood, D. D. (**1990**), "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am., **87**(6), 2592–2605.

Kalluri, R. and Shera, C. A. (**2001**), "Distortion-product source unmixing: A test of the two-mechanism model for DPOAE generation," J. Acoust. Soc. Am., **109**(2), 622–637.

Keshishzadeh, S., Garrett, M., and Verhulst, S. (**2019**a), "The Derived-Band Envelope Following Response and its Sensitivity to Sensorineural Hearing Deficits," bioRxiv, 820704.

Keshishzadeh, S., Vasilkov, V., and Verhulst, S. (**2019**b), "Tonotopic Sensitivity to Supra-Threshold Hearing Deficits of the Envelope Following Response Evoked by Broadband Stimuli," 23rd International Congress on Acoustics, 6548–6553.

Kujawa, S. G. and Liberman, M. C. (**2009**), "Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss," J. Neurosci., **29**(45), 14077–14085.

Möhrle, D., Ni, K., Varakina, K., Bing, D., Lee, S. C., Zimmermann, U., Knipper, M., and Rüttiger, L. (**2016**), "Loss of auditory sensitivity from inner hair cell synaptopathy can be centrally compensated in the young but not old brain," Neurobiol. Aging, **44**, 173–184.

Parthasarathy, A. and Kujawa, S. G. (**2018**), "Synaptopathy in the aging cochlea: Characterizing early-neural deficits in auditory temporal envelope processing," J. Neurosci., **38**(32), 7108–7119.

Shaheen, L. A., Valero, M. D., and Liberman, M. C. (**2015**), "Towards a diagnosis of cochlear neuropathy with envelope following responses," J. Assoc. Res. Oto., **16**(6), 727–745.

Shera, C. A. and Guinan Jr, J. J. (**2007**), "Cochlear traveling-wave amplification, suppression, and beamforming probed using noninvasive calibration of intracochlear distortion sources," J. Acoust. Soc. Am., **121**(2), 1003–1016.

Shera, C. A. and Zweig, G. (**1993**), "Noninvasive measurement of the cochlear traveling-wave ratio," J. Acoust. Soc. Am., **93**(6), 3333–3352.

Valero, M., Burton, J., Hauser, S., Hackett, T., Ramachandran, R., and Liberman, M. (**2017**), "Noise-induced cochlear synaptopathy in rhesus monkeys (Macaca mulatta)," Hearing Res., **353**, 213–223.

Verhulst, S., Altoe, A., and Vasilkov, V. (**2018**), "Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss," Hearing Res., **360**, 55–75.

Verhulst, S., Jagadeesh, A., Mauermann, M., and Ernst, F. (**2016**), "Individual differences in auditory brainstem response wave characteristics: relations to different aspects of peripheral hearing loss," Trends Hear., **20**, 2331216516672186.

Neural health in cochlear implant users with ipsilateral residual hearing

MARINA IMSIECKE[1,*], ANDREAS BÜCHNER[1,2], THOMAS LENARZ[1,2] AND WALDO NOGUEIRA[1,2]

[1] *Clinic for Laryngology, Rhinology and Otology, Hanover Medical School, Germany*

[2] *Cluster of Excellence 'Hearing4All', Hanover, Germany*

Studies in cochlear implant (CI) users have shown a correlation between neural health and speech reception performance. Recently, electrically evoked compound action potentials (eCAP) with varying interphase gaps (IPG) have been used to estimate neural health. In the present study, we investigated eCAP characteristics in CI users with ipsilateral residual hearing (electric-acoustic stimulation, EAS). We hypothesized that neural health is better in apical areas in EAS users than in basal areas, due to increased hair cell survival. Amplitude growth functions (AGF) with varying IPGs of 2.1 and 10 μs were measured in 19 MED-EL Flex recipients with residual hearing. The eCAP characteristics slope, N_1 latency and stimulus level at 50% maximum eCAP amplitude were investigated for the effect of IPG across electrode positions and were correlated to speech perception outcomes and duration of hearing loss. CI users without residual hearing were used as a control group to compare the patterns of slope, latency and 50% maximum amplitude between both IPGs. IPG showed a significant effect on the eCAP characteristics. The change in stimulus level for the 50% maximum amplitude showed a significant difference between electrode 1 and 3 as well as 1 and 4 in EAS users, maybe indicating impaired neural health in the medial region and validating the measurement in EAS users.

INTRODUCTION

Progressive auditory nerve degeneration is known in patients suffering from severe hair cell loss, and survival of the auditory nerve (i.e. neural health), is partially assumed to be responsible for the variability in speech reception performance among cochlear implant (CI) users (Seyyedi *et al.*, 2014; Pfingst *et al.*, 2015). Several studies found a correlation between speech reception performance and indirect measures of neural health, such as duration of hearing loss (Holden *et al.*, 2013; Nadol *et al.*, 2001). However, a strong variability persists, and the state of the auditory nerve cannot be quantified in living humans. Studies in animals that employed objective measures such as the characteristics of the electrically evoked compound action potential (eCAP) have suggested that it can be used to determine the state of the auditory nerve (Prado-Guitierrez *et al.*, 2006). Recently, eCAP amplitude growth functions (AGF) with

*Corresponding author: imsiecke.marina@mh-hannover.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

varying inter-pulse gaps (IPG) have been suggested by Ramekers *et al.* (2014) to analyze the refractory ability of the auditory nerve, which is impaired when the auditory nerve suffers from degeneration. They found that in guinea pigs several characteristics of eCAP and AGF recordings correlated significantly with quantified histological measures of the auditory nerve. For single pulses, the difference in N_1 latency, averaged across the three highest current levels, AGF slope and stimulus intensity needed to reach 50% of the maximum eCAP amplitude, also called current offset, all correlated highly to the spiral ganglion cell (SGC) packing density with varied IPG for normal hearing and deafened animals. The difference in latency decreased with higher density (i.e. better neural survival), and the difference in slope increased with better neural health.

An objective measure of neural health could help to understand and predict speech reception in CI users. The electric-acoustic stimulation (EAS) population, combined with available imaging data of the implant, offers the possibility to validate neural health measures under the assumption that this population has better neural health in the apex than in the base of the cochlea.

METHODS

19 EAS subjects participated in the measurement. A control group of 19 CI users without residual hearing (250 Hz $>$ 90 dB HL) was matched in age and duration of hearing loss to the EAS users. There was a mean difference of 2 years for age, whereas there was a difference of 12 years for duration of hearing loss.

AGFs were measured using the automatized, continuous eCAP measurement function of the MAESTRO (MED-EL, Innsbruck, Austria) fitting software called AutoART for two different IPGs of 2.1 and 10 μs. The amplitude of the single pulse electric stimulation in proprietary charge units (qu) was steadily increased until the subject indicated the loudest acceptable loudness level (LAPL) (Gärtner *et al.*, 2018), upon which the stimulation of the current electrode was stopped. The algorithm of the software then determined the threshold and slope of the AGF by fitting a sigmoid function. Additionally, the latency of the first negative peak was determined. The four most apical electrodes (numbers 1-4) and one basal electrode (9, or 8 if not possible) were measured in EAS users and all electrodes in CI users.

The eCAP characteristics slope, N_1 latency and 50% maximum amplitude were analyzed towards changes per IPG, and these changes were compared across the electrode array. In a second step, the results of the two subject groups with and without residual hearing were compared to each other by identifying differences in the patterns across cochlear location. The eCAP characteristics were analyzed for correlation to indirect measures of neural health, such as duration of hearing loss and speech reception performance. The latter was either obtained by a matrix sentence test in 65 dB noise (Wagener *et al.*, 1999, OLSA) in the EAS users, or with a sentence test (Hochmair-Desoyer *et al.*, 1997, HSM) in quiet at 65 dB presentation level for CI users.

	EAS users				CI users		
ID	**Age**	**CI use**	**Dur HL**	**Electrode**	**Age**	**CI use**	**Dur HL**
c01	43	0.9	9	Flex 20	39	2.3	18
c02	66	2.7	63	Flex 20	67	0.9	66
c03	62	1.5	36	Flex 28	67	3.1	64
c04	39	1.9	26	Flex 28 PI	42	2.9	29
c05	82	3.8	10	Flex 24	82	2.3	4
c06	48	1.4	16	Flex 16	49	1.8	47
c07	49	2.9	35	Flex 20	50	1.5	27
c08	52	1.6	20	Flex 16	54	4.4	14
c09	61	1.7	21	Flex 24	60	3.3	48
c10	68	2.7	12	Flex 16	67	3.3	58
c11	62	2.6	9	Flex 16	63	1.7	12
c12	54	2.5	50	Flex 28	49	2.8	46
c13	46	1.7	11	Flex 24 PI	50	3.2	1
c14	71	1.5	21	Flex 24 PI	71	2.3	NA
c15	46	1.5	10	Flex 28 PI	45	2.7	6
c16	44	2.2	40	Flex 16	42	3.3	29
c17	56	2.0	10	Flex 24	59	2.8	15
c18	78	8.7	9	Flex 20	80	2.8	5
s03	64	1.2	44	Flex 24	61	1.9	58

Table 1: Subject data with subject ID, age at testing, duration of implant use, and duration of hearing loss (dur HL), all in years, for electric-acoustic stimulation (EAS) users and control group of CI users without residual hearing. Electrode type for EAS users is given, for CI users was Flex 28.

RESULTS

The AGFs for two electrodes (i.e. apical electrode contact number 1 and basal electrode number 9, or in one case 8) and for both IPGs are shown in Figure 1 for all subjects with residual hearing. A variability in dynamic range is visible, most AGFs stop between 20 to 30 qu, at which subjects indicated LAPL. Also, a high variability in the slope of the individual AGFs can be observed, with a very pronounced case of no elicited response in the basal electrode of subject ID c13. The eCAP response of some subjects did not exceed noise levels before LAPL was reached. Thus the estimation of the eCAP is missing in these subjects (IDs c03, c16, c17, s03) for different electrodes.

Differences in the AGFs, elicited by the different combinations of electrode position and IPG, become apparent, and these characteristics were further compared and analyzed. Based on the findings of Ramekers *et al.* (2014) and Schvarzt-Leyzac and Pfingst (2018), the characteristic slope (i.e. increase in eCAP amplitude per charge unit), stimulus intensity needed to reach 50% maximum amplitude and the latency of the N_1 component, which is not shown in the AGFs, were chosen to be further

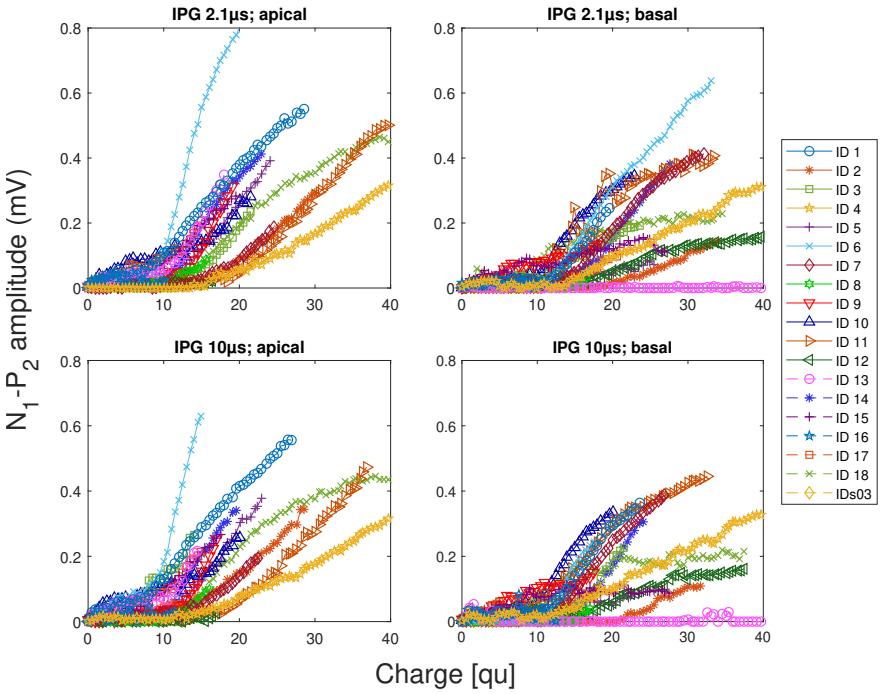


Fig. 1: eCAP amplitude growth functions in dependency of stimulus charge for individual EAS users for apical (left) and basal electrodes (right) and IPGs of 2.1 (top) and 10 μs (bottom).

analyzed.

The changes of these three characteristics due to changes in IPG were obtained in EAS subjects and in a control group of CI recipients without residual hearing. For both groups, all three characteristics were significantly higher with 10 μs than with 2.1 μs IPG (paired t-test $p < 0.01$ for EAS, $p < 0.001$ for CI) across all measured electrodes. The differences in slope, latency and maximum amplitude were also significantly different for the two groups ($p < 0.001$). These differences were also assessed on a basis of the insertion angle of the electrode, which was measured, as comparing the electrode contact number is not feasible for EAS and CI users with very different electrode types and insertion depths. Figure 2 shows the changes in eCAP characteristics across the measured range of insertion angles for both groups EAS (red circles) and CI (blue diamonds) users. Insertion in EAS users was more shallow, so that values only reached up to 400°. The overall trend showed different patterns for the three different characteristics and the two groups across insertion angle. For the change in slope the values decreased towards larger insertion angles (i.e. towards apical locations in EAS users). In contrast, this measure was lower in CI users in basal regions, but increased towards apical locations, resulting in reverse patterns for EAS and CI users. For the change in latency (Fig. 2 middle), the results of EAS and CI users were similar, and almost constant across the insertion angle. For 50% amplitude,

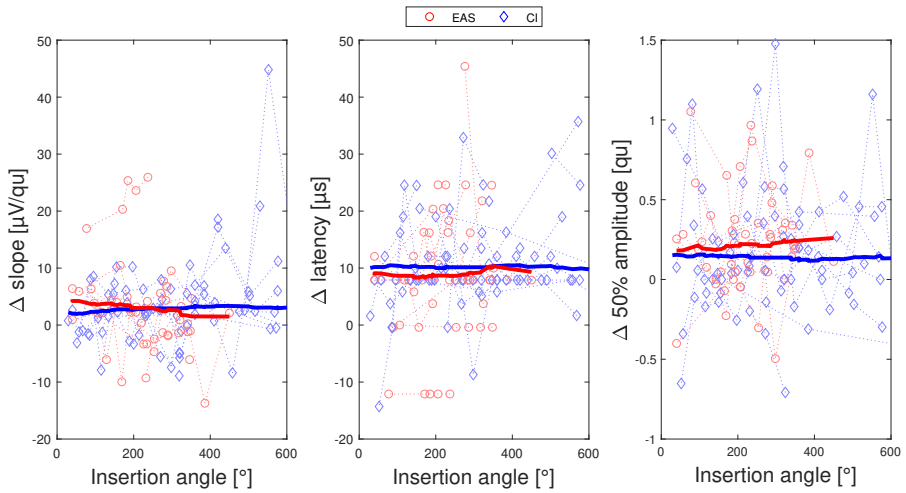


Fig. 2: Differences (Δ) in slope, N_1 latency and 50% maximum amplitude between the two measured IPGs for CI (blue diamonds) and EAS (red circles) users for individual insertion angles and the running average across insertion angle for both groups (lines).

results of CI users were constant across the electrode position, and values in EAS users increased. A statistical analysis between pooled apical and basal regions by median split did not show significant changes for any characteristic or between EAS and CI users, after correcting for the number of comparisons with a Bonferroni correction (significant $p < 0.05/3$). However, in EAS users, a significant difference ($p < 0.05$) between electrodes 1 and 3 as well as 1 and 4 was found for the change in stimulus intensity for 50% maximum amplitude. The data for the basal electrode was reduced due to missing eCAP responses, so that this might prevent a statistically significant difference. No difference was found in CI users, indicating a difference between apical and middle electrodes in EAS users. Additionally, there was a statistically significant difference between the most apical electrode of CI and EAS users.

Duration of hearing loss and speech reception outcomes in each subject were correlated to the change due to IPG of the three characteristics, in Figure 3 this is shown for the results of the most apical electrode, but results are similar for other electrodes or means across electrodes. Duration of hearing loss did not show significant correlation in either CI nor EAS users, and no consistent effect of duration of hearing loss is visible. As the available results were reduced due to missing data in many subjects, the statistical power is reduced.

For speech test outcomes, two different measurements were used, as the performance of the two groups is highly variable. EAS subjects reach ceiling performance in the HSM sentence test in quiet and also in noise at 10 dB SNR, while SRTs are not commonly tested in the clinical routine at the Hannover Medical School (MHH), so that OLSA results are not available in CI users. Each group was individually analyzed

for correlations between IPG effects and speech performance, without significant results.

However, no significant effects could be observed in the current data. A multiple linear regression that took into account the effect of age and duration of hearing loss showed a significant correlation to speech reception performance in EAS group (R^2 = 0.49, p = 0.005), but not in CI users (R^2 = 0.28, p = 0.081). However, none of the eCAP characteristics predicted the speech reception performance or significantly improved the regression model with age and duration of hearing loss ($p > 0.05/3$), indicating that no information about neural health could be gained with these characteristics in the current subject group.

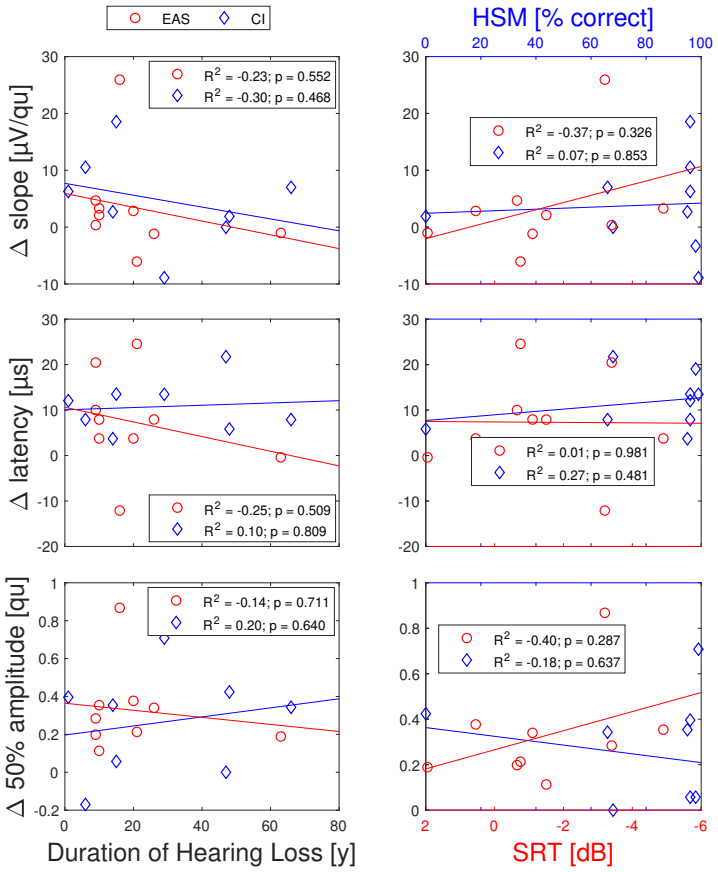


Fig. 3: Differences (Δ) in slope, N_1 latency and 50% maximum amplitude between the two measured IPGs for CI (blue diamonds) and EAS (red circles) users in dependency of duration of hearing loss (left) and speech reception performance (right, speech reception threshold (SRT) of OLSA for EAS and HSM in quiet for CI users).

DISCUSSION AND CONCLUSION

Neural health in EAS and CI users was assessed with eCAP recordings of different IPGs, assuming that the difference in eCAP characteristics caused by the IPG is an indication of the state of neural health (Ramekers et al., 2014). IPG had a significant effect on all characteristics. IPG elicited changes to N_1 latency, slope and maximum amplitud, whiche were compared across electrode location for each individual subject and between the two matched groups. The change in slope and 50% maximum amplitude showed reverse patterns across electrode position for EAS and CI groups, but no significant effects between pooled apical and basal areas. A difference in stimulation intensity for 50% maximum amplitude in apical electrodes of EAS users corresponds to findings of better neural health by Ramekers *et al.* (2014). Duration of deafness did not show a clear influence on eCAP characteristics, opposing the results in animals (Ramekers *et al.*, 2014). The change in slope seems to decrease with increasing duration of hearing loss for both EAS and CI users, but the lack of more data limits the statistical power. Speech reception performance also did not correlate to any of the characteristics. It is possible that large inter-subject variability due to differences in cognitive ability confounds the effect of peripheral neural health on speech performance outcomes, and thus, the results reported by Schvarzt-Leyzac and Pfingst (2018) could not be extended to the inter-subject level. The quite low number of successful AGF fits for some electrodes reduced the strength of the results, and furthermore, the influence of other factors on the indirect measures such as duration of hearing loss and speech perception seems to be higher than the effect of neural health. Thus, the effect of IPG could not be shown in this study.

The hypothesis of better neural health in EAS subjects was confirmed in 50% maximum amplitude in the most apical electrode. IPG effects are limited to a difference between the apical and medial electrodes in EAS users. Possibly larger differences in IPG would have been more sensitive, regarding the high variability across subjects. A study with bilateral EAS/CI subjects is also feasible to investigate the effect of residual hearing. Significant differences in the change in 50% maximum amplitude were found in EAS users between the most apical electrode to middle electrodes, but not within CI users or in comparison to basal electrodes. It seems that 50% maximum amplitude is the most sensitive measure in humans with multicausal hearing loss and should be further investigated in higher numbers of subjects and viable electrodes.

ACKNOWLEDGEMENTS

The authors thank Stefan Strahl (MED-EL) for the helpful input on methods and analysis. This project was funded by the Cluster of Excellence 'Hearing4All' and MED-EL GmbH.

REFERENCES

Gärtner, L., Lenarz, T., and Büchner, A. (**2018**). "Fine-grain recordings of the electrically evoked compound action potential amplitude growth function in cochlear implant recipients," Biomed. Eng., **17**(1), 140, doi: 10.1186/s12938-018-0588-z.

Hochmair-Desoyer, I., Schulz, E., Moser, L., and Schmidt, M. (**1997**). "The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users," Am. J. Otol., **18**(6), 83.

Holden, L., Finley, C., Firszt, J., Holden, T., Brenner, C., Potts, L., Gotter, B., Vanderhoof, S., Mispagel, K., Heydebrand, G. et al. (**2013**). " Factors affecting open-set word recognition in adults with cochlear implants," Ear Hear., **34**(3), 342, doi: 10.1097/AUD.0b013e3182741aa7.

Nadol, J., Burgess, B., Gantz, B., Coker, N., Ketten, D., Kos, I., Roland, J., Shiao, J., Eddington, D., Montandon, P. et al. (**2001**). " Histopathology of cochlear implants in humans," Ann. Otol. Rhinol. Laryng., **110**(9), 883–891.

Pfingst, B.E., Zhou, N., Colesa, D.J., Watts, M.M., Strahl, S.B., Garadat, S.N., Schvartz-Leyzac, K.C., Budenz, C.L., Raphael, Y. and Zwolan, T.A. (**2015**). " Importance of cochlear health for implant function," Hear. Res., **322**, 77-88, doi: 10.1016/j.heares.2014.09.009.

Prado-Guitierrez, P., Fewster, L.M., Heasman, J.M., McKay, C.M., and Shepherd, R.K. (**2006**). "Effect of interphase gap and pulse duration on electrically evoked potentials is correlated with auditory nerve survival," Hear. Res., **215**, 47-55, doi: 10.1016/j.heares.2006.03.006.

Ramekers, D., Versnel, H., Strahl, S.B., Smeets, E.M., Klis, S.F.L., and Grolman, W. (**2014**). "Auditory-nerve responses to varied inter-phase gap and phase duration of the electric pulse stimulus as predictors for neuronal degeneration," J. Assoc. Res. Oto., **15**(2), 187-202, doi: 10.1007/s10162-013-0440-x.

Schvartz-Leyzac, K.C., and Pfingst, B.E. (**2018**). "Assessing the Relationship Between the Electrically Evoked Compound Action Potential and Speech Recognition Abilities in Bilateral Cochlear Implant Recipients," Ear Hearing, **39**(2), 344-358, doi: 10.1097/AUD.0000000000000490.

Seyyedi, M., Viana, L.M. and Nadol, J.B. (**2014**). " Within-subject comparison of word recognition and spiral ganglion cell count in bilateral cochlear implant recipients," Otol. Neurotol., **35**(8), 1446, doi: 10.1097/MAO.0000000000000443.

Wagener, K., Kühnel, V., and Kollmeier, B. (**1999**). "Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test," Zeitschrift Für Audiologie, **38**, 4-15.

Towards unblinding the surgeons: Complex electrical impedance for electrode array insertion guidance in cochlear implantation

NAUMAN HAFEEZ[1,*], XINLI DU[1], NIKOLAOS BOULGOURIS[1], PHILIP BEGG[2], RICHARD IRVING[2], CHRIS COULSAN[2] AND GUILLAUME TOURRELS[3]

[1] *Institute of Environment, Health and Societies, Brunel University, London, UK*

[2] *University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK*

[3] *Oticon Medical Neurelec Inc., Vallauris, France*

The complications during electrode array insertion in scala tympani for cochlear implantation may cause trauma, residual hearing loss and affect speech outcomes. The inner ear is like a black box for surgeons during the insertion process with no real-time feedback and must rely on radiation-based extraoperative imaging. Impedance measurement of electrodes during insertion is a simple yet effective method to assess array position. For this, an impedance meter has been designed which can measure magnitude ($|Z|$), phase (θ), real (R) and imaginary (X_c) parts of impedance. A switching circuit can sequentially scan all electrode pairs at regular intervals during insertion. An Evo® straight electrode array is inserted in a transparent 2:1 scaled up 2D cochlear model (11 trials) filled with 0.9% saline using a 3-degrees-of-freedom actuation system. Bipolar impedance measurements of 8 pairs (40 samples each) are taken at regular intervals during 25 mm insertion at speed of 0.05mm/sec. A notable increase in $|Z|$ and R is observed in the apical 3 electrode pairs when they first get in to contact with the lateral wall. At the same time, the phase gets less negative (more resistive impedance) and X_c increases (less capacitance). These results show that impedance can be used for electrode array localization in cochlea and impedance change due to electrode proximity to different materials can have application in other electrode implants.

INTRODUCTION

A cochlear implant (CI) is an electronic device that provides restoration of auditory perception in patients with sensorineural hearing loss (Eshraghi *et al.*, 2012). The CI mechanism directly stimulates the nerve system by electric means bypassing the damaged sensory hair cells which are responsible for the transconduction of acoustic signals into electrical signals in a normal human ear. An electrode array (EA) inserted into scala tympani (ST) membrane of the inner ear is responsible for stimulating the nerve fibre. There are three objectives of EA insertion for better hearing outcomes:

*Corresponding author: nauman.hafeez@brunel.ac.uk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

(1) deep insertion into cochlear to cover lower frequency range, (2) close proximity to the modiolus wall to ensure greater operating efficiency and (3) to preserve residual hearing by preserving inner ear structure (Rebscher *et al.*, 2008). There is always a high chance of insertion trauma to accomplish the first two goals. This includes injury to the lateral or modiolar wall and distortion of the basilar membrane which may result in loss of residual hearing, poor speech outcomes and limited device performance (Roland and Wright, 2006). Insertion failure such as tip fold over and buckling of electrode array can also have severe effects. With the advent of electro-acoustic cochlear implants and relaxation of eligibility criteria for implantation, it is even more important to save residual hearing.

The reason for these mishaps is the unavailability of real-time intraoperative feedback systems for surgeons during EA insertion. The surgery often only involves preoperative planning and postoperative evaluation using CT scan images. There is also a method to get intraoperative information through fluoroscopy but it involves radiation and its exposure could be harmful to the patient's health. There is, therefore, no safe, inexpensive and intraoperative procedure available to point out trauma or faulty array placement during surgery.

Recent advances in EA design and surgical procedure and tooling were made to keep intra-cochlear trauma to be minimum during implantation (Dhanasingh and Jolly, 2017). EA with softer material, pre-curved perimodiolar arrays, and Advance Off Stylet insertion technique are some examples. Trauma prevention, ease in insertion, better visualisation and human ear anatomy are major factors to choose insertion through the round window or cochleostomy (Richard *et al.*, 2012). These techniques have helped to relatively reduce the risk of trauma, however, failed to completely avoid it and poses other complications, for example, frequent tip folder-over during pre-curved EAs.

A magnetically guided system was presented by Clark *et al.* (2012) in which a magnetically tipped electrode array is guided by a manipulator magnet placed near the patient's head which applies magnetic torque to the tip causing it to bend away from the ST walls. One of the primary focus to avoid trauma is by reducing the insertion forces during insertion. It has been shown in studies that robotic insertion can help control insertion forces by varying insertion speed (Zhang *et al.* (2010); Kontorinis *et al.* (2011); Zhang *et al.* (2009); Rau *et al.* (2010)). These systems may reduce trauma, however, they need local position information of EA in cochlea during insertion.

6The impedance of electrodes during insertion is another feature that could be useful for array localization. Impedance magnitude measurement mechanism is built into all commercially available cochlear implants, however, it is only employed postoperatively to check the correct functioning of each electrode. Tan *et al.* (2013) and Pile *et al.* (2017) conducted experiments to observe the change in impedance magnitude before and after stylet removal of a perimodiolar electrode array

during insertion. According to their results, an increase in impedance magnitude was observed when EA comes closer to the modiolar wall after stylet removal. Giardina *et al.* (2018) looked into the relationship between the monopolar impedance magnitude of electrodes with insertion depth and proximity to the cochlear wall. The change in component values of the equivalent impedance model namely access resistance (R_a), polarization resistance (R_p) and polarization capacitance (C_p) had also been observed during insertion.

Impedance phase has not been looked into in the above studies. The objective of this study is to design an impedance meter that is able to measure bipolar impedance magnitude as well as phase. From these two measures, it is also possible to calculate the real and imaginary parts of the impedance. The proposition is that these properties would change when the electrodes interact with the walls due to disturbances in electrochemical reactions on electrode-electrolyte interface.

METHODOLOGY

Electrode Array and Cochlear Model

Oticon Medical's soft straight EVO® electrode array was used. It is a long (Insertion Length: 25 mm, Active Length: 24 mm), thin (proximal diameter = 0.5 mm; distal diameter = 0.4 mm), flexible array with a smooth silicone surface carrying 20 micro-machined titanium-iridium electrodes as shown in Fig. 1. Each electrode length is 0.47 mm and the gap between two consecutive electrodes is 0.73 mm. Electrodes are numbered from E1 to E20 where former is the basal and latter is the apical electrode. A 2:1 scaled-up plastic 2D cohlear model was used which was filled with the saline solution of 0.9% concentration as an alternative to perilymph fluid.

Actuation System and Impedance Meter

The actuation system used has two translational stages (vertical and horizontal movements) and a rotational stage and is controlled by a custom MATLAB® GUI based application. The insertion speed of the linear actuators and the insertion angle of the rotational stage can be controlled. Physik Instrumente (PI) M-404 and M-061 devices were used for translational and rotational stages respectively and PI's C-863 servo controller was used as a driver of these stages.

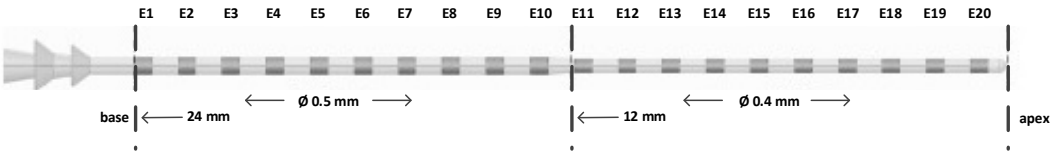


Fig. 1: Evo® Electrode Array with 20 electrodes E1-E20 (www.oticonmedical.com).

The impedance meter was designed using National Instrument's (NI) Data Acquisition

(DAQ) 6211 device having 8 differential I/O channels (or 16 single-ended), 16 bits of ADC resolution and single-channel maximum sampling rate of 250 kS/s. A simple voltage divider circuit was constructed with a known resistance of 3.3 Ω as one of its components and other being the electrode pair. DAQ device is controlled by custom software written in MATLAB®. A sine wave V_{in} of 1V amplitude was applied across the circuit using one of the channels of DAQ device and output voltages V_{out1} and V_{out2} were recorded across electrode pair and known resistance respectively through separate output channels of DAQ device. Since there is a known resistance in the circuit, current I through it can be measured as $I = V_{out2}/R$. The same current flows through the series circuit and impedance magnitude of the electrode pair can be measured using the same current I as $|Z| = V_{out1}I$. Impedance phase θ was measured by taking phase difference between the voltage across electrode pair V_{out1} and current I through them as $\angle\theta = \angle\theta_V - \angle\theta_I$.

Complex impedance is represented by its real and imaginary components in the Cartesian form is given as $Z = R + X_c$. The real and imaginary parts of impedance can be calculated using $|Z|$ and θ as $R = |Z|cos\theta$ and $X_c = |Z|sin\theta$, respectively.

The bipolar impedance of 8 electrode pairs (E20-E19, E18-E17 and so on) was recorded with this impedance meter at room temperature. For switching between 8 electrode pairs, two 74HC4051 8-to-1 multiplexers were used. Multiplexers were controlled by digital output signals from another DAQ device (6009). Six select lines of two multiplexers were connected to the digital output ports of 6009 DAQ device as shown in Fig. 2 as a block diagram.

The overall setup of the complete system is shown in Fig. 3. The plastic cochlear model glued on a support was placed inside a glass filled with saline. the electrode

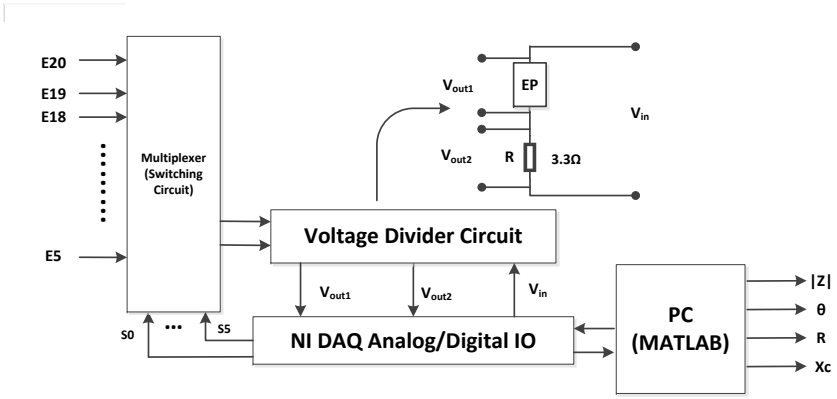


Fig. 2: Impedance Meter and Switching Circuit Setup: 16 Electrodes E20-E5 are connected to multiplexers, 6 select signals S0-S5 are controlling selection of electrode pair for measurement, NI DAQ device is generating and measuring voltages to calculate impedance magnitude, phase, real and imaginary parts.

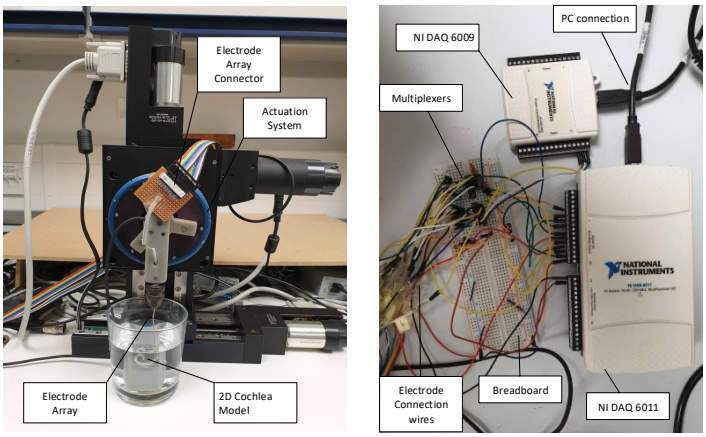


Fig. 3: Experimental Setup (a) Actuation System (b) Impedance meter with the switching circuit connected to electrode wires.

array was placed on a holder attached to the actuation system. Experiments were performed by inserting the electrode array at a speed of 0.05 mm/s for 25 mm depth and bipolar impedance magnitude and phase measurements of each pair were recorded sequentially during the insertion. It took 500 seconds to complete the insertion. The impedance meter samples a pair every 1.5 s, so 41 samples of each electrode pair were taken during the insertion process and analysed offline using Python 3.6.

RESULTS AND DISCUSSION

Bipolar complex impedance of apical 16 electrodes in pairs of 8 were recorded sequentially during the insertion of the electrode array in a plastic ST model. Figure 4 shows four instances during insertion; A) 4 electrodes inserted, B) 8 electrodes inserted, C)16 electrodes inserted, D) 20 electrodes inserted. The arrows show the location of the tip of EA. Figure 5 shows impedance magnitude ($|Z|$), phase (θ), real (R) and imaginary (X_c) parts during insertion. The recordings not only depicts the changes in values when EA is closely placed to the wall (compared to when it is not) but also shows changes when a specific electrode is rubbing (exerted pressure/force) along the wall. It is important to note that when an electrode pair is not in the saline solution (not entered ST model), it gave a high open-circuit impedance value and these values are ignored and not included in the graphs.

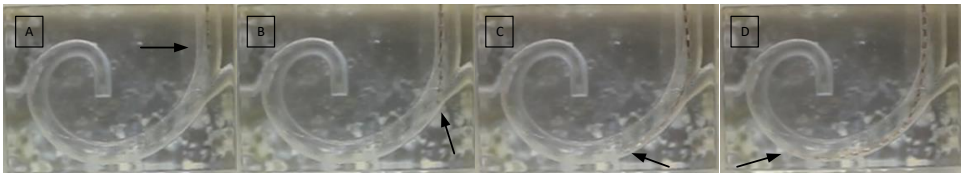


Fig. 4: Electrode array insertion in 2D scala tympani plastic model.

Comparing Fig.4 and Fig.5, EP1 (E19-E20) and EP2 (E17-E18) were inside the model

and away from the wall at instance A. At instance B, E20 (most apical electrode) was starting to touch the lateral wall for the first time and all measurements of EP1 (E19-E20) were starting to change. $|Z|$ from 2.5 kΩ to 3.2 kΩ, θ is getting less negative from -36° to -33°, impedance real part (resistance R) increased from 2 kΩ to 2.7 kΩ, and impedance imaginary part (reactance X_c) from 1.5 kΩ to 1.8 kΩ. In the same way, after 4 samples (between instance B and C) second pair EP2 (E17-E18) came in to contact with the wall and its electrical properties starting to change in the same way. A significant change in measured values were also observed in subsequent electrode pairs EP3(E15-E16), EP4(E13-E14), EP5(E11-E12) just before instance C until instance D, however, there is no significant change in complex impedance of EP6(E9-E10), EP7(E7-E8) and EP8 (E5-E6) as they do not come in to contact with the wall during the insertion process. Figure 6 shows percent change of values during whole insertion (when a particular pair is in saline filled cochlear model) in $|Z|$, θ, R, and X_c of EP1 (apical), EP4 (middle) and EP7 (basal) which clearly shows EP1 has more pronounced change than EP4 due to more contact pressure on it whereas EP7 has no significant change as this pair did not come in to contact with the outer wall.

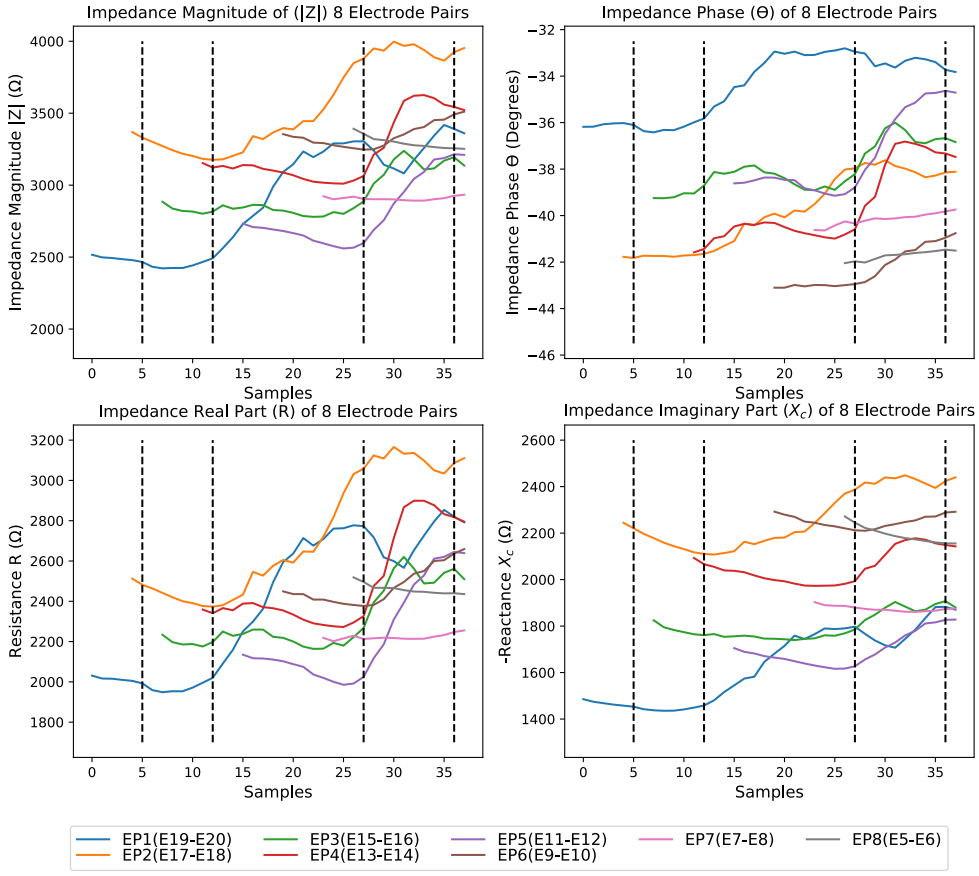


Fig. 5: Impedance magnitude $|Z|$, phase θ, resistance R and capacitive reactance X_c of 8 electrode pairs during insertion.

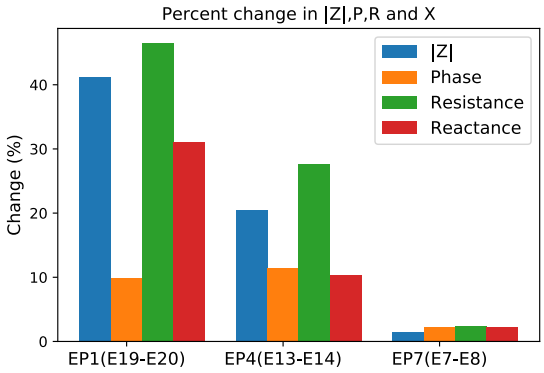


Fig. 6: Percent change (from minimum to maximum values) in impedance magnitude $|Z|$, phase θ, resistance R and capacitive reactance X_c of electrode pair 1, 4 and 7.

We are considering an impedance model of an electrode pair with polarization impedance (resistance and capacitance in parallel) of each electrode contact and an access resistance R_a in series with these polarization impedances $(R1||C1)R_a(R2||C2)$. An increase in $|Z|$ due to wall contact suggested that there is more resistance path between the electrodes. The more negative phase θ implies that impedance is more resistive than capacitive and this phenomenon can also be seen in R and X_c graphs where the change in R is more pronounced than in reactance. Also, an increase in reactance means a decrease in polarization capacitance according to relation $X_c = 1/2\pi fC$.

These results are mainly due to three reasons: 1) disturbance in the chemical reaction at the electrode-electrolyte interface due to wall contact pressure/force, 2) when the array gets closer to the wall, there would be a decrease in reacting electrolyte with electrodes and between electrodes, and 3) impedance of the plastic material is higher than saline.

CONCLUSION

According to current research, there is enough evidence of a relation between EA placement procedure with hearing outcomes and bipolar complex impedance measurements may be used as a sensing technology for localization of EA during cochlear implant surgery. Impedance change due to electrode proximity to different material and application of pressure/force can have application in other electrode implants. These results may be used as feedback control for the actuation system for EA insertion in implantation to manoeuvre it precisely.

REFERENCES

Clark, J.R., Leon, L., Warren, F.M., and Abbott, J.J. (**2012**). "Magnetic guidance of cochlear implants: Proof-of-concept and initial feasibility study," J. Med. Devices, **6**(3). doi: 10.1115/1.4007099.

Dhanasingh, A. and Jolly, C. (**2017**). "An overview of cochlear implant electrode array designs," Hear. Res., **356**, 93–103. doi:10.1016/j.heares.2017.10.005.

Eshraghi, A.A., Nazarian, R., Telischi, F.F., Rajguru, S.M., Truy, E., and Gupta, C. (**2012**). "The cochlear implant: Historical aspects and future prospects," Anat. Rec., **295**, 1967–1980. doi: 10.1002/ar.22580.

Giardina, C.K., Krause, E.S., Koka, K., and Fitzpatrick, D.C. (**2018**). "Impedance measures during in vitro cochlear implantation predict array positioning," IEEE. Trans. Biomed. Eng., **65**, 327–335.

Kontorinis, G., Lenarz, T., Stöver, T., and Paasche, G. (**2011**). "Impact of the insertion speed of cochlear implant electrodes on the insertion forces," Otol. Neurotol., **32**, 565–570. doi: 10.1097/mao.0b013e318219f6ac.

Pile, J., Sweeney, A.D., Kumar, S., Simaan, N., and Wanna, G.B. (**2017**). "Detection of modiolar proximity through bipolar impedance measurements," Laryngoscope, **127**, 1413–1419. doi: 10.1002/lary.26183.

Rau, T.S., Hussong, A., Leinung, M., Lenarz, T., and Majdani, O. (**2010**). "Automated insertion of preformed cochlear implant electrodes: Evaluation of curling behaviour and insertion forces on an artificial cochlear model," Int. J. Comput. Assist. Radiol. Surg., **5**, 173–181.

Rebscher, S.J., Hetherington, A., Bonham, B., Wardrop, P., Whinney, D., and Leake, P.A. (**2008**). "Considerations for the design of future cochlear implant electrode arrays: Electrode array stiffness, size and depth of insertion," J. Rehabil. Res. Dev., **45**, 731.

Richard, C., Fayad, J.N., Doherty, J., and Linthicum Jr, F.H. (**2012**). "Round window versus cochleostomy technique in cochlear implantation: Histological findings," Otol. Neurotol., **33**, 1181. doi: 10.1097/mao.0b013e318263d56d.

Roland, P.S. and Wright, C.G. (**2006**). "Surgical aspects of cochlear implantation: Mechanisms of insertional trauma," In *Cochlear and Brainstem Implants* (Karger Publishers), **64**, 11–30. doi: 10.1159/000094642.

Tan, C.T., Svirsky, M., Anwar, A., Kumar, S., Caessens, B., Carter, P., Treaba, C., and Roland Jr, J.T. (**2013**). "Real-time measurement of electrode impedance during intracochlear electrode insertion," Laryngoscope, **123**, 1028–1032.

Zhang, J., Bhattacharyya, S., and Simaan, N. (**2009**). "Model and parameter identification of friction during robotic insertion of cochlear-implant electrode arrays," Robot. Autom., **2009**, 3859–3864. doi: 10.1109/robot.2009.5152738.

Zhang, J., Wei, W., Ding, J., Roland Jr, J.T., Manolidis, S., and Simaan, N. (**2010**). "Inroads toward robot-assisted cochlear implant surgery using steerable electrode arrays," Otol. Neurotol., **31**, 1199–1206.

Timing of turn taking between normal-hearing and hearing-impaired interlocutors

A. JOSEFINE MUNCH SØRENSEN[1,*], EWEN N MACDONALD[1], THOMAS LUNNER[1,2]

[1] *Department of Health Technology, Technical University of Denmark (DTU), DK-2800 Kgs. Lyngby, Denmark*

[2] *Eriksholm Research Centre, Oticon A/S, DK-3070 Snekkersten, Denmark*

Having a conversation requires more resources than just understanding speech. Previous studies of the timing of turn taking in conversations suggest that in order to sustain normal, fluid turn taking, interlocutors have to predict the end of each other's turns. Thus, while noise and hearing loss should make understanding speech more difficult, it should also reduce the resources available for speech planning and possibly reduce the saliency of cues used to predict turn ends, resulting in delayed and more variable turn taking. We recorded conversations between 12 pairs of native-Danish young normal-hearing (NH) and older hearing-impaired (HI) listeners with mild presbyacusis in quiet and multitalker babble at three levels. The interlocutors conducted a Diapix task, finding differences in two near-identical pictures. Both HI and NH talkers responded more slowly and with more variability with increasing noise level, and the HI with more variability than the NH. We saw indications that the younger NH adopted a more careful communication strategy, likely to ease the effort on their older HI interlocutor, by adapting their speech rates to their interlocutor and overlapping less.

INTRODUCTION

Traditionally, speech understanding and production is studied in isolation where people are passively listening and reporting back what they heard, or producing speech with no addressee. However, real communication is not just the sum of production and listening, it is an interaction between two or more participants who use dynamic feedback and adaptation to increase understanding and information sharing. Recent studies, however, seek to measure speech understanding and production simultaneously in studies of conversational interaction (e.g., Beechey *et al.*, 2018; Hadley *et al.*, 2019). In this study, we investigated conversational turn-taking between younger normal-hearing (NH) and older hearing-impaired (HI) interlocutors solving the Diapix task (Baker and Hazan, 2011) in quiet and in three levels of a multitalker babble noise: 60, 65, and 70 dBA SPL. Earlier studies of conversational interactions suggest that interlocutors predict the end of their partner's turn to sustain normal, rapid turn-taking (e.g., Levinson and Torreira, 2015). We hypothesised that hearing loss and noise interference should increase listening difficulty, reducing the resources available

*Corresponding author: ajso@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

for speech planning and reducing the saliency of predicting cues, resulting in delayed and more variable response times.

METHOD

Participants

Twelve unacquainted mixed- and same-gender pairs of younger normal-hearing (NH) and older hearing-impaired (HI) interlocutors were recruited (9 females, 7 mixed-gender pairs). The NH participants ($\mu = 26$ years, $\sigma = 2.7$ years) had hearing threshold levels below 20 dB HL between 125 Hz and 8 kHz. The HI participants ($\mu = 73$ years, $\sigma = 4.4$ years) had mild presbyacusis (see Figure 1 for their audiograms), and were unaided during the experiment. All participants provided informed consent and the experiment was approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). The participants were compensated for their time.

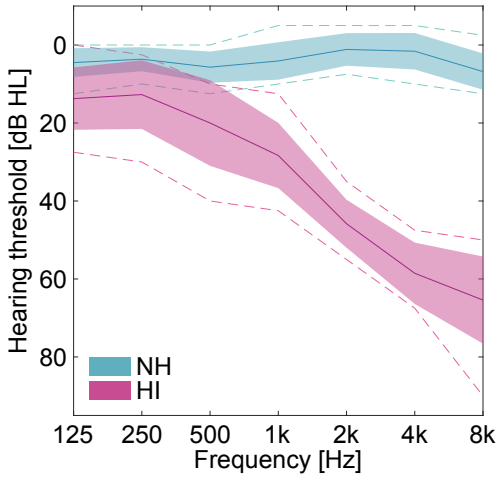


Fig. 1: Audiometric thresholds of the younger normal-hearing and older hearing-impaired listeners. The solid line indicates the mean hearing threshold, the coloured regions indicate one standard deviation, and the dotted lines indicate minimum and maximum measured thresholds.

Setup

Seated in separate booths, the participants wore Shure PGA31 wireless cardioid microphones (transmitted by Shure GLXD14 wireless system) and Sennheiser HD650 open headphones, over which they communicated with each other. The gains were calibrated such that the resulting presentation levels over the headphones were the same as the A-weighted broadband levels one meter away from the talker in the same room. A 20-talker babble was created by taking 20 minutes of recordings from 20 talkers balanced in genders from the NH/NH recordings from Sørensen *et al.* (2020).

The recordings were normalized to the same RMS level as the recording with lowest RMS level, and pauses were removed using voice activity detection (VAD). Finally, they were added together. Auditive verification ensured it was impossible to resolve any content from the individual talkers.

Task and procedure

Similarly to the task in Sørensen *et al.* (2020), participants solved the DiapixUK task (Baker and Hazan, 2011) to elicit dialogue. A training round was conducted outside the booths to familiarize the participants with the task. Inside the booths, the participants had another test round with 65 dBA SPL background noise to familiarize them with the setup and procedure. In the test, the participants solved the Diapix task in three replicates of four conditions: quiet, 60, 65 and 70 dBA SPL background noise. The order of the conditions was randomized within each replicate, and they had a break in between each replicate. The participants were given a maximum of 10 minutes to find 10 differences between the Diapix.

Analysis of recordings

During a turn-taking there is a change in the conversational floor termed a floor-transfer. The duration of such a floor-transfer is termed a floor-transfer offset (FTO) measured from the offset of one person's speech to the onset of the next person's speech. This can either be negative, termed an overlap-between, or positive, termed a gap. Following the procedure in Sørensen *et al.* (2020), each of the conversations were categorized into conversational states: 1) gaps, 2) overlaps-between, 3) utterances, which are speech tokens separated by silence of less than 180 ms, 4) pauses, which are joint silence not followed by a floor-transfer, and 5) overlaps-within, which are joint speech during utterances of one talker that does not result in a floor-transfer.

Mixed-effects regression models were fit to the variables in *R* using the *lme4* package, with background, hearing and replicate as main effects, and pair as random intercept. Denominator degrees-of-freedom were Satterthwaite approximated for the F-tests for the fixed effects. Pairwise comparisons were computed using the *lsmeans* function (*lmerTest* package) comparing least-squares means of the significant effects using the Satterthwaite approximated df.

RESULTS

The average speaking levels of the participants is seen in Figure 2, left panel. A random intercept for participants was added to the mixed effects model. All participants increased their speaking levels significantly in background noise $[F(3,258) = 584.95, p < 2.2\text{e-}16]$, and there was a significant interaction between hearing status and background $[F(3,269) = 14.54, p < 8.91\text{e-}9]$. A multiple comparison post-hoc analysis revealed that the difference was driven by the level differences in quiet between NH and HI, where the HI spoke significantly louder than the NH $[t(24.7) = -2.091, p < 0.047]$.

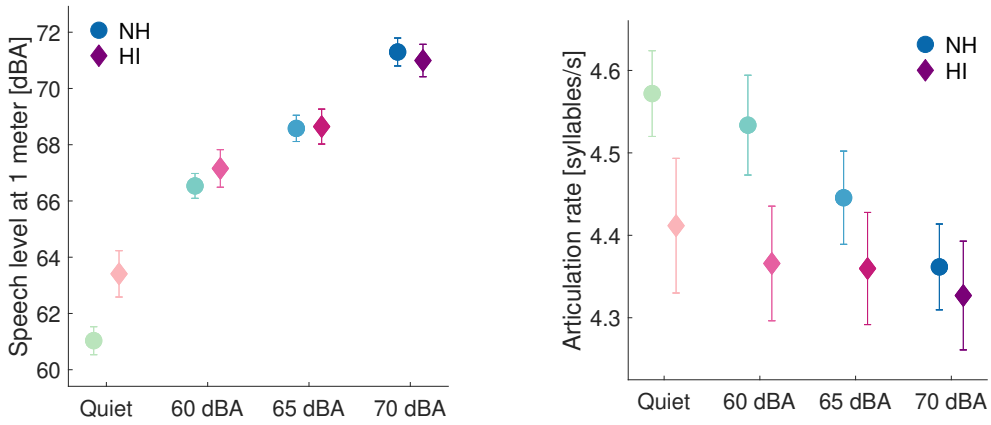


Fig. 2: Speech level (left panel) and articulation rate (right panel) in the four backgrounds (quiet, 60, 65 and 70 dBA SPL) for younger normal-hearing and older hearing-impaired listeners. The bars indicate standard error.

In Figure 2, right panel, the participant's articulation rates for the NH and HI in the four backgrounds is plotted (computed using the Praat script presented in de Jong and Wempe (2009) with default parameter settings). Again, a random intercept for participants was added. There was a significant main effect of background $[F(3,258) = 10.97, p < 8.88\text{e-}7]$, and a significant interaction between hearing and background $[F(3,258) = 2.75, p < 0.043]$. With increasing noise level, the articulation rates of both groups of talkers decreased, and the articulation rates of the NH talkers approached those of the HI.

As an indication of who tended to dominate the conversation, the average proportion of time each person in the two hearing status groups was speaking was computed and can be seen in the left panel of Figure 3. The proportion is measured as the total duration of active speech from the participant (determined by VAD) divided by the total duration of active speech in the conversation from both participants. There was a statistically significant effect of hearing, with the HI speaking more than the NH: $[F(1,286) = 80.4, p < 2.2\text{e-}16]$.

The HI group produced more overlaps-within than the NH and for both groups the rate of overlaps-within decreased with increasing background noise level, confirmed by a statistically significant main effect of both hearing $[F(1,272) = 5.44, p < 0.0204]$ and background $[F(3,272) = 7.47, p < 8.05\text{e-}5]$.

The FTO distributions, collapsed across participants within the NH and HI groups, in the four backgrounds, are seen in the left panel of Figure 4. The distributions were estimated using 100 ms bin widths. By visual inspection, the distributions seem broader for the HI group than the NH group, and slightly broader with increasing noise level. A two-sample Kolmogorov-Smirnov test rejected the null-hypothesis that the samples came from the same distributions for the NH and HI in the four conditions:

$[D = 0.103, p < 2.2\text{e-}16]$ in quiet, $[D = 0.103, p < 1\text{e-}13]$ in 60 dBA noise, $[D = 0.088, p < 1.18\text{e-}10]$ in 65 dBA noise and $[D = 0.096, p < 1.4\text{e-}12]$ in 70 dBA noise.

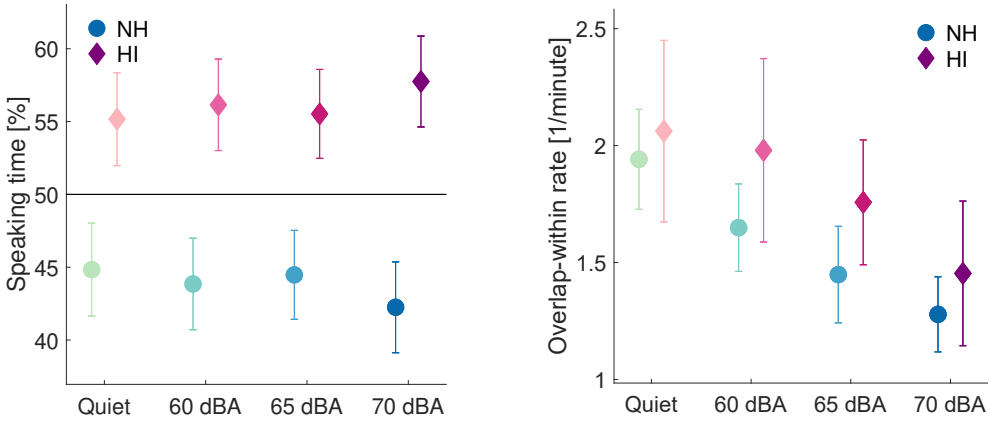


Fig. 3: Speaking time (the percentage of time a person speaks during the conversation) (left panel) and rate of occurrence of overlaps-within (i.e., turns from one talker that occur completely within a turn of the other talker) (right panel) in the four backgrounds (quiet, 60, 65 and 70 dBA SPL) for younger normal-hearing and older hearing-impaired listeners. Note that the rate has been normalized by the total phonation time rather than duration of the conversation. The bars indicate standard error.

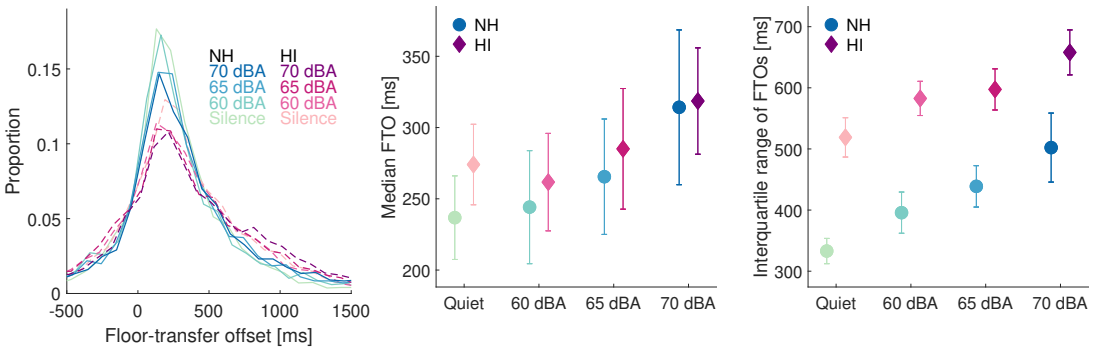


Fig. 4: Normalized distributions (left panel) of floor-transfer offsets (FTOs) along with the median (middle panel) and interquartile range (right panel) for the four combinations of language and noise. The bars indicate standard error.

In the middle and right panels of Figure 4, the median FTO and interquartile range (IQR) of FTOs are plotted. There was a statistically significant main effect of background on the median FTO $[F(3,261) = 10.89, p < 9.14\text{e-}7]$, but no significant main effect of hearing status or replicate and no interactions. For the IQR, there

were significant main effects of both background [$F(1,272) = 15.4, p < 2.78\text{e-}9$] and hearing status [$F(1,272) = 135.4, p < 2.2\text{e-}16$], confirming the visual impression that the distributions were broader for the HI group.

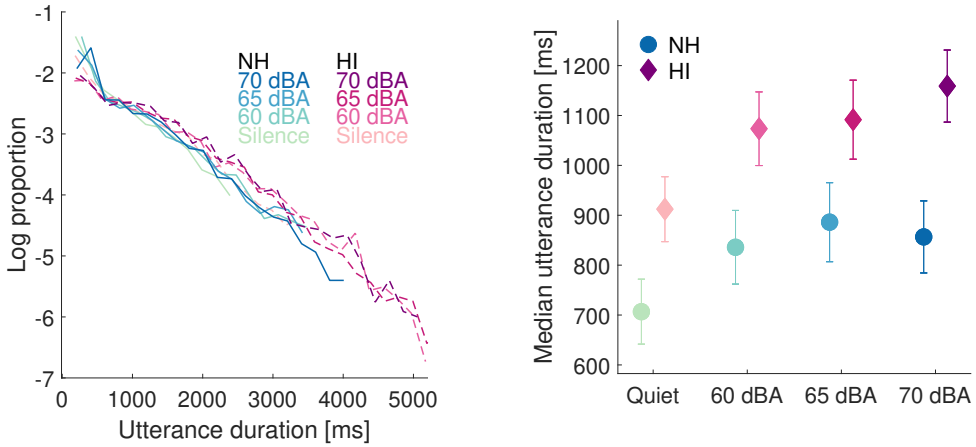


Fig. 5: Normalized distributions (left panel) and median (right panel) of the duration of utterances for younger normal-hearing and older hearing-impaired talkers in the four backgrounds. The bars indicate standard error. The density in the left panel has been log-transformed to more easily compare the slopes.

The distributions of utterances (estimated using 200 ms bin widths) along with the median utterance duration for the two hearing statuses and four backgrounds are seen in the left and right panels of Figure 5, respectively. There was a significant main effect of both hearing status [$F(1,272) = 52.48, p < 4.48\text{e-}12$] and background [$F(3,272) = 7.49, p < 7.7\text{e-}5$] on the median utterance duration, with both groups increasing their utterance durations in noise and the HI group producing about 25 % longer utterances than their NH interlocutor.

DISCUSSION

Speaking levels

On average, the older HI spoke about 2.5 dB louder than the younger NH when there was no background noise, but the two groups increased their speaking levels to achieve almost the same SNR in background noise. With increasing noise level, there was a decrease in the SNR. At 60, 65, and 70 dBA SPL, the average SNR was 7, 3.5 and 1 dB, respectively. It is physically strenuous to speak at a high sound pressure level, so talkers may trade off speech understanding and physical effort. However, in all conditions the participants spoke at positive SNRs, whereas in the NH/NH conversations in Sørensen *et al.* (2020) the participants spoke at -2.5 dB SNR in 70 dBA noise. This shows adaptive behavior from the younger NH talker to their older HI interlocutor, adjusting to their hearing difficulty. The HI may both speak at a

positive SNR for their own auditory feedback, but also to signal difficulty so that their interlocutors increase their voice level.

Utterances

Similarly to the experiments with NH/NH of Sørensen *et al.* (2020) and Watson *et al.* (2020), the participants lengthened their utterances in noise. This may be a strategy to give themselves and their interlocutor more time to plan their response. It may also be a strategy to meet appropriate response times by initiating turns while still not fully planned, resulting in lengthened turns. While the NH increased their duration of utterances to the same extent as those in Sørensen *et al.* (2020), the HI lengthened their turns significantly more. This supports the interpretation that the older HI talkers are more challenged than the younger NH talkers. From the distributions of utterances it seems that the overall utterance duration of the older HI are longer, indicated by the shallower slope. An immediate interpretation is that it could be explained by the lower articulation rate of the older HI. However, the articulation rates of NH are similar to those of the older HI in the 70 dBA condition, yet utterance durations remain different.

Articulation rates

In other studies, when talking to an NH partner, NH talkers increased their articulation rates in noise (Sørensen *et al.*, 2020; Watson *et al.*, 2020). However, in this study we saw the opposite trend: with increasing noise level, the NH participants decreased their articulation rates. In general, the HI talkers spoke slower than the NH talkers. This may just be an effect of age, but it may also be a signalling strategy to their NH interlocutors to slow down articulation to ease speech understanding and reduce the communication challenge for the HI interlocutor.

Overlaps-within

Another indication of the NH's adaptive and accommodating behaviour is the rate at which overlaps-within occur. With increasing background noise level, the rate goes down for both groups. In both Sørensen *et al.* (2020) and Watson *et al.* (2020), when talking to an NH partner, the NH increased their rate of overlaps-within in background noise, regardless of whether they were unacquainted or not, or if they spoke in free conversation or solved the Diapix task. This was attributed to an increased stress. However, more overlaps are likely to decrease the information transmission and increase the cognitive load on participants. It was observed in Watson *et al.* (2020) that when participants spoke freely, they had a higher rate of overlaps-within than when they solved the Diapix task, where information transfer is presumably more crucial. Here, the younger NH may have adopted an even more careful turn taking strategy in conversations with older HI talkers to increase information transfer. This may also be why we found a delay in FTOs, not just because of increased cognitive load, but also to actively reduce overlaps that likely reduce speech understanding.

SUMMARY

For both the younger NH and older HI talkers, floor-transfer offsets were longer and more variable in background noise, and the older HI showed more variability than the NH talkers. There were indications that the NH adapted their speech to accommodate their interlocutor's difficulty. For example, they adapted their speaking rates to speak slower with increasing noise level, opposite to what was found in our previous studies with NH interlocutors (Sørensen *et al.*, 2020; Watson *et al.*, 2020). Moreover, both groups decreased their rates of overlaps-within with increasing noise level, but the NH decreased their rates significantly more than the older HI. The NH also produced significantly shorter utterances than the HI and spoke less of the time than the HI.

ACKNOWLEDGEMENTS

A.J.M.S. and a portion of this study was supported by the William Demant Foundation (16-3968).

REFERENCES

Baker, R., and Hazan, V. (**2011**). "DiapixUK: Task Materials for the Elicitation of Multiple Spontaneous Speech Dialogs," Behav. Res. Methods, **43**(**3**), 761–70. doi: 10.3758/s13428-011-0075-y.

Beechey, T., Buchholz, J. M., and Keidser, G. (**2018**). "Measuring communication difficulty through effortful speech production during conversation," Speech Commun., **100**, 18–29. doi: 10.1016/j.specom.2018.04.007.

de Jong, N. and Wempe, T. (**2009**). "Praat script to detect syllable nuclei and measure speech rate automatically," Behav. Res. Methods, **41**, 385-90. doi: 10.3758/BRM.41.2.385.

Hadley, L. V., Brimijoin, W. O., and Whitmer, W. M. (**2019**). "Speech, movement, and gaze behaviours during dyadic conversation in noise," Sci. Rep., **9**(1), 10451. doi: 10.1038/s41598-019-46416-0.

Levinson, S. C., and Torreira, F. (**2015**). "Timing in turn-taking and its implications for processing models of language," Front. Psychol., **6**, 731. doi: 10.1038/s41598-019-46416-0.

Sørensen, A. J. M., Fereczkowski., M, and MacDonald, E. N. (**2020**). "Effects of noise and L2 on the timing of turn taking in conversation," Proc. ISAAR, **7**, 85-92.

Watson, S., Sørensen, A. J. M., and MacDonald, E. N. (**2020**). "The effect of conversational task on turn taking in dialogue," Proc. ISAAR, **7**, 61-68.

The effect of harmonic number and pitch salience on the ability to understand speech-on-speech based on differences in fundamental frequency

SARA M. K. MADSEN[1,*], TORSTEN DAU[2] AND ANDREW J. OXENHAM[1]

[1] *Department of Psychology, University of Minnesota, 75 East River Parkway, Minneapolis, MN, 55455, USA*

[2] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Ørsteds Plads, Building 352, 2800 Lyngby, Denmark*

Differences in fundamental frequency (F0) between competing voices facilitate the ability to segregate a target voice from interferers, thereby enhancing speech intelligibility. Although lower-numbered harmonics produce greater pitch salience than higher-numbered harmonics, it remains unclear whether differences in harmonic ranks, and therefore pitch salience, affect the benefit of pitch differences. Earlier studies have not reported an effect of pitch salience, but have generally used only conditions where the difference in average F0 (ΔF0) between the two competing voices was large. It is possible that the effect of pitch salience is greater in more challenging conditions, in which the ΔF0 is relatively small. This study tested speech intelligibility in the presence of one speech masker for ΔF0s of 0, 2, and 4 semitones. The speech was presented in a broadband condition or was highpass or lowpass filtered to manipulate the pitch salience of the voicing. Results showed no interaction between filter type and ΔF0, suggesting little or no effect of harmonic rank or pitch salience in the ability to use F0 to segregate voices, even with smaller ΔF0s between competing voices. The results suggest some benefit of ΔF0 between competing voices, even in the absence of low-numbered spectrally resolved harmonics.

INTRODUCTION

Pitch differences between competing voices can enhance our ability to segregate target speech from a background of other speakers (Bird and Darwin, 1998; Brokx and Nooteboom, 1982). It is, for example, easier to understand a female speaker masked by a male speaker than by another female speaker. Therefore, it seems plausible that the ability to make use of pitch differences would improve with the strength (salience) of the pitch of the speech. However, the importance of pitch salience for understanding speech in a speech background is unclear.

Pitch salience has been shown to affect the ability to discriminate small differences in F0 between consecutive complex tones in studies that varied the pitch salience by

*Corresponding author: madse399@umn.edu

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

varying the number of the lowest harmonic component (rank; low harmonic rank is associated with high pitch salience and vice versa) included in the stimuli (Bernstein and Oxenham, 2006; Hoekstra and Ritsma, 1977; Shackleton and Carlyon, 1994). More specifically, these studies found that thresholds are lowest for tones with low harmonic rank and increase with increasing lowest harmonic rank, often reaching a plateau when the lowest harmonic present exceeds the 10^{th}. This is also the point at which no harmonics are thought to be spectrally resolved, although the link between spectral resolvability and harmonic number remains uncertain (Bernstein and Oxenham, 2003; Graves and Oxenham, 2019).

Psychoacoustic studies of sound segregation based on fundamental frequency (F0) differences have often been carried out with interleaved sequences of tones. One such study did not find a difference in amount of perceived segregation as a function of F0 difference between tones consisting only of high harmonic ranks and tones containing low harmonic ranks (Vliegen and Oxenham, 1999), whereas others did find a significant effect of harmonic rank on segregation (Grimault *et al.*, 2000; Madsen *et al.*, 2018). Moreover, one of the studies found a correlation between F0 difference limens (DLs) and performance in a stream segregation task (Madsen *et al.,* 2018), supporting the idea that perceptual salience of cues used for segregation is important for the ability to segregate sounds.

However, even if pitch salience is important for sound segregation, it is not necessarily important for speech-on-speech perception. One study explored the effect of harmonic rank on speech intelligibility by comparing conditions where the target and one speech masker had been either lowpass (LP) or highpass (HP) filtered to retain or remove resolved components, respectively (Oxenham and Simonson, 2009). Surprisingly, similar speech intelligibility and masking release were found in both conditions, suggesting no benefit of having resolved harmonic components in the speech. However, this study only tested conditions where the long-term average F0 difference between the voices (ΔF0) was four or eight semitones (ST), according to recent F0 estimates obtained with Praat (Boersma and Weenink, 2009). It may be that pitch salience is only relevant for more challenging conditions, i.e. for conditions with smaller values of ΔF0.

The aim of the present study was to determine whether there is an effect of harmonic rank on the ability of listeners to use differences in F0 between a target talker and an interfering speech masker to understand speech. Speech from a target speaker and a masker was either LP filtered (low harmonic rank condition) or HP filtered (high harmonic rank condition) and the masker was manipulated with Praat to obtain conditions where the long-term average F0 of the target and masker was separated by 0, 2, or 4 STs. F0DLs were measured for a subset of the participants to confirm that the filter cutoff frequencies used yielded conditions with and without spectrally resolved harmonics, yielding good and poor pitch discrimination, respectively.

METHODS

General methods

The experiments were conducted in a double-walled acoustically shielded booth. The stimuli were generated in MATLAB (The Mathworks, Natick, MA, USA) and were presented at a sampling rate of 48000 Hz via a Fireface UCX sound card (RME, Haimhausen Germany) and Sennheiser HD 650 headphones (Sennheiser, Wedemark, Germany). The experimental protocols were approved by the Scientific Ethical Committees of the Capital Region of Denmark (H-16036391).

Speech experiment

Eighteen native-Danish-speaking participants (9 female) between 20 and 28 years (mean = 23.67, SD = 2.45) were tested. All participants had audiometric thresholds at octave frequencies between 250 and 8000 Hz no greater than 20 dB HL.

Speech intelligibility was tested for sentences masked by one speech masker for conditions where the speech material was either HP filtered, LP filtered, or unfiltered (broadband). The masker speech was manipulated in Praat to generate conditions where the difference in average long-term F0 of the target and masker (ΔF0) varied. The target consisted of sentences from the CLUE speech corpus (Nielsen and Dau, 2009). These are short contextual sentences similar to the HINT sentences (Nilsson *et al.*, 1994) that had a duration between 1.23 and 1.86 s. Speech from recordings of conversations (Sørensen *et al.*, 2018) was used as maskers. The recordings from two speakers were concatenated separately and all gaps exceeding 100 ms, non-Danish words, loud exclamations, and other sounds such as laughter were removed. The maskers for the main test were generated from the remaining speech material from one of the speakers that had a duration of 222.3 s and was divided into 180 overlapping segments of 2.47 s. Similarly, the maskers for the training (30 blocks of 2.47 s) were made from the remaining speech material from the other speaker. The maskers started 500 ms before the target and ended at least 100 ms after the target and were gated with 50 ms raised-cosine onset and offset ramps. One CLUE sentence was presented in quiet immediately before each trial to guide the participant towards the target voice. The guide sentence was always the same and was never the same as the target sentence. All speakers were male. The long-term average F0 was approximately 110 Hz for the target, 139 Hz for the masker used for testing, and 148 Hz for the masker used for training. The F0s of the maskers were manipulated in Praat to obtain differences between the long-term average F0 of the target and masker (ΔF0) of 1, 3, and 5 STs for the training and 0, 2, and 4 STs for the main test. The average long-term F0 of the masker was always the same as or higher than that of the target. The speech maskers were filtered to have the same long-term spectrum as the CLUE sentences. For the HP- and LP-filtered conditions, the target and masker were filtered with a 4^{th}-order Butterworth filter after being combined. The guide sentences were filtered with the same filter. The conditions with ΔF0 of 4 STs were used as a reference since it is the only ΔF0 tested both here and in the study by Oxenham and Simonson (2009).

Cutoff frequencies of 800 Hz and 1500 Hz were chosen for the LP- and HP-filtered conditions, respectively, since pilot experiments indicated that they would yield similar performance. A target-to-masker ratio (TMR) of 0 dB was used for the filtered conditions and a TMR of -15 dB was used for the broadband conditions to obtain similar performance in the filtered and broadband conditions for ΔF0 = 4 STs.

A Gaussian noise with the same long-term spectrum as the CLUE sentences (before filtering) was filtered with a 4th-order Butterworth filter and added to the filtered speech stimuli. For the LP-filtered condition, the noise was HP filtered with a cutoff frequency of 800 Hz and for the HP-filtered condition, the noise was LP filtered with a cutoff frequency of 1500 Hz. The level of the noise before filtering was 12 dB lower than the unfiltered target speech. The target and maskers combined were presented at an overall sound pressure level (SPL) of 70 dB.

In the main test, each of the nine conditions (three filter conditions and three ΔF0s) was tested with two lists each containing 10 sentences. The order of the conditions was randomized within each of two consecutive blocks, both containing all of the nine conditions. The training consisted of three runs presented in the following order: 1) Broadband with ΔF0 of 5 STs presented at a TMR of -12 dB; 2) HP filtered with ΔF0 of 3 STs, presented at a TMR of 3 dB; 3) LP filtered with ΔF0 of 1 ST, presented at a TMR of 0 dB.

The participants were instructed to listen for the voice of the guide sentence and were asked to type what they heard that voice said, after each trial. The speech scores were transformed into rationalized arcsine units (RAU) before statistical analysis.

F0 discrimination limens

F0 discrimination limens (F0DLs) were measured with a two-interval, three-down, one-up adaptive procedure where each interval contained four 200-ms tones, presented immediately after each other as in earlier studies (Madsen *et al.*, 2019; Madsen *et al.*, 2017). In the reference interval, all tones had the same F0, the reference F0, that was roved over two semitones and centred on 131 Hz (corresponding to one standard deviation above the long-term average F0 of the target speech). In the target interval, the F0 of the first and third tone was higher and the F0 of the second and fourth tone was lower than the reference F0. The difference in F0 between the high and low tones was varied adaptively, while the F0s of the tones remained geometrically centered on the reference F0. All tones were gated with raised-cosine ramps of 20 ms and the two intervals were separated by a 400 ms pause. The participants were asked to indicate which interval contained the changes in pitch. Feedback was provided after each trial.

The harmonic components were added in either sine or random phase. For the latter, the phase was for each component chosen randomly and independently from a uniform distribution from 0 to 2π. As in the speech experiment, the tones were either LP or HP filtered with a fourth-order Butterworth filter using cutoff frequencies of

800 Hz and 1500 Hz, respectively. Moreover, as in the speech experiment, a HP-filtered Gaussian noise with cutoff frequency of 800 Hz was added in the LP filtered condition and a LP-filtered Gaussian noise with cutoff frequency of 1500 Hz was added in the LP filtered condition. The average overall level of the tones was 70 dB SPL but the level was roved independently for each tone over a uniform range of 6 dB. The noise was presented at a level 12 dB below the nominal level of the tones before LP or HP filtering.

For each run, the thresholds were calculated as the geometric mean across the last six reversals. The experiment contained three blocks with one run for each condition and the order of the conditions were randomized within a block. The first run was used for training and the final thresholds were defined as the geometrical mean across the two last runs.

RESULTS

Speech intelligibility was measured as the proportion of words reported correctly in each condition. All deviations except obvious misspellings or homophones were considered incorrect. Additional words and differences in word order were not penalized.

The scores for the reference conditions ($\Delta F0 = 4$ STs) and broadband conditions are shown in the left and middle panel of Fig. 1, respectively. It can be seen that the mean scores are very similar for the three reference conditions, and a repeated-measures ANOVA with filter type as the within-subjects factor found no significant effect of filter condition [$F(2,34) = 0.47$, $p = 0.63$]. This finding confirmed that the cutoff frequencies chosen for the HP- and LP-filtered conditions and the TMRs chosen for the filtered and broadband conditions yielded similar performance in the three reference conditions. The speech scores for the broadband conditions were analyzed separately since the filtered conditions were measured at a higher TMR than the broadband conditions. For the broadband condition, there was a tendency for the scores to increase slightly with increasing $\Delta F0$ even though the scores for $\Delta F0 = 0$ and $\Delta F0 = 2$ were very similar to each other. Analysis of the speech scores with $\Delta F0$ as the within-subject factor showed a small but significant effect of $\Delta F0$ [$F(2,34) = 3.35$, $p = 0.047$]. Moreover, Bonferroni-corrected post-hoc tests showed a significant difference between the conditions with $\Delta F0$ of 0 and 4 semitones [$t(34) = -2.55$, $p = 0.046$] but not between either of the other pairs of conditions. The right panel of Fig. 1 shows individual and mean speech scores for the LP- and HP-filtered conditions for $\Delta F0$s of 0, 2, and 4 STs. As expected, scores generally increased with increasing $\Delta F0$. Furthermore, the scores were generally higher for the HP than for the corresponding LP conditions especially for $\Delta F0 = 0$ STs and for $\Delta F0 = 2$STs. Analysis with $\Delta F0$ and filter condition as within-subject factors show a significant effect of both $\Delta F0$ [$F(2, 34) = 24.37$, $p < 0.0001$] and filter condition [$F(1,17) = 9.51$, $p = 0.0067$] but no interaction between $\Delta F0$ and filter condition [$F(2, 34) = 1.25$, $p = 0.3$], indicating that low and high harmonics both facilitate improvements in performance with F0 differences at similar rates in the presence of competing voices.

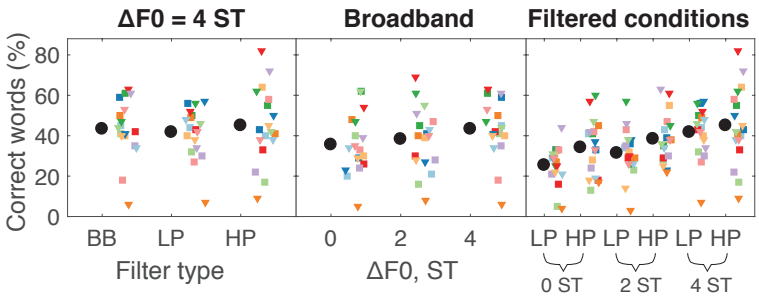


Fig. 1: Speech scores for the reference conditions (left panel), the broadband condition (middle panel), and the LP- and HP-filtered conditions (right panel). Larger circles represent the mean across participants and the smaller symbols show the individual scores.

F0 discrimination

F0 discrimination thresholds are shown in Fig. 2. As expected, F0DLs were higher in the HP conditions than in the LP conditions, and phase affected F0DLs in the HP, but not in the LP, conditions. There were significant effects of both filter type [$F(1, 5) = 40.51$, $p = 0.00011$] and phase [$F(1,5) = 11.99$, $p = 0.018$] and a significant interaction between phase and filter condition [$F(1,5) = 7.60$, $p = 0.040$]. The results are consistent with expectations based on high-ranked unresolved harmonics being present when the stimuli were HP filtered with a cutoff frequency of 1500 Hz, as in the speech experiment. The results confirm that the pitch of the speech was less salient under HP conditions than under LP conditions.

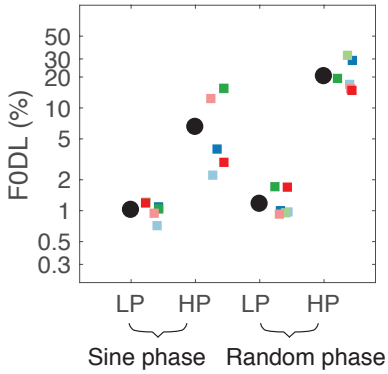


Fig. 2: F0 discrimination thresholds for complex tones with components added in sine or random phase and then LP or HP filtered at 800 Hz or 1500 Hz, respectively.

DISCUSSION

In the speech experiment, there were significant effects of ΔF0 and filter type but no significant interaction between ΔF0 and filter condition. The improvement in speech scores with increasing ΔF0 is consistent with results from earlier studies (Bird and Darwin, 1998; Brokx and Nooteboom, 1982; Madsen *et al.,* 2017). The lack of a main effect of filter type when considering the reference conditions confirms that we were successful at selecting filter cutoff frequencies that produced roughly equal

performance in the LP and HP conditions when ΔF0 = 4 ST. Due to this equalization of performance, it is more relevant to compare the slopes of the scores for the HP-filtered conditions relative to the LP-filtered conditions as a function of ΔF0 instead of the absolute scores. The seemingly steeper slope for the LP condition compared to the HP condition suggests that pitch is more important for the LP than for the HP condition. This support the idea that the more salient pitch cues in the LP condition lead to better separation of the voices and therefore better speech intelligibility. However, the lack of a significant interaction between filter condition and ΔF0 indicates that this effect is not robust. The lack of interaction is consistent with the results from Oxenham and Simonson (2009), which showed similar performance for a HP- and a LP-filtered conditions for ΔF0s of 4 STs and 8 STs. This may be explained by the different forms of speech information conveyed in the low and high spectral region, respectively, or by the difference strength of masker modulation in the two spectral regions as proposed by Oxenham and Simonson (2009). Another possible explanation is that, despite testing the smallest possible long-term average ΔF0 of 0 STs, the momentary differences in ΔF0 might have been too large for differences in pitch salience to affect speech intelligibility. This would suggest that pitch salience would not be an issue for understanding speech-on-speech in real-life situations.

In summary, this study tested speech intelligibility in a background of a speech masker and found a small effect of ΔF0 but a similar relation between performance in LP- and HP-filtered conditions for different ΔF0s. This suggest that the difference in pitch salience between low-numbered and high-numbered harmonics is not a determining factor for the ability to use F0 differences between competing talkers to better understand speech.

ACKNOWLEDGEMENTS

Work supported by the Demant Foundation and NIH grant R01 DC005216.

REFERENCES

Bernstein, J.G.W., & Oxenham, A.J. (**2003**). "Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number?" J. Acoust. Soc. Am., **113**(6), 3323–3334. doi: 10.1121/1.1572146

Bernstein, J.G.W., & Oxenham, A.J. (**2006**). "The relationship between frequency selectivity and pitch discrimination: Sensorineural hearing loss," *J*. Acoust. Soc. Am., **120**(6), 3929–3945. doi: 10.1121/1.2372452

Bird, J., & Darwin, C.J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," *Psychophysical and Physiological Advances in Hearing*, 263–269.

Boersma, P., & Weenink, D. (**2009**). Praat: Doing phonetics by computer (Version 5.1.3.1).

Brokx, J.P.L., & Nooteboom, S.G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon., **10**, 23–36.

Graves, J.E., & Oxenham, A.J. (**2019**). "Pitch discrimination with mixtures of three concurrent harmonic complexes," J. Acoust. Soc. Am., **145**(4), 2072-2083.

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., & Collet, L. (**2000**). "Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency," J. Acoust. Soc. Am., **108**(1), 263–271. doi: 10.1121/1.429462

Hoekstra, A., & Ritsma, R.J. (**1977**). "Perceptive hearing loss and frequency selectivity," *Psychophysics and Physiology of Hearing*, 263–271.

Madsen, S.M.K., Dau, T., & Moore, B.C.J. (**2018**). "Effect of harmonic rank on sequential sound segregation," Hear. Res., **367**, 161–168. doi: 10.1016/j.heares.2018.06.002

Madsen, S.M.K., Marschall, M., Dau, T., & Oxenham, A.J. (**2019**). "Speech perception is similar for musicians and non-musicians across a wide range of conditions," Sci. Rep., **9**(1), 1-10. doi: 10.1038/s41598-019-46728-1

Madsen, S.M.K., Whiteford, K.L., & Oxenham, A.J. (**2017**). "Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds," Sci. Rep., **7**, 1-9. doi: 10.1038/s41598-017-12937-9

Nielsen, J.B., & Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol., **48**(10), 729–741. doi: 10.1080/14992020903019312

Nilsson, M., Soli, S.D., & Sullivan, J.A. (**1994**). "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am., **95**, 1085–1099. doi: 10.1121/1.408469

Oxenham, A.J., & Simonson, A.M. (**2009**). "Masking release for low- and high-pass-filtered speech in the presence of noise and single-talker interference," J. Acoust. Soc. Am., **125**(1), 457–468. doi: 10.1121/1.3021299

Shackleton, T.M., & Carlyon, R.P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination," J. Acoust. Soc. Am., **95**(6), 3529–3540. doi: 10.1121/1.409970

Sørensen, A.J., Fereczkowski, M., & MacDonald, E.N. (**2018**). "Task dialog by native-Danish talkers in Danish and English in both quiet and noise," Zenodo. doi: 10.5281/zenodo.1204951

Vliegen, J., & Oxenham, A.J. (**1999**). "Sequential stream segregation in the absence of spectral cues," J. Acoust. Soc. Am., **105**, 339–346. doi: 10.1121/1.424503

Perceptual learning and speech perception: A new hypothesis

KAREN BANAI[1,*] AND LIMOR LAVIE[1]

[1]*Department of Communication Sciences and Disorders, University of Haifa, Haifa, Israel*

Perceptual learning for speech remains substantial even in older adults, but the functional significance of this observation is not well understood. It has been suggested that perceptual learning might serve to support listening in adverse conditions by promoting behavioural and neural plasticity, but this hypothesis is not consistent with the acoustic specificity of learning. Instead, we now suggest that in the context of speech perception, perceptual learning might be best viewed as one of the capacities that, like working memory, support speech perception in an on-line fashion. Consistent with this hypothesis, we present data that rapid perceptual learning of one speech task accounts for substantial individual differences in other speech tasks even after accounting for the potential correlations between different indices of speech perception.

INTRODUCTION

Perceptual learning, defined as relatively long lasting experience-dependent changes in the processing of sensory stimuli, has been documented across sensory modalities and age groups (Green *et al.,* 2019). Speech, the focus of this paper, is also subject to perceptual learning that can occur rapidly, even following few minutes of exposure (for review see Samuel and Kraljic, 2009). Over the last decades, attempts were made to develop perceptual training regimens for hearing rehabilitation (e.g., Sweetow and Sabes, 2006), but a more recent systematic review (Henshaw and Ferguson, 2013) concluded that the evidence for the efficacy of such programs is weak. One of the reasons cited is that while robust, perceptual learning is also quite specific. For example, although older adults with and without hearing loss retain substantial learning of both speech in noise (Burk and Humes, 2008; Karawani *et al.,* 2016) and time-compressed speech (Manheim *et al.,* 2018), transfer of learning is limited by the acoustic and semantic similarity of new materials to those experienced in training. These findings of robustness and specificity raise the question of the role of perceptual learning in speech recognition. If past learning fails to modify future speech recognition to a substantial degree, what (if any) is the functional role of perceptual learning?

While struggling with the implications of the specificity of training-induced learning, we observed strong correlations between speech perception and perceptual learning across age and hearing levels. First, although the transfer of learning on time-compressed speech is limited, rapid learning (following < 5 minutes of listening) of

*Corresponding author: kbanai@research.haifa.ac.il

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

time-compressed speech is highly correlated with the recognition of natural-fast speech (Manheim *et al.,* 2018). Second, the amount of within-session learning on one speech in noise task, was strongly correlated with naïve performance on another speech in noise task (Karawani *et al.,* 2017). It could have been suggested that the correlations we observed reflect either rapid learning on the tasks which we used to measure 'perception', or the transfer of learning during the rapid learning phase to the other, different, task. Both explanations seem unlikely because, in the two studies we refer to, the perceptual tests were too brief to elicit rapid learning on either speech in noise or natural-fast speech. Furthermore, the learning phase in these two studies seems too brief to result in transfer (Adank and Janse, 2009; Wright *et al.,* 2010; Banai and Lavner, 2014). We therefore proposed an alternative view, that rapid perceptual learning, which underlies perceptual adjustment to challenging or unusual speech (e.g., Mattys *et al.,* 2012), might serve as one of the factors contributing to speech perception in adverse conditions. By this account, perceptual learning might play a role in speech perception, similar to that played by cognitive functions such as attention and working memory (Ronnberg *et al.,* 2019). The study described here is an attempt to test one of the predictions of this account. Namely, if rapid learning is some form of general capacity, the correlations observed in our previous studies should be replicated under conditions with less similarity between the stimuli used to assess rapid learning and those used to estimate speech perception. To this end, we used a time-compressed speech (TCS) task to estimate rapid perceptual learning, and speech in noise (SIN) and naturally fast speech (NFS) as indices of speech perception under adverse conditions. TCS was selected as the learning task due to the large number of studies that documented rapid, robust and long lasting learning with this task across age groups (Altmann and Young, 1993; Dupoux and Green, 1997; Peelle and Wingfield, 2005).

METHODS

Participants

Seventy-eight young adults, all native Hebrew speakers (ages 18-35, M = 26, SD = 4; 38 female) participated in this study as unpaid volunteers. By self-report, all participants had normal hearing and no history of learning, language or neurological disorder. Most participants were undergraduate students at the University of Haifa and other academic institutes in the region or recent graduates.

Procedure

Each participant completed two sessions, held 5-9 days apart. On the first session, we assessed the perception of time-compressed speech. On the second session, the recognition of time-compressed speech was tested again in order to calculate a between-session learning index. Subsequently, each participant completed a speech in noise and a natural-fast speech test. The order of these tests was counterbalanced across participants. Stimuli were presented diotically through headphones (Sennheiser HD-205 or 215) in a quiet room on campus or at their homes. All aspects of the study

were approved by the ethics committee of the Faculty of Social Welfare and Health Sciences, University of Haifa (IRB 199/12).

Tasks and stimuli

For all tasks, we used 5-6 word sentences in Hebrew (Prior and Bentin, 2006), produced by two native female Hebrew speakers, recorded and amplitude normalized using the Audacity software. Each sentence was presented only once throughout the two study sessions. During each task participants were instructed to transcribe each sentence after it was played. The percentage of correct words was computed and used in data analysis. Only perfectly reported words were counted as correct (for further details on the word-scoring method see Manheim *et al.*, 2018).

Rapid perceptual learning of time-compressed speech (TCS).

On each session, 10 different sentences were presented. We defined rapid learning as the difference in transcription accuracy between the two sessions. Stimuli for this task were recorded by talker 1 at an average natural speech rate of 111 words/minute (SD = 17) and then compressed to 30% of their natural duration using a WSOLA algorithm (Verhelst and Roelands, 1993).

Natural-fast speech perception.

Twenty sentences were presented by talker 2 at an average natural fast rate of 214 words/minute (SD = 26).

Speech-in-noise perception.

Twenty sentences were presented diotically by talker 1 mixed in 4-talker babble noise (for details see Karawani *et al.*, 2016). Speech and noise were presented simultaneously, in each participant's most comfortable level. Signal-to-noise ratio was -6 dB.

RESULTS

Rapid perceptual learning of time-compressed speech

As shown in Fig. 1, 75% of participants improved their performance between the two sessions with a median improvement of 10% (*IQR* = 1-20%; $Z = 6.05$, $p < 0.001$). These data are consistent with previous findings on the rapid learning of time-compressed speech and its retention over time, and suggest that there are substantial individual differences in the magnitude or rate of rapid learning across participants.

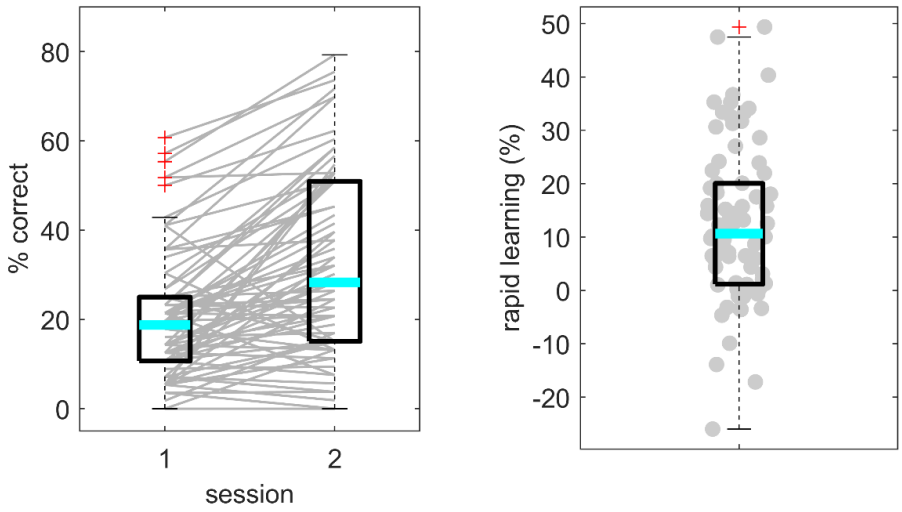


Fig. 1: Recognition (left) and rapid across-session perceptual learning (right) of time compressed speech. Box edges mark the inter-quartile range; thick line within each box marks the median; Whiskers are 1.5 the IQR. Grey symbols/lines show individual data.

Rapid learning and individual differences in speech recognition

Rapid learning and speech recognition were significantly correlated (SIN: $r = 0.35$, $p = 0.002$; NFS: $r = 0.44$, $p < 0.001$). These correlations (especially with NFS) may have been expected because rapid learning was also assessed using a speech task. Therefore, we attempted to statistically partial out the contribution of the TCS to these correlations. To this end, linear regression models were used in which baseline recognition of TCS (the first 5 sentences of the first session) were entered to the model first, and the rapid learning index was entered on a second stage. Although this is not a perfect control, baseline performance was not correlated with the rapid learning index ($r = 0.08$, $p = 0.48$). Details of these models, which generally conformed with the assumptions of liner models (tolerance > 0.95, VIF < 1.5), are shown in Table 1. The unique contribution of rapid learning to the perception of each type of perceptually difficult speech is depicted on Fig 2.

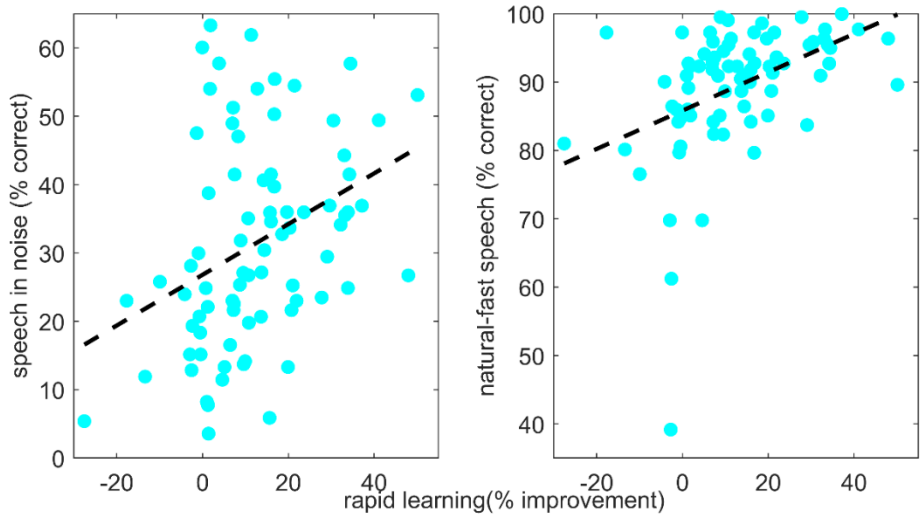


Fig. 2. Speech perception as a function of rapid learning of time-compressed speech. Partial correlation after controlling for baseline recognition of time-compressed speech is shown (for details of the formulas used see Manheim et al., (2018)).

	Predictor	R^2	F	B	t
SIN	TCS perception			0.30	2.93**
	Rapid learning	0.11	10.22**	0.33	3.20**
	Full model	0.21	10.23***		
NFS	TCS perception			.16	1.58
	Rapid learning	0.18	17.61***	.43	4.20***
	Full model	0.20	10.66***		

Table 1. Speech perception as a function of baseline speech perception (TCS) and rapid perceptual learning. R^2 - the unique contribution of rapid perceptual learning followed the proportion of variance explained by the full model; β – standardized regression coefficients. $^{**}p < 0.005$, $^{***}p < 0.001$.

DISCUSSION

In the current study, we tried to broaden our understanding of the role of perceptual learning in speech perception under adverse conditions. Previous observations led us to suggest that age-related declines in rapid learning may contribute to speech perception deficits in older adults (Manheim *et al.*, 2018). This reflects a more general view of rapid perceptual learning as one of the factors contributing to individual differences in speech perception under adverse conditions. By this account, under a broad array of adverse conditions, rapid learning allows listeners to quickly calibrate speech processing based on the acoustic demands of the ongoing situation. The data presented in this manuscript is consistent with this idea. Although correlational in

nature, it nevertheless shows that individual differences in rapid learning of time-compressed speech are associated with individual differences in the perception of natural-fast speech and speech in noise. The correlation for speech in noise is especially telling because time-compressed speech and speech in noise represent different domains of speech perception (as opposed to natural-fast and time-compressed speech which represent two forms of rapid speech) and possibly rely on different sensory processes and top-down strategies. Furthermore, the use of TCS as baseline in the statistical model should have accounted for the contribution of the processes shared by TCS and SIN.

More studies are required to test our hypothesis further. Going forward, it will be important to control for additional variables that are thought to be involved in both speech perception and perceptual learning (e.g., working memory and inhibition), as well as to account for the relationships between rapid learning and longer-term learning. It will also be of interest to consider populations with more variance in age and hearing levels than tested here. In an ongoing study in our lab we are looking at older adults with presbycusis to determine whether the pattern of correlations reported here is modified by age, hearing status and hearing aid use, and whether rapid learning is associated with clinically relevant indices of speech perception (preliminary outcomes have been submitted by Rotman *et al.* (2020) to these proceedings). If our hypothesis is correct, it follows that rapid learning might be one of the factors partially predicting how well listeners will adapt to new hearing aids. We acknowledge that none of these proposals can provide definitive proof of our hypothesis, but negative findings can certainly falsify it.

Our hypothesis was driven to a great extent by our frustration with the literature on the potential clinical application of auditory training (Pichora-Fuller and Levitt, 2012). Specifically, we maintain that neither bottom-up nor top-down approaches were successful in broadening the scope of learning generalization. Nevertheless, although auditory training studies failed to prove effective under rigorous scrutiny (Henshaw and Ferguson, 2013; Saunders *et al.,* 2016), there are individuals who report training-related benefits (see Lavie *et al.,* 2013; Karawani *et al.,* 2016). It is therefore interesting to end with speculating that perhaps, for some individuals, training may have pushed rapid perceptual learning to support speech perception under ecological conditions, rather than have a direct impact on speech perception. Testing this speculation requires further studies.

ACKNOWLEDGEMENTS

This study was supported by the Israel Science Foundation grant 206/18. We thank our 2017-2018 audiology students who collected the data for this manuscript as part of their undergraduate research projects.

REFERENCES

Adank, P. and Janse, E. (**2009**). "Perceptual learning of time-compressed and natural fast speech," J. Acoust. Soc. Am., **126**, 2649-2659. doi: 10.1121/1.3216914

Altmann, T. and Young, D. (**1993**). "Factors affecting adaptation to time-compressed speech," EUROSPEECH '93. Berlin**,** 333-336.

Banai, K. and Lavner, Y. (**2014**). "The effects of training length on the perceptual learning of time-compressed speech and its generalization," J. Acoust. Soc. Am., **136**, 1908. doi: 10.1121/1.4895684

Burk, M. H. and Humes, L. E. (**2008**). "Effects of long-term training on aided speech-recognition performance in noise in older adults," J. Speech Lang. Hear. Res., **51**, 759-771. doi: 10.1044/1092-4388(2008/054)

Dupoux, E. and Green, K. (**1997**). "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes," J. Exp. Psychol. Human, **23**, 914-927. doi: 10.1037/0096-1523.23.3.914

Green, C. S., Banai, K., Lu, Z. and Bavelier, D. (**2019**). "Perceptual Learning," *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. J. T. Wixted, John Wiley & Sons, Inc.**,** 1-47. doi: 10.1002/9781119170174.epcn217

Henshaw, H. and Ferguson, M. A. (**2013**). "Efficacy of individual computer-based auditory training for people with hearing loss: a systematic review of the evidence," PLoS One **8**, e62836. doi: 10.1371/journal.pone.0062836

Karawani, H., Bitan, T., Attias, J. and Banai, K. (**2016**). "Auditory perceptual learning in adults with and without age-related hearing loss," Front Psychol., **6**, 2066. doi: 10.3389/fpsyg.2015.02066

Karawani, H., Lavie, L. and Banai, K. (**2017**). "Short-term auditory learning in older and younger adults," Proc. ISAAR, **6**, 1-8.

Lavie, L., Attias, J. and Karni, A. (**2013**). "Semi-structured listening experience (listening training) in hearing aid fitting: influence on dichotic listening," Am. J. Audiol., **22**, 347-350. doi: 10.1044/1059-0889

Manheim, M., Lavie, L. and Banai, K. (**2018**). "Age, Hearing, and the Perceptual Learning of Rapid Speech," Trends Hear., **22**, 2331216518778651. doi: 10.1177/2331216518778651

Mattys, S. L., Davis, M. H., Bradlow, A. R. and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cogn. Process., **27**, 953-978. doi: 10.1080/01690965.2012.705006

Peelle, J. E. and Wingfield, A. (**2005**). "Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech," J. Exp. Psychol. Hum. Percept. Perform., **31**, 1315-1330. doi: 10.1037/0096-1523.31.6.1315

Pichora-Fuller, M. K. and Levitt, H. (**2012**). "Speech comprehension training and auditory and cognitive processing in older adults," Am. J. Audiol., **21**, 351-357. doi: 10.1044/1059-0889

Prior, A. and Bentin, S. (**2006**). "Differential integration efforts of mandatory and optional sentence constituents," Psychophysiology, **43**, 440-449. doi: 10.1111/j.1469-8986.2006.00426.x

Ronnberg, J., E. Holmer and Rudner, M. (**2019**). "Cognitive hearing science and ease of language understanding," Int. J. Audiol., **58**, 247-261. doi: 10.1080/14992027.2018.1551631

Rotman, T., Lavie, L., and Banai, K. (**2020**). "Rapid perceptual learning of time-compressed speech and the perception of natural fast speech in older adults with presbycusis," Proc. ISAAR, **7**, 93-100.

Samuel, A. G. and Kraljic, T. (**2009**). "Perceptual learning for speech," Atten. Percept. Psychophys., **71**, 1207-1218. doi: 10.3758/APP.71.6.1207

Saunders, G. H., Smith, S. L. , Chisolm, T. H. , Frederick, M. T. , McArdle, R. A. and Wilson, R. H. (**2016**). "A randomized control trial: Supplementing hearing aid use with listening and communication enhancement (LACE) auditory training," Ear Hearing, **37**, 381-396. doi: 0.1097/AUD.0000000000000283

Sweetow, R. W. and Sabes, J. H. (**2006**). "The need for and development of an adaptive Listening and Communication Enhancement (LACE) Program," J. Am. Acad. Audiol., **17**, 538-558. doi: 10.3766/jaaa.17.8.2

Verhelst, W. and Roelands, M. (**1993**). "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Minneapolis, MN, USA**,** 554-557.

Wright, B. A., Wilson, R. M. and Sabin, A. T. (**2010**). "Generalization lags behind learning on an auditory perceptual task," J. Neurosci. **30**, 11635-11639. doi: 10.1523/JNEUROSCI.1441-10.2010

The effect of conversational task on turn taking in dialogue

SAM WATSON[1], A. JOSEFINE MUNCH SØRENSEN[1], AND EWEN N. MACDONALD[1,*]

[1]*Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

In previous studies, several methods have been used to elicit conversation between talkers. Some involved participants solving a shared task (e.g., describing a map or finding differences between two near-identical pictures), while others have recorded more spontaneous dialogue (e.g., telephone calls). Since the goals of the talkers, and thus the definition of successful conversation, varies across these methods, it is thought likely that turn-taking behaviour will vary depending on how conversations are elicited. The present study investigated this by eliciting English conversations from 7 pairs of native-Danish talkers using two methods: solving a Diapix task and engaging in unguided "small talk". For each method, in both quiet and 70 dBA babble, two conversations were recorded for each pair. Overall, several differences in conversational behaviour were observed. When engaged in "small talk", participants spoke more rapidly, produced longer utterances, and replied more quickly than compared to when they were solving the Diapix task. These within-pair differences indicate that comparisons of behaviour across studies should also consider the method by which conversations were elicited.

INTRODUCTION

Recent studies investigating the effects of noise and hearing loss on interactive communication have suggested conversational effort could be assessed using measures of speech production and turn-taking behaviour (Beechey *et al.*, 2018; Hadley *et al.*, 2019; Sørensen *et al.*, 2020a,b). However, for some proposed metrics, the pattern of results vary substantially between studies (e.g., utterance duration increasing in noise for some studies vs. decreasing in others). A possible explanation for this could be differences across studies in the method used to elicit conversations.

When talkers switch turns (i.e., there is a transfer of who has the floor), the acoustic signals produced by each talker may partially overlap or be separated by a silent gap. The length of this interval (with a negative sign for overlap and positive for gap) is termed the floor-transfer offset (FTO). It has been hypothesized that in conditions where communication difficulty is increased, the FTO distribution should shift to the right when speech planning is delayed due to limited resources (e.g., Sørensen *et al.*, 2020a,b). In addition, if increased difficulty decreases the saliency of acoustic cues used to predict the timing of turn ends, then the FTO distribution should become more broad.

*Corresponding author: emcd@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

In the present study, we investigate the potential effect of task on several metrics of speech production and turn-taking behaviour when participants were engaged in both free conversation ("small talk") and when solving the Diapix task (Baker and Hazan, 2011), where the participants find differences between two almost identical pictures by describing them to each other.

METHOD

Fourteen normal-hearing native-Danish talkers were recruited for the study (mean age 23). They were divided in pairs (3 male-male, 3 male-female, and 1 female-female), and individuals in each pair did not know each other before the experiment. All participants reported normal hearing and were comfortable communicating in English. The procedure was approved by Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391), and all participants gave informed consent.

During the experiment, participants were seated in separate isolated sound booths and had no visual contact with each other. They spoke into Shure SM35 microphones that were connected with the GLXD15 wireless systems. The microphone signals were mixed using an RME Fireface 802 sound card and presented over Sennheiser HD650 headphones such that each individual heard his/her partner's voice at the same level as if he/she were standing 1m away.

Each pair produced two conversations in each of four conditions: Diapix task in quiet, Diapix task in noise, five minutes of "small talk" in quiet, and five minutes of "small talk" in noise. The noise used in this experiment was a 20-talker babble presented at 70 dBA and was the same as that used by Sørensen *et al.* (2020b). The conversations were recorded in two blocks. In each block, a conversation in each of the four conditions was collected, with the conditions randomized in order.

The recorded conversations were analyzed in the same manner as Sørensen *et al.* (2020a,b). For each talker, average speech levels, articulation rates, and utterance durations were measured. Here, utterances are defined as portions of speech that are separated by acoustic silences of more than 180 ms. In addition, two measures related to turn taking were recorded: FTOs and overlaps-within. As described above, the FTO is the interval between when one talker stopped and the other started speaking. However, in natural dialogue, turns do not always alternate between talkers. Sometimes the turn of one talker occurs completely within that of the other talker. We term these overlap-within, because the utterance is temporally overlapped within the turn of the other talker, who continues to maintain the floor.

RESULTS

Articulation rates, averaged across talkers, in each of the four conditions are plotted in the left panel of Figure 1. While no effect of noise was observed on articulation rate, talkers spoke more quickly during free conversation. A repeated measures ANOVA confirmed a significant main effect of task $[F(1,107) = 26.445, p < 0.001]$. No

significant main effect of noise $[F(1,107) = 0.171, p = 0.68]$ or significant interaction $[F(1,107) = 0.015, p = 0.902]$ was observed.

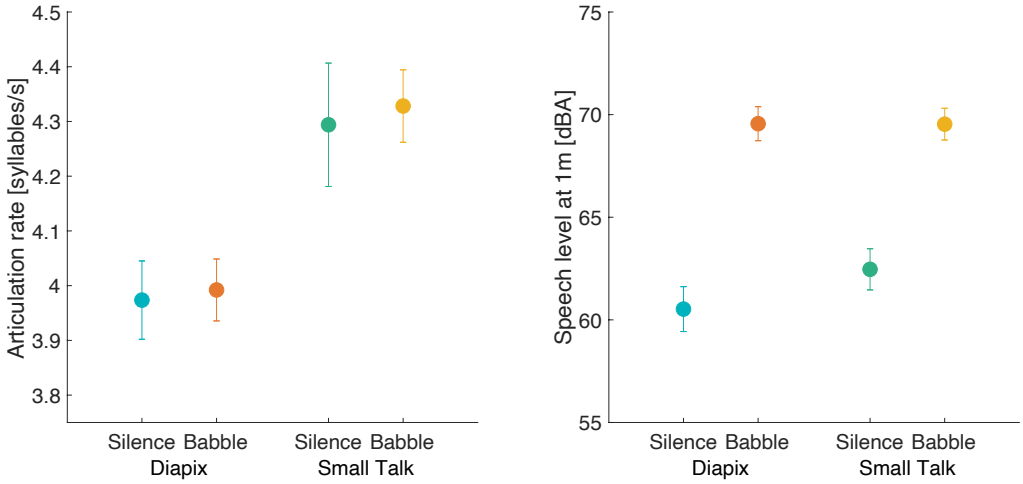


Fig. 1: Average articulation rate (left panel) and speech level (right panel) produced in the four combinations of task and noise. The bars indicate standard error.

Speech levels, averaged across talkers, in each of the four conditions, are plotted in the right panel of Figure 1. Consistent with the Lombard effect, talkers increased speech levels in noise. However, speech levels were similar in the two tasks. A repeated measures ANOVA confirmed a significant main effect of noise $[F(1,107) = 117.175, p < 0.001]$. No significant main effect of task $[F(1,107) = 1.656, p = 0.201]$ or significant interaction $[F(1,107) = 1.1738, p = 0.19]$ was observed.

For every instance where talkers switched turns, the floor-transfer offset (FTO) was calculated. The left panel of Figure 2 presents normalized FTO distributions for each of the four combinations of task and noise (i.e., results were normalized by the total number of floor transfers recorded in that condition after averaging across talker pairs and repetition). The median and interquartile range of these distributions, averaged across talker pairs and repetition, are plotted in the middle and right panels.

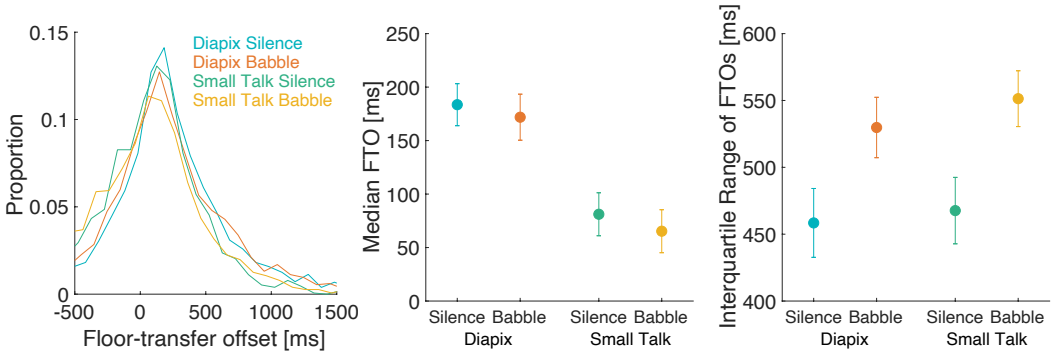


Fig. 2: Normalized distributions (left panel) of floor transfer offsets (FTO) along with the median (middle panel) and interquartile range (right panel) of these distributions for the four combinations of task and noise. The bars indicate standard error.

From Figure 2, it can be seen that task and noise had different effects on the distribution. The median FTO was shorter during small talk than during the Diapix task, but did not change in the presence of babble. A repeated measures ANOVA confirmed a significant main effect of task $[F(1,34.537) = 5.665, p < 0.001]$. No significant main effect of noise $[F(1,51) = 1.135, p = 0.442]$ or significant interaction $[F(1,51) = 0.014, p = 0.908]$ was observed. In contrast, while the FTO interquartile range was similar across tasks, it increased in noise. A repeated measures ANOVA confirmed a significant main effect of noise $[F(1,51) = 25.577, p < 0.001]$. No significant main effect of task $[F(1,51) = 1.008, p = 0.32]$ or significant interaction $[F(1,51) = 0.161, p = 0.689]$ was observed.

The distributions of utterance durations in the four conditions is plotted in the left panel of Figure 3. Note that here, utterances that were categorized as overlaps-within have been excluded. The median utterance duration increased both in babble and in small talk (see the right panel of Figure 3). A repeated measures ANOVA confirmed a significant main effects of noise $[F(1,51) = 20.396, p < 0.001]$ and task $[F(1,51) = 14.79, p < 0.001]$ and no significant interaction was observed $[F(1,51) = 0.024, p = 0.877]$.

The rate at which overlaps-within occurred increased in small talk (see Figure 4). A repeated measures ANOVA confirmed a significant main effects of task $[F(1,51) = 10.617, p < 0.01]$. No significant main effect of noise $[F(1,51) = 1.175, p = 0.28]$ or interaction $[F(1,51) = 0.035, p = 0.852]$ was observed.

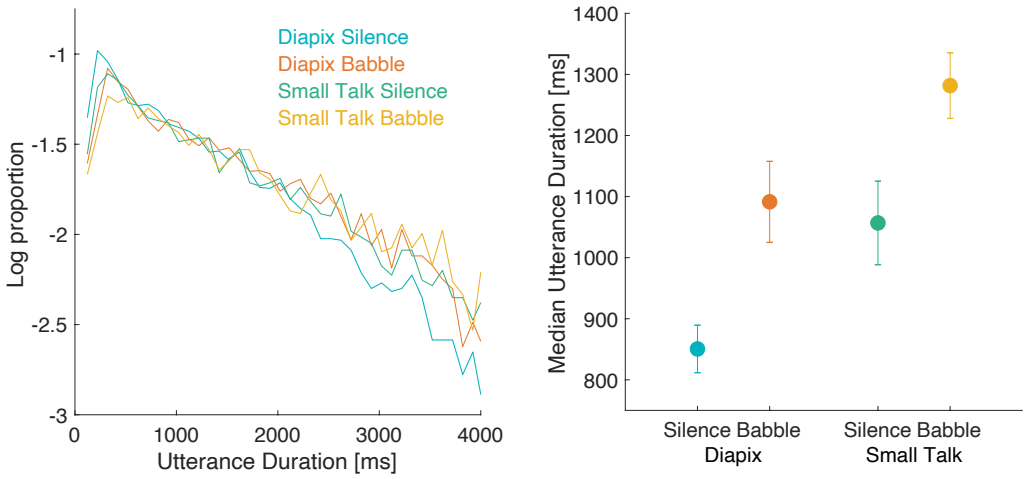


Fig. 3: Normalized distributions of utterance duration (left panel) and median utterance duration of these distributions (right panel) for the four combinations of task and noise. The bars indicate standard error.

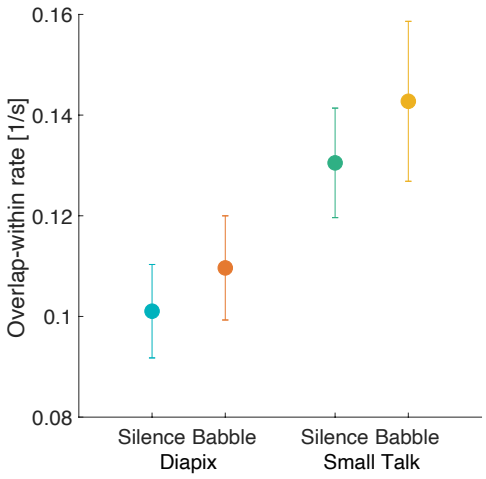


Fig. 4: Mean rate of occurrence of overlaps-within (i.e., turns from one talker that occur completely within a turn of the other talker) for the four combinations of task and noise. Note that the rate has been normalized by the total phonation time rather than duration of the conversation. The bars indicate standard error.

DISCUSSION

The purpose of the present study was to investigate if the method of eliciting dialogue between two talkers affected various measures of speech production and turn-taking behaviour. Over the course of the study, pairs of talkers, who were not familiar with each other prior to the experiment, produced eight conversations in four different conditions. In half the conditions, talkers were instructed to participate in small talk (i.e., a free conversation). In the other half, they conducted a Diapix task, where they had to find differences between two almost identical pictures. Half of the conversations were conducted in quiet, the other half were conducted in a background of multi-talker babble noise. Overall, changes in speech production and turn-taking behaviour were observed across the four conditions. Further, the pattern of results indicated that while both background noise and conversational task influence dialogue behaviour, they have different effects.

Speech production

Consistent with the Lombard effect, talkers increased speech levels in the presence of noise, but the levels were not influenced by the task. In contrast, talkers spoke more rapidly when participating in free conversation than when solving the Diapix task. However, their speech rate was not influenced by the noise.

The influence of noise on articulation rate in previous studies of conversation has been inconsistent. While the same Diapix task was used in Sørensen *et al.* (2020a) and Sørensen *et al.* (2020b), the normal-hearing talker pairs in Sørensen *et al.* (2020a) increased their rate of speech in noise, whereas the normal-hearing talkers in Sørensen *et al.* (2020b), who conversed with hearing-impaired talkers, decreased their rate of speech when talking in noise, indicating different behaviour depending on conversational partner.

Floor-transfer offset (FTO)

It has been hypothesized that in conditions where communication difficulty is increased, the FTO distribution should shift to the right when speech planning is delayed due to limited resources (e.g., Sørensen *et al.*, 2020a,b). Further, if increased difficulty decreases the saliency of acoustic cues used to predict the timing of turn ends, then the FTO distribution should become more broad.

In the present study, the median FTO during the Diapix task was longer than during free conversation. It is tempting to conclude that conducting the Diapix task is more challenging than holding free conversation. However, no change was observed between the quiet and noise conditions. If it was conversational effort that was responsible for the longer median FTO observed when the Diapix task was conducted in quiet, then one would expect that adding noise would further increase the difficulty and result in an even longer median FTO. However, this was not observed.

One possible explanation for these results is that participants are communicating differently between the conversational tasks. To solve the Diapix task quickly may require more accurate information transmission than is needed in free conversation. Thus, talkers might adjust behaviour and target a longer FTO to reduce the number of speech overlaps. Another possible explanation is that solving the Diapix task may involve more question-answer constructions than free conversation, some of which may require a visual search to be completed (e.g., "Do you see a red ball?"), delaying the response from a talker.

While the interquartile range of the FTO distributions increased in noise, there were no differences across conversational tasks. Since the FTO distributions for free conversation and solving the Diapix task were similar in breadth, these results suggest that the ability to predict the timing of turn ends was not influenced by task. The broader FTO distributions observed in the presence of noise are consistent with a reduction in ability to predict the timing of turn ends, which is likely due to a reduction in the saliency of acoustic cues used to make the predictions.

Utterance duration

The median utterance duration was observed to be longer during free conversation and also increased in the presence of noise. Sørensen *et al.* (2020a) also observed increased utterance duration in noise and suggested that this was due to talkers holding their turn longer, providing more time for interlocutors to conduct speech planning and speech understanding.

In that study, the slopes of the distributions of utterance duration were different in quiet vs. noise. However, in the present study, the differences in median utterance duration across conditions appear to be driven mainly by differences in the frequency of very short utterances (i.e, approximately 500 ms or shorter, which corresponds to 1-2 syllables). For utterance durations ranging between 750-2000ms, the slopes of the distributions are similar across the four conditions. This is consistent with a possibly increase in the number of simple short responses during the Diapix task (e.g., "Yes", "Uh...","Yep", "Huh...")

Overlap-within rate

In natural dialogue, turns do not always alternate between talkers. Sometimes the turn of one talker occurs completely within that of the other talker (i.e., it is overlapped within the turn of the other talker who continues to maintain the floor).

In the present study, overlaps-within occurred more frequently during small talk than when conducting the Diapix task. One possible explanation for this is a difference in the conversational goals between small talk and solving a Diapix task. As mentioned above, to solve a Diapix task rapidly, participants should aim to maximize the rate of information transfer. As a consequence, they may attempt to reduce the rate at which

they interrupt their partner. In contrast, during small talk, the quality of the social interaction may be prioritized over the rate at which information is transmitted.

However, for free conversation, both longer utterance durations and a shift of the FTO distribution to the left were observed. Thus, it is also possible that the increase in the rate of overlaps-within are a natural consequence of these changes rather than a change in conversational goals.

SUMMARY

When participating in small talk compared to the Diapix task, talkers spoke more rapidly, produced longer utterances, produced overlaps-within more frequently, and when a turn switched, the floor-transfer offset was shorter. When holding conversation in noise, talkers increased the level of voice, produced longer utterances, and the distribution of floor-transfer offsets was more broad.

ACKNOWLEDGEMENT

Parts of this study were supported by the William Demant Foundation (16-3968).

REFERENCES

Baker, R. and Hazan, V. (**2011**). "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs," Behav. Res. Methods, **43**, 761-770. doi: 10.3758/s13428-011-0075-y

Beechey, T., Buchholz, J. M., and Keidser, G. (**2018**). "Measuring communication difficulty through effortful speech production during conversation," Speech Commun., **100**, 18-29. doi: 10.1016/j.specom.2018.04.007

Hadley, L. V., Brimijoin, W. O., and Whitmer, W. M. (**2019**). "Speech, movement, and gaze behaviours during dyadic conversation in noise," Sci. Rep., **9**, 10451. doi: 10.1038/s41598-019-46416-0

Sørensen, A. J. M., Fereczkowski., M, and MacDonald, E. N. (**2020a**). "Effects of noise and L2 on the timing of turn taking in conversation," Proc. ISAAR, **7**, 85-92.

Sørensen, A. J. M., Lunner, T., and MacDonald, E. N. (**2020b**). "Timing of turn taking between normal-hearing and hearing-impaired interlocutors," Proc. ISAAR, **7**, 37-44.

Duration threshold for identifying speech samples for different phonemes

HENDRIK HUSSTEDT[1,*], SIMONE WOLLERMANN[1,2], DANIEL BANK[1,2], MARIO SCHINNERL[1,2], MARLITT FRENZ[1] AND JÜRGEN TCHORZ[3]

[1] *German Institute of Hearing Aids, Lübeck, Germany*

[2] *Universität Lübeck, Lübeck, Germany*

[3] *Technische Hochschule Lübeck, Lübeck, Germany*

The identification or classification of acoustic objects is important to decide in which way a sound needs to be interpreted and to rate its importance or relevance. In recent studies, it has been shown that the minimal duration of a sound, which is required for a correct identification, could be a useful audiological parameter, e.g. providing information about the hearing ability of a person. In this work, we want to investigate which cues are used by humans to classify a sound correctly as speech. For this purpose, the duration thresholds for the identification of speech samples starting with different phonemes are analyzed for elderly listeners with normal and impaired hearing. To this end, a two-alternative forced choice (2-AFC) method was used, where, as an alternative to speech, a noise signal with a matched frequency spectrum was presented. In contrast to previous studies, there were no frequency cues available and we found no correlation to the pure tone average (PTA) or speech understanding in noise. As one main conclusion, the results suggest that humans primarily exploit the temporal envelope (ENV) rather than the temporal fine structure (TFS) for the identification of short speech samples above hearing threshold and without frequency cues.

INTRODUCTION

In daily life, the identification or classification of sounds enables the construction of an acoustic scenery with certain objects in our brain. It allows us to decide in which way a sound needs to be interpreted, e.g., the sound of a car has other features than a speech signal. Moreover, in certain listening situations, the identification of sounds is required to rate the importance of the corresponding acoustic object, e.g., speech in a classroom, and car sounds in a traffic situation are of high importance. Consequently, the ability to identify sounds, which is sometimes referred to as auditory gnosis, is an important ability of our acoustic perception (Akelaitis, 1944; Hirsh and Watson, 1996). An impairment of this ability can have serious consequences. If cognitive disabilities are the reason, we speak about auditory agnosia (Pietro *et al.*, 2016). Besides, from an audiological point of view, it is also interesting to investigate how a limited hearing ability can impair the identification of sounds. As depicted by

*Corresponding author: h.husstedt@dhi-online.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Ballas (1993), Gygi *et al.* (2004), and McDermott and Simoncelli (2011), frequency and temporal information are used for the identification, which are both impaired for people with hearing loss (Moore, 1984). For an audiological evaluation, a test paradigm for the identification task has to be defined. To this end, the minimal duration of a sound required for a correct identification can be measured, which was applied in recent studies (Gray, 1942; Bank *et al.*, 2019; Budathoki *et al.*, 2019). These studies report duration thresholds for speech in the range of 20-40 ms, which is remarkably short compared to technical algorithms, e.g., used for the automatic selection of hearing aid programs (Husstedt *et al.*, 2018). Both studies used an alternative forced choice (AFC) procedure with multiple sound samples of different classes, e.g., speech, music, or animal sounds. One drawback of this approach is that the results strongly depend on the sound samples and sound classes chosen (Husstedt *et al.*, 2019). Therefore, another study used the method of adjustment where single sound files could be analyzed independently. However, this method has a high variation due to individual ratings (Husstedt *et al.*, 2019). In this work, we present another measurement procedure to determine the minimum duration of a speech sound required for a correct identification. For this purpose, a 2-AFC task was defined where speech has to be correctly distinguished from a noise signal with matched frequency spectrum. The goal is to investigate which temporal cues are used by humans to classify a sound correctly as speech, and how the perception of those cues is impaired by a hearing loss. To this end, the duration thresholds for the correct identification of speech samples starting with different phonemes are analyzed for a population of 30 people with an age between 60-85 years and a pure tone average (PTA) between 0-80 dB.

MATERIAL AND METHODS

Sound samples

As sound samples, ten German words spoken by a male speaker have been recorded for the test. Different initial phonemes have been selected to investigate their influence on the duration threshold (see Tab. 1).

Abteilung	**Ak**tivitäten	**Al**ternativen	**An**gebote	**Ar**beitsplätze
Bakterien	**Ka**ndidaten	**La**ndwirtschaft	**Na**chmittage	**Ra**hmenbedingungen

Table 1: Ten German words used for the tests spoken by one male speaker.

Five words start with the consonants “B”, “K”, “L”, “N”, and “R” followed by the same vowel “A”. Thus, if there is any effect of the second phoneme, this is almost constant for all of these five words. In contrast, the other five words start with the vowel ”A” followed by the five consonants mentioned so that the effect of the second phoneme can be investigated. For the cutting of the words out of the recorded data, the moving root mean square (RMS) with a window size of 10 ms has been computed,

and the point where the sound pressure level first exceeds a level of 45 dB was set as the starting point. For all words, the starting point is kept constant and the end point is varied to provide sound samples with different durations. Moreover, for each duration of a speech sample, an individual noise signal was generated by randomizing the phase information. Thus, the absolute value of the frequency spectrum is equal for both the speech and the corresponding noise sample. Before the presentation a fade in and fade out of 0.5 ms was applied to all signals. Furthermore, since previous studies indicate that a logarithmic representation of the duration is preferable, the duration of the samples is varied in steps of 1 dB and the absolute duration is provided in Decibel relative to 1 ms (e.g., as illustrated in Fig. 1, 36 dB rel 1 ms corresponds to approx. 63 ms).

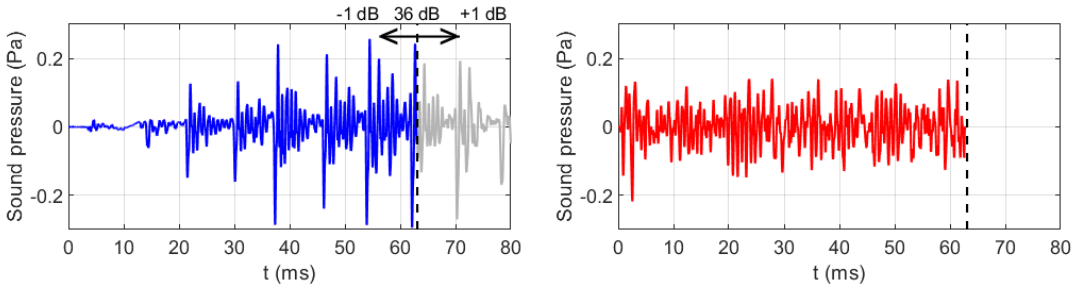


Fig. 1: Visualization of a speech sample (left) and the corresponding noise signal (right). For each duration of the speech sample, the noise signal was generated by randomizing the phase information so that the absolute value of the frequency spectrum is preserved.

Measurement procedure

A touch screen was used for the interaction with the test persons, and the sounds were presented with Sennheiser HDA 280 headphones. First, a dialog was shown where the test persons could listen to different sound samples of different duration, and could adjust the sound pressure level for comfortable loudness. For the test, this selected sound pressure level was applied to each speech and noise sample.

After adjusting the presentation level, the 2-AFC-procedure started where the speech and noise signal were presented in randomized order, and the test person needed to answer which signal was speech (see Fig. 2 a, b). The duration of the samples was adaptively changed with the weighted up-down method according to Kaernbach (1991). The starting point was always 46 dB rel 1 ms (approx. 200 ms), and the end of the test was reached after 12 reversal points (see Fig. 2 c). Until the fourth reversal point, step sizes of 2 dB down and 6 dB up, and afterwards 1 dB down and 3 dB up were used. As result, the average of the last four reversal points was computed and saved as duration threshold.

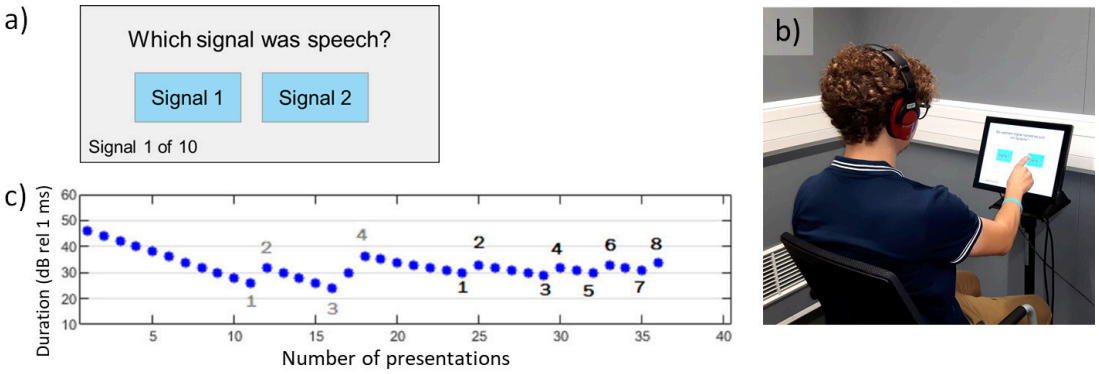


Fig. 2: Visualization of the 2-AFC procedure applied: a) Dialog presented on the touch screen; b) Photograph of a test person typing in an answer; c) Example result for the adaptively changed duration of the samples with the weighted up-down method according to Kaernbach (1991).

Study design

Overall, 30 elderly people with an age between 60-85 years and a pure tone average (PTA) between 0-80 dB were tested. We have defined the PTA as the average of the hearing thresholds for the frequencies 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. To distinguish effects of hearing loss from age related effects, a population was selected, which shows no significant correlation ($p = 0.26$) between age and hearing loss (see Fig. 3).

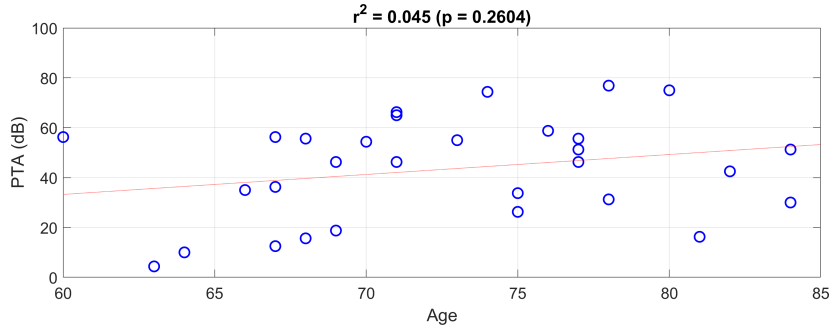


Fig. 3: Scatter plot of the pure tone average (PTA) against age of the 30 test persons.

During the appointment various audiological assessments were performed in the following order: otoscopy, pure tone audiometry, Montreal cognitive assessment (MOCA, Nasreddine *et al.*, 2005), speech in noise with the German sentence test GÖSA (Kollmeier and Wesselkamp, 1997), and finally the measurement of the duration thresholds as depicted in Fig. 2 for the ten German words listed in Tab. 1. The order of the words was chosen according to a Latin square so that after ten test

persons every word has been presented at each position. Furthermore, the order of the columns and rows of the Latin square was randomized for every ten test persons so that the order of the words and their position relative to each other was also randomized.

RESULTS

Initial phonemes

In Fig. 4, the duration thresholds for all words are depicted in a boxplot, and significant differences are indicated with stars. Since the Lilliefors test suggests a normal distribution only for some of the words, the Friedman test including the Tukey-Kramer method was used for a comparison. There are no significant differences between all of the words starting with the same phoneme "A". On the contrary, various significant differences can be observed for the words starting with different phonemes.

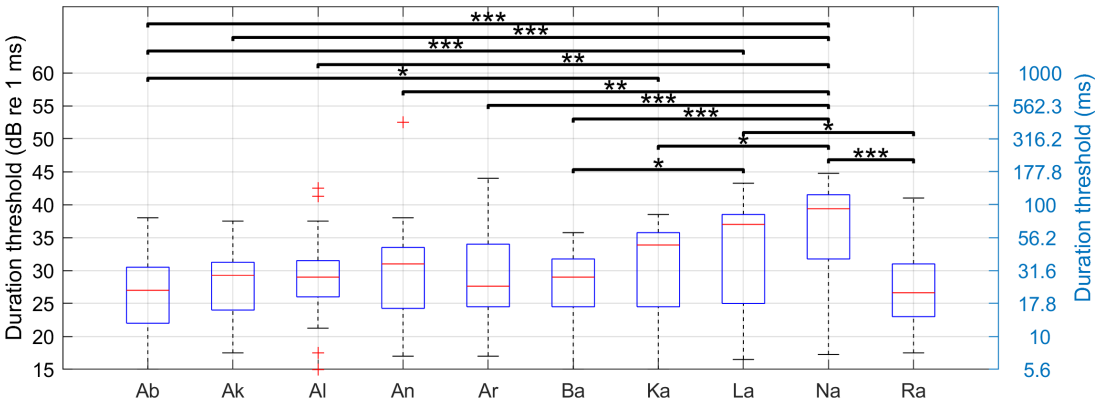


Fig. 4: Boxplot of the duration threshold for all words. For the comparison, the Friedman test including the Tukey-Kramer method was performed. The stars indicate significant differences where "*" corresponds to $p < 0.05$, "**" to $p < 0.01$, and "***" to $p < 0.001$.

Duration thresholds against PTA

In Fig. 5, the duration thresholds of all words are depicted in scatter plots against PTA. For each plot, the Spearman correlation coefficient was computed and it was tested for significance. Since ten hypotheses are tested, the level of significance shall be corrected according to the Bonferroni method ($\tilde{\alpha} = \alpha/10 = 0.05/10$). With this correction, there are no significant correlations between the duration threshold of a word and the PTA. The same holds, if we average the duration threshold over all words.

Speech in noise and cognition

As expected, the results of the speech in noise test (GÖSA) show a high correlation to the PTA whereas there is no correlation to the duration thresholds. Furthermore,

we found a significant correlation between the results of the MOCA and the duration thresholds.

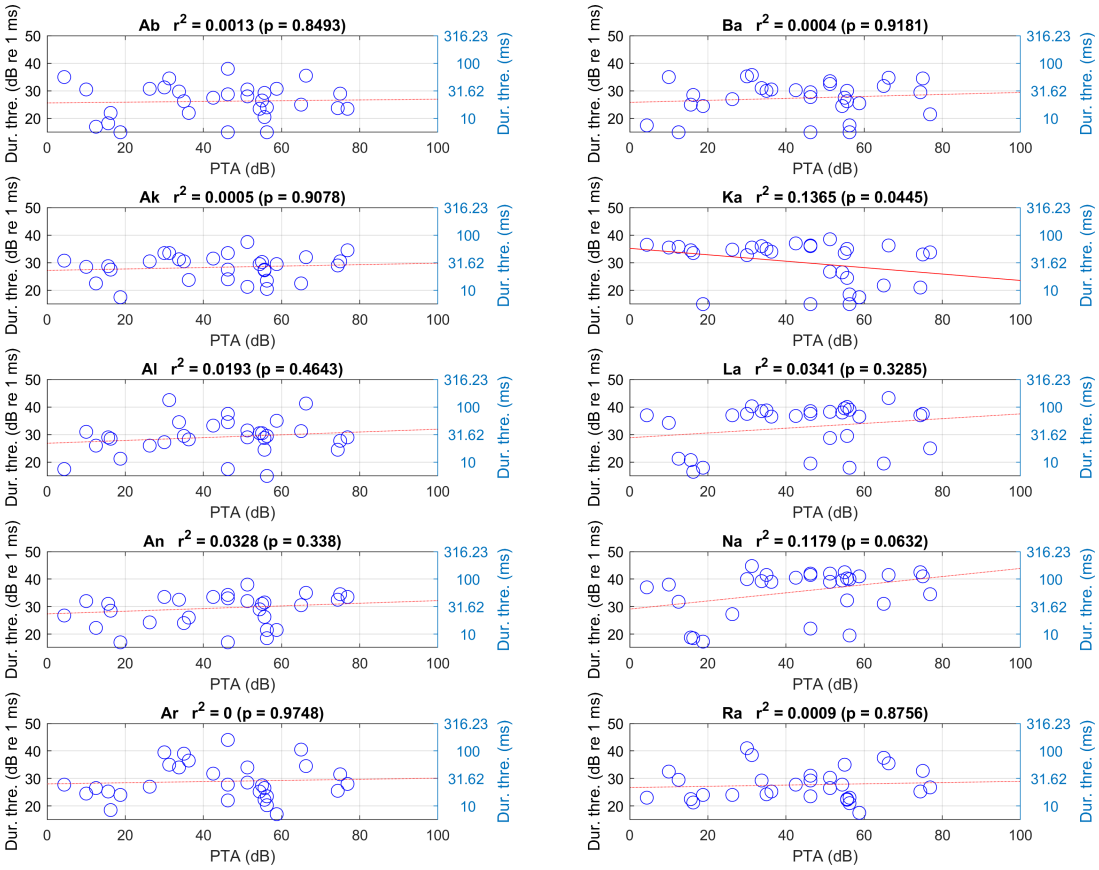


Fig. 5: Scatter plot of the duration threshold against pure tone average (PTA) for all words. Since the α-value is adapted, the p-values indicated are without Bonferroni correction.

DISCUSSION

Initial phonemes

The results clearly demonstrate that different initial phonemes have an influence on the duration threshold. The median is lowest for “Ra” with 21.4 ms and highest for “Na” with 93.1 ms. There are no significant differences between those words starting with the same phoneme “A” followed by different phonemes. This indicates that the second phoneme has a minor or negligible influence on the duration thresholds. A subjective evaluation of the duration of the initial phonemes results in values of approx. 77-120 ms for all words starting with “A”. Since this is longer than the duration thresholds measured (see Fig. 4), this a reasonable explanation for the minor impact of the second phoneme. Except for the word starting with “Ba”, the same holds for all other words

considered. Consequently, in most cases, the initial phoneme is sufficient to correctly detect the signals as speech.

Hearing ability

One main research question of this study is how the hearing ability affects the identification of short speech samples. In contrast to previous studies (Bank *et al.*, 2019; Budathoki *et al.*, 2019), we found no correlation between PTA and duration threshold. As one main difference, in our study there were no frequency cues available due to the matched frequency characteristic of the speech and noise samples. Consequently, one could conclude that a limited frequency resolution of hearing impaired listeners has led to higher duration thresholds in Bank *et al.* (2019) and Budathoki *et al.* (2019).

Without frequency cues, people have to only rely on temporal cues such as the temporal envelope (ENV) and the temporal fine structure (TFS). The results of Lorenzi *et al.* (2006) indicate that hearing impaired people can exploit the information provided by the ENV in the same way as people without hearing loss whereas hearing impaired people show a greatly reduced ability to use information provided by the TFS. Since we see no effect of the hearing loss, the results suggest that the ENV rather than the TFS is used for the identification of short speech samples, if frequency cues are not available and the samples are presented above hearing threshold.

CONCLUSION

In this work, the duration thresholds for the identification of short speech samples starting with different phonemes are analyzed for elderly listeners with normal and impaired hearing. To this end, a 2-AFC method was used where as alternative to speech a noise signal with matched frequency spectrum was presented. The results show that mostly within the first phoneme, people can correctly identify a signal as speech. Moreover, the initial phoneme has a significant influence on the time required for a correct identification. In contrast to previous studies, there were no frequency cues available and we found no correlation to the pure tone average (PTA) or speech understanding in noise. As one main conclusion, the results suggest that humans primarily exploit the temporal envelope (ENV) rather than the temporal fine structure (TFS) for the identification of short speech samples, which include no frequency cues and which are presented above hearing threshold.

REFERENCES

Akelaitis, A. J. (**1944**). "A Study of Gnosis, Praxis and Language Following Section of the Corpus Callosum and Anterior Commissure," J. Neurosurg., **1(2)**, 94-102. doi: 10.3171/jns.1944.1.2.0094

Ballas, J. A. (**1993**). "Common factors in the identification of an assortment of brief everyday sounds," J. Exp. Psychol. Hum. Percept. Perform., **19(2)**, 250-267. doi: 10.1037/0096-1523.19.2.250

Bank, D., Schinnerl, M., Frenz, M., Gassenmeyer, F., and Husstedt, H. (**2019**). "Duration Threshold for Identifying Sound Samples of Elderly Hearing Impaired," The Student Conference of the BioMedTec Science Campus, Lübeck, Mar., 2019

Budathoki , D., Tchorz, J., and O'Beirne , G. (**2019**). "Duration Thresholds for Identifying Different Sound Types," 22. DGA Jahrestagung, Heidelberg, Germany, 2019.

Gray, G. W. (**1942**). "Phonemic Microtomy: The Minimal Duration of Perceptible Speech Sounds," Speech Monogr., **9(1)**, 75-90. doi: 10.1080/03637754209390064

Gygi, B., Kidd, G. R., and Watson, C. S. (**2004**). "Spectral-temporal factors in the identification of environmental sounds," J. Acoust. Soc. Am., **115(3)**, 1252-1265. doi: 10.1121/1.1635840

Hirsh, I. J., and Watson, C. S. (**1996**). "Auditory psychophysics and perception," Ann. Rev. Psychol., **47(1)**, 461-484. DOI: 10.1146/annurev.psych.47.1.461

Husstedt, H., Bank, D., and Schinnerl, M. (**2019**). "Comparison of Two Procedures to Measure the Duration Threshold for Identifying Sound Samples," 22. DGA Jahrestagung, Heidelberg, Germany, 2019.

Husstedt, H., Wollermann, S., and Tchorz, J. (**2018**). "Analysis of the Transition of the Automatic Selection of Hearing Aid Programs," 45th Erlanger Kolloquium, Erlangen, Germany, 2018.

Kaernbach , C. (**1991**). "Simple adaptive testing with the weighted up-down method," Percept. Psychophys., **49(3)**, 227-229. doi: 10.3758/BF03214307

Kollmeier, B., and Wesselkamp, M. (**1997**). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," J. Acoust. Soc. Am., **104(2)**, 2412-2421. doi: 10.1121/1.419624

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U. S. A., **103(49)**, 18866-18869. doi: 10.1073/pnas.0607364103

McDermott, J. H., and Simoncelli, E. P. (**2011**). "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis," Neuron, **71(5)**, 926-940. doi: 10.1016/j.neuron.2011.06.032

Moore, C. J. M. (**1984**). "Frequency selectivity and temporal resolution in normal and hearing-impaired listeners," Br. J. Audiol., **19(3)**, 189-201. doi: 10.3109/03005368509078973

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (**2005**). "The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment," J. Am. Geriatr. Soc., **53**, 695-699. doi: 10.1111/j.1532-5415.2005.53221.x

Pietro, M., Laganaro, and M., Schnider, A. (**2016**). "Auditory agnosia," in *Neuropsychological Research: A Review*, Edited by P. Marien and J. Abutalebi (Psychology Press), chap. 15.

Assessing the impact of fundamental frequency on speech intelligibility in competing-talker scenarios

PAOLO A. MESIANO[1,*], JOHANNES ZAAR[1], LARS BRAMSLØW[2], NIELS H. PONTOPPIDAN[2], AND TORSTEN DAU[1]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Augmented Hearing, Eriksholm Research Centre, 3070 Snekkersten, Denmark*

When only monaural cues are available in competing-talker scenarios, normal-hearing (NH) listeners are able to identify and understand the target speech while hearing-impaired listeners often experience difficulties. A good understanding of the role of monaural cues in speech segregation is therefore essential for developing hearing-aid compensation strategies. Earlier studies with NH listeners showed that differences in fundamental frequency (ΔF_0) between the target talker and one interfering talker can facilitate the segregation of the speech signals. However, most of these studies used speech materials that bear little resemblance with everyday speech. Furthermore, the F_0 was either defined by talker sex or measured as a talker-specific average, thus ignoring the significant F_0 variability across sentences. The present study instead used everyday-speech type sentences from the Danish Hearing in Noise Test (HINT) and employed a more accurate method for assessing the impact of F_0 on intelligibility for NH listeners. Compared to previous studies, the overall effect of ΔF_0 was found to be smaller and it was hypothesised that the previously employed speech materials might have enhanced the effect of ΔF_0 beyond its real-life importance.

INTRODUCTION

When several talkers speak simultaneously, it can be challenging to identify and understand one specific target-speech signal. In such situations, usually referred to as competing-talker scenarios, the healthy auditory system shows exceptional abilities of segregating the target speech from the interfering speech by making use of several auditory cues. For example, binaural cues are known to be beneficial for separating signals arriving from different directions. However, in situations where binaural cues are not present or unreliable, the auditory system must rely on monaural cues only. In fact, it has been demonstrated that, in such conditions, normal-hearing (NH) listeners are still able to identify and understand the target speech, whereas hearing-impaired (HI) listeners typically experience substantial difficulties (Bramsløw *et al.*, 2015). The fundamental frequency (F_0) is a monaural auditory cue with a strong impact on competing-talker scenarios: when the target speech signal and a competing speech

*Corresponding author: pamesi@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

signal differ in F_0, the perception of the target is typically facilitated (e.g., Brokx and Nooteboom, 1982, Darwin *et al.*, 2003, Summers and Leek, 1998, Assmann and Summerfield, 1990, Assmann, 1999). Several approaches have been used to investigate the role of F_0 differences between two competing talkers. However, the employed experimental scenarios have typically been different from everyday-life listening situations.

For example, in Brokx and Nooteboom (1982) and Darwin *et al.* (2003), the F_0 of the competing sentences was assessed as a talker-specific average, thus ignoring the considerable F_0 variability across sentences spoken by a given talker. Brokx and Nooteboom (1982) generated the F_0-separated condition by pairing target speech from a male talker with interfering speech from the same talker who was asked to imitate the (higher) pitch of a female voice. The resulting interfering signal had a higher F_0 on average, but the F_0 contour (i.e., F_0 as a function of time) showed large variations, often crossing the F_0 contour of the target. To obtain a more controlled F_0 difference between competing voices, Summers and Leek (1998) employed monotonized F_0 contours (constant F_0 over time). The advantage of this approach is a perfectly controlled F_0 difference for each pair of sentences, but the monotonized voice may sound unnatural to the listener. In Assmann (1999), pairs of sentences with naturally-varying F_0 contours were used and F_0 separations were generated as the difference between the across-time average F_0 of each sentence, rather than based on the average of the talker, obtaining a more accurate control of F_0 separation. However, Assmann (1999) aligned the competing sentences at their offsets, potentially introducing a strong cue due to the onset differences.

Furthermore, the speech corpora have typically been chosen to maximize the influence of the F_0 separation. Darwin *et al.* (2003) used competing sentences from the coordinate-response measure (CRM), a speech corpus comprised of time-aligned closed-set sentences with a pre-defined structure and two scoring keywords per sentence. They observed a significant improvement in speech intelligibility when two competing sentences differed in F_0 by more than two semitones and reported performance improvements of more than 20% for a 9-semitone separation. However, these strong benefits induced by F_0 differences might be exaggerated in relation to real-life speech, due to the high degree of word alignment in the CRM corpus.

The present study aimed to overcome the above mentioned limitations by (i) employing a more accurate method for measuring and generating the F_0 separation between competing sentences, taking into account the variability across sentences for each talker, and by (ii) using a more realistic and less constrained speech corpus as compared to the mentioned reference studies.

METHOD

The Danish Hearing in Noise Test (HINT) (Nielsen and Dau, 2011) was used to generate the experimental stimuli. The HINT speech corpus consists of 200 open-set five-word everyday-type natural sentences split into ten phonetically-balanced

lists. Recordings of the speech material from twelve different talkers (six males and six females) were used. First, F_0 contours were extracted with the software Praat (Boersma, 1993): the instantaneous F_0 was estimated in 10-ms time steps, within a frequency range between 30 and 550 Hz. Then, the across-time median and standard deviation of the F_0 contour for each sentence was computed. The median F_0 of a talker can vary significantly across the speech corpus, by up to four semitones. Therefore, the F_0 separation between paired sentences was measured as the difference between their median F_0 values, regardless of the average F_0 of the specific talker.

The stimuli were generated by pairing sentences spoken by the same talker, taken from different lists, presented simultaneously and aligned at their onsets. The sentences were processed in Praat to obtain a difference in median F_0 (ΔF_0) of 0, 3, 6 or 12 semitones. Each value of the F_0 contour was multiplied by a scaling factor $s > 0$. This approach preserved the natural increase in frequency range observed for increasing median F_0. To avoid shifting sentences to extreme and unnatural F_0 values, the desired ΔF_0 difference was split across the two sentences in each pair, as indicated in Table 1, where the F_0 shifts in semitones refer to the separation from the talker's average F_0. In each sentence pair, the larger F_0 shift was applied towards higher frequencies ($s > 1$) if the average F_0 of the talker was above the average F_0 of the entire speech corpus, and towards lower frequencies ($s < 1$) otherwise. Figure 1 shows an example of a sentence pair with different ΔF_0s, indicated in the different panels.

ΔF_0	Talker F_0 > overall F_0		Talker F_0 < overall F_0	
	F_0 shift (s>1)	F_0 shift (s<1)	F_0 shift (s>1)	F_0 shift (s<1)
0	0	0	0	0
3	1	-2	2	-1
6	2	-4	4	-2
12	4	-8	8	-4

Table 1: F_0 shifts applied to each sentence in a pair to obtain the desired ΔF_0.

The stimuli were presented using the competing-voices test (CVT) framework developed by Bramsløw *et al.* (2019), where the listeners were provided with the first word of the target sentence on a screen prior to the stimulus playback (text pre-cue). After playback, the listeners were asked to repeat as many words as possible from the target sentence. The target sentence was presented at an average sound pressure level (SPL) of 65 dB with level roving of $\pm$5 dB. Target-to-masker ratios (TMRs) of -12, -8, -4, 0, 4 dB were combined with the four ΔF_0 values to produce a total of 20 experimental conditions. Each condition was tested using 20 sentence pairs. The performance in the experiment was measured as the proportion of correctly repeated words in the target sentence, averaged over the 20 sentence pairs presented in each experimental condition.

To avoid any effect of presentation order or sentence repetition in the group results, the test conditions (ΔF_0 and TMR) were balanced across listeners using a Latin square

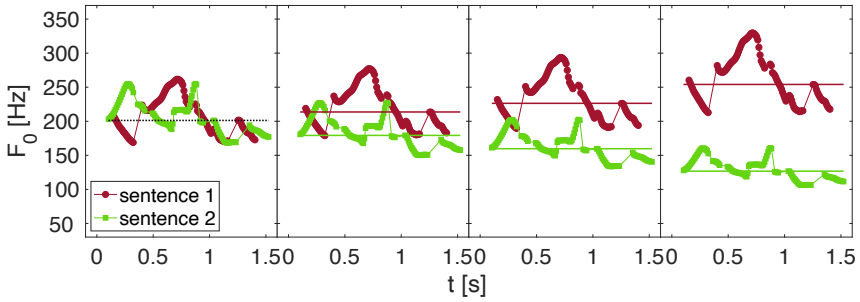


Fig. 1: Example of F_0 contour processing for a pair of HINT sentences. Panels from left to right show ΔF_0 of 0, 3, 6 and 12 semitones. The median F_0 of each sentence is displayed as a straight line, except for $\Delta F_0 = 0$ semitones (left panel) where the median F_0 of both sentences coincide with the talker's median F_0, shown as a dashed black line.

design while sentence-list and talker were randomized across conditions. In each pair, the target sentence was assigned randomly to either the sentence with higher or lower median F_0. Given the limited number of sentences in the HINT speech corpus, sentences have been repeated during the experiment, either as target or masker. However, the process of generating sentence pairs was randomized in a way that a sentence was presented a second time as late as possible and never within a given experimental condition.

The stimuli were free-field equalized and presented diotically over Sennheiser HDA200 headphones to the listeners seated in a sound-proof booth. Fifteen young native speakers of Danish with NH (pure-tone thresholds below 20 dB hearing level between 125 Hz and 8 kHz) were tested in the experiment.

RESULTS

The left panel of Figure 2 shows average group results, with the proportion of correct words displayed as a function of TMR. Each function represents results obtained for a particular ΔF_0 condition. Overall, speech intelligibility was found to improve with increasing TMR for all ΔF_0 values. The strongest effect of ΔF_0 was observed for intermediate TMR values (at -8 and -4 dB), with a maximum effect at TMR=-8 dB where the performance improved with increasing ΔF_0. At this TMR, the proportion of correct words increased by 15% from $\Delta F_0 = 0$ semitones to $\Delta F_0 = 12$ semitones. At the limits of the TMR range tested, only a minor effect of ΔF_0 was measured.

For comparison, the data from Darwin *et al.* (2003) are shown in the right panel of Figure 2. The range of TMR values in common between their experiment and the present study is indicated by the grey area in the two panels. Darwin *et al.* (2003) observed an overall stronger effect of ΔF_0 between target and interfering sentences, with the largest speech intelligibility improvement of about 30% at TMR = -3 dB for a separation of 12 semitones. In their study, the F_0 separation was effective for TMRs

ranging from -6 dB to 3 dB, decreasing for increasing TMR, and becoming negligible at the higher applied TMRs where a ceiling effect was observed.

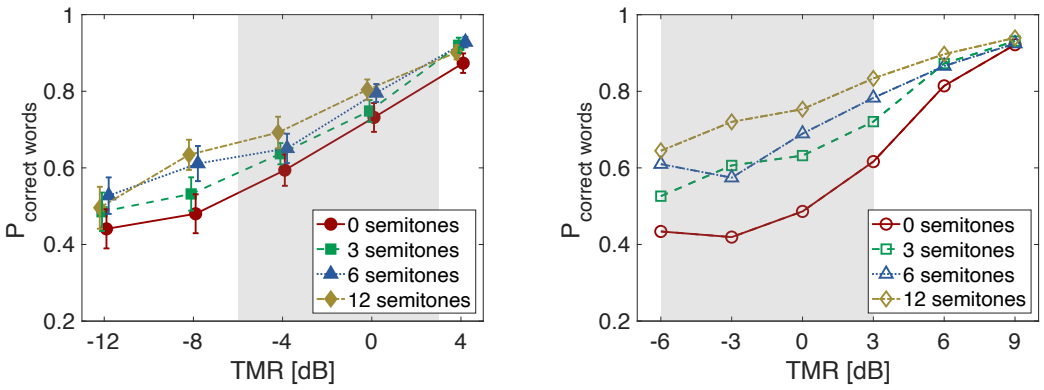


Fig. 2: Left panel: Group average proportion of correct words as a function of TMR. Different ΔF_0 conditions are shown as different curves. Error bars represent standard errors across listeners. Right panel: Data from Darwin *et al.* (2003) for comparison. The grey areas in the two panels indicate the group of TMR values that was common in the two studies.

A mixed-model analysis of variance (ANOVA) was conducted on rau-transformed data, including listener, ΔF_0 and TMR as main factors. The listener was treated as a random factor while ΔF_0 and TMR as fixed factors. To analyse the interactions between the main experimental factors, two-way interactions were also included. The results of the ANOVA are reported in Table 2. All main factors were significant ($p < 0.001$), indicating that performance differs across listeners and that both ΔF_0 and TMR affect performance significantly. However, no significant interactions between factors were observed.

Factor	**Fixed/Random**	**df**	**F**	**p**
Listener	Random	14	11.17	<0.001
ΔF_0	Fixed	3	7.85	<0.001
TMR	Fixed	4	118.24	<0.001
Listener*ΔF_0	Random	42	1.21	0.201
Listener*TMR	Random	56	1.4	0.055
ΔF_0*TMR	Fixed	12	0.77	0.685
Error		167		

Table 2: Results of mixed-model analysis of variance performed on RAU-transformed proportion of correctly repeated words.

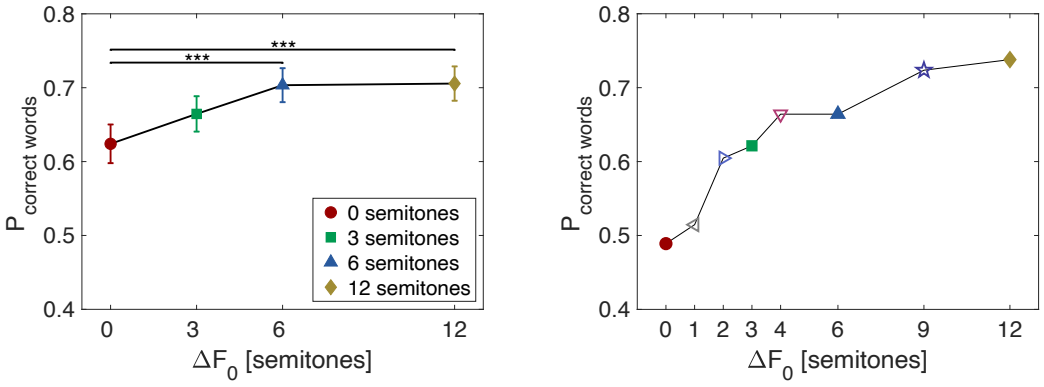


Fig. 3: Left panel: Word-recognition performance as a function of ΔF_0, averaged across TMRs. Error bars represent 95% confidence intervals. Right panel: Data from Darwin *et al.* (2003) averaged across their four lowest TMRs. The filled symbols represent ΔF_0 values that were in common with those of the current study.

Figure 3 shows the word recognition performance as a function of ΔF_0, averaged across TMRs. The results of the present study are shown in the left panel and those by Darwin *et al.* (2003) are displayed in the right panel. A post-hoc pairwise comparison analysis of the data from the present study showed that the difference between $\Delta F_0 = 0$ semitones and ΔF_0 = 6 or 12 semitones was statistically significant at the $p < 0.001$ level. The effect of ΔF_0 increased by 8% for ΔF_0 = 6 semitones and saturated above it, such that an increase in F_0 separation above 6 semitones did not provide any additional increase of speech intelligibility. In contrast, the data from Darwin *et al.* (2003), averaged across their four lowest TMRs, showed that a two-semitone separation was sufficient to obtain a 12% increase of speech intelligibility relative to the zero-semitone separation condition.

DISCUSSION

Overall, the results from the present study demonstrate that the benefit of substantial F_0 separations between competing sentences is relatively small in NH listeners when using a realistic speech corpus, in particular at conversational (positive) TMRs. The strongest effect of ΔF_0 was found at the negative TMRs and no effect of ΔF_0 was observed for the extreme values of TMR (-12 dB and 4 dB). It is possible that the task was too difficult at TMR = -12 dB, therefore making the ΔF_0 information ineffective, whereas it was too easy at TMR = 4 dB, making the ΔF_0 cue superfluous.

A strictly monotonic increase of speech intelligibility with increasing TMR was found for all ΔF_0s. This is in contrast to previous findings where results showed a non-monotonic relation with a local minimum at about 0 dB TMR (Brungart, 2001). This non-monotonic behavior was attributed to the low target level at negative TMRs that might have facilitated the segregation cue.

The high degree of sentence synchrony in the CRM speech corpus might have led to large amounts of energetic masking and thus to an overall more challenging task compared to a similar experiment that employed the HINT sentences. This can be noticed by comparing performances for $\Delta F_0 = 0$ semitones between the current and the reference study (left and right panels in Figure 2, respectively). Performances observed by Darwin *et al.* (2003) are lower than those obtained with the HINT speech corpus at the same TMR, meaning that to obtain a given speech intelligibility, more beneficial TMRs were needed in the reference study. A high degree of energetic masking may have emphasised the effect of F_0 separation beyond its real-life importance, resulting in an overestimation of the effects of ΔF_0.

Further work is required to prove this hypothesis. Furthermore, additional research is needed to investigate which other cues at the level of the F_0 contribute to speech separation besides the ΔF_0 and how these cues are affected by hearing impairment. The results may then be relevant for the development of hearing-aid processing strategies targeted to restore speech intelligibility in competing-talker scenarios.

REFERENCES

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am., **88**(2), 680-697.

Assmann, P. F. (**1999**). "Fundamental frequency and the intelligibility of competing voices," in Proc. International Congress of Phonetic Sciences, 179-182.

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in Proceedings of the Institute of Phonetic Sciences, **17**(1193), 97-110.

Bramsløw, L., Vatti, M., Hietkamp, R.K. and Pontoppidan, N. (**2015**). "Binaural speech recognition for normal-hearing and hearing-impaired listeners in a competing voice test," Proc. Speech in Noise, Copenhagen, Denmark.

Bramsløw, L., Vatti, M., Rossing, R., Naithani, G., and Henrik Pontoppidan, N. (**2019**). "A Competing Voices Test for Hearing-Impaired Listeners Applied to Spatial Separation and Ideal Time-Frequency Masks," Trends Hear., **23**, 1-12.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon., **1**(1), 23-36.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am., **109**(3), 1101-1109.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am., **114**(5), 2913-2922.

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol., **50**(3), 202-208.

Summers, V., and Leek, M. R. (**1998**). "F0 processing and the seperation of competing speech signals by listeners with normal hearing and with hearing loss," J. Speech Lang. Hear. Res., **41**(6), 1294-1306.

Effects of noise and L2 on the timing of turn taking in conversation

A JOSEFINE MUNCH SØRENSEN[1,*], MICHAL FERECZKOWSKI[1,2], AND EWEN N MACDONALD[1]

[1] *Department of Health Technology, Technical University of Denmark (DTU), DK-2800 Kgs. Lyngby, Denmark*

[2] *Institute of Clinical Research, University of Southern Denmark, DK-5230 Odense M, Denmark*

Previous studies of floor-transfer offsets (FTO), the offset from when one talker stops talking to the next one starts, suggest that normal conversation requires interlocutors to predict when each other will finish their turn. We hypothesized that increasing the difficulty of holding a conversation by adding noise and/or speaking in a second language (L2) would result in longer FTOs. Conversations from 20 pairs of normal-hearing (NH), native-Danish talkers were elicited using the Diapix task in four conditions consisting of combinations of language (Danish vs. English) and noise background (quiet vs. ICRA 7 noise presented at 70 dBA). Overall, participants took longer to complete the task in both noise and in L2 indicating that both factors reduced communication efficiency. In contrast to our predictions, in the presence of noise, the median of the FTO distribution decreased by approximately 40 ms and the standard deviation decreased by approximately 60 ms. However, the average median duration of utterances increased by 40% in noise. These findings are consistent with talkers holding their turn for longer, which may allow more time for their own speech planning. Overall, the results suggest that talkers may prioritise maintaining social norms for turn-taking fluency when communicating in difficult environments.

INTRODUCTION

When talkers take turns (i.e., there is a transfer of who has the floor), the acoustic signals produced by each talker may partially overlap or be separated by a silent gap. The length of this interval is termed the floor-transfer offset (FTO) with a negative value indicating an overlap and a positive indicating a gap. In previous studies of conversational interaction, the FTOs are shorter than the latencies of speech planning and articulation, suggesting that in addition to conducting speech understanding and speech planning in parallel, interlocutors must also predict when their partner will end their turn (Levinson and Torreira, 2015; Stivers *et al.*, 2009). In difficult communication conditions (e.g., in the presence of noise), speech understanding may require more cognitive resources, resulting in fewer resources available for speech

*Corresponding author: ajso@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

planning and reduced saliency of the acoustic cues used to predict turn ends (for a discussion of these cues see Gravano and Hirschberg, 2011). As a consequence, we hypothesized that speech planning may be delayed and prediction precision may be reduced, which would be observed by a shift to the right and a broadening of the FTO distribution. To test these hypotheses, conversations were recorded in conditions that varied in the degree of expected communication difficulty. Conversations were recorded both in the absence and presence of background noise, with talkers speaking both in their native language (Danish) and in a second language (English).

METHODS

Participants

40 young normal-hearing (NH) native Danish talkers ($\mu_{\text{age}} = 26$ years, $\sigma^2 = 3.7$ years, 12 women) participated in acquainted pairs (four mixed-gender pairs). All participants had hearing threshold levels below 20 dB HL between 125 Hz and 8 kHz, self-reported as being "comfortable" in English, and had all participated in at least one university-level class taught in English. All participants provided informed consent and the experiment was approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). The participants were compensated for their time.

Setup and Recordings

The participants were seated in separate listening booths and wore Sennheiser HD650 open headphones and head-worn Shure WH20 microphones. The microphone levels were calibrated such that the resulting presentation levels over headphones were the same as if their interlocutor was one meter away from them in the same room. All calibrations were done in dBA. The noise presented in the experiment was a 6-talker speech-shaped noise (ICRA 7, Dreschler *et al.* (2001)) calibrated to an average presentation level of 70 dBA. In the headphones, the participants heard a linear mix of themselves, their interlocutor, and the background noise (in noise conditions).

Task and Procedure

The participant pairs elicited dialogue by conducting the Diapix UK task (Baker and Hazan, 2011), a spot-the-difference task in which participants are given almost identical cartoon pictures, and they have to work together to find the differences between them. To familiarize participants with the task, they conducted a Diapix task using pictures from the original Diapix corpus Van Engen *et al.* (2010) facing each other outside the audiometric booths under the experimenter's supervision. Following this, they moved to the two separate booths and conducted a second Diapix task in background noise. The test session consisted of finding 10 differences between Diapix UK pairs in three repetitions of four conditions consisting of the combinations of conversing in either their first (L1, Danish) or second language (L2, English) in quiet or noise. The order of the conditions was randomized within each replicate. After each

replicate the participants had a break. All the recordings for which we have received consent have been made publicly available (Sørensen *et al.*, 2018).

Analysis of Recordings

Voice activity detection was performed on the individual microphone tracks to get binary speech activity arrays. For each conversation, the binary speech arrays from the two talkers were fed into a classification algorithm developed by the author. The algorithm categorized the conversations into utterances (speech tokens by each individual separated by silence of less than 180 ms), gaps (joint silence following a floor transfer), overlaps-between (joint speech during floor transfer), overlaps-within (joint speech during the utterance of one talker that does not result in a floor transfer), and pauses (joint silence not followed by a floor transfer), see Fig. 1. For analysing the effects of noise, second language and replicate on various measures, mixed-effects regression models were fitted to the variables of interest using the *lme4* package in *R*. ANalysis Of VAriance (ANOVA) tables were provided with Satterthwaite approximated denominator degrees-of-freedom (df) corrected F-tests for the fixed effects. The *lsmeans* function from the *lmerTest* package was used to compute pairwise comparisons of least-squares means of the significant effects using the Satterthwaite approximated df.

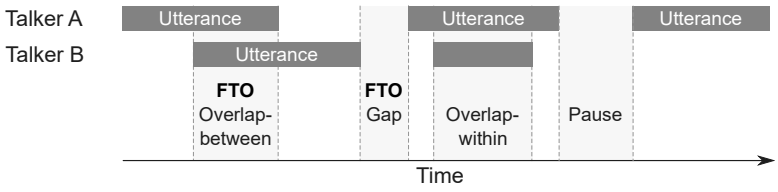


Fig. 1: Illustration of the classification of gaps, overlaps-within, overlaps-between, pauses, and utterances during conversations between Talker A and B. There are two floor-transfer offsets (FTOs): the overlap-between and gap.

RESULTS

Speech levels at a one-meter distance from each talker were estimated by computing the root mean square (RMS) level of all speech units, excluding pauses, and are plotted in Fig. 2, right panel. On average, talkers spoke 8.9 dBA louder in background noise than they did in quiet, resulting in an average SNR of -2.5 dB. There was a main effect of background [$F(1,39) = 1123.6, p <$ 2e-16], and of replicate [$F(1,39) = 3.91, p < 0.0283$]. A multiple comparison post-hoc analysis revealed that there was only a significant difference in level between replicate 1 and 3 [$t(39) = 2.648, p < 0.0116$], where the talkers spoke significantly softer by on average 0.44 dBA in the third replicate.

The task-completion time, i.e. time it took each pair to find 10 differences between the Diapix, was measured as a proxy for communication efficiency. The left panel

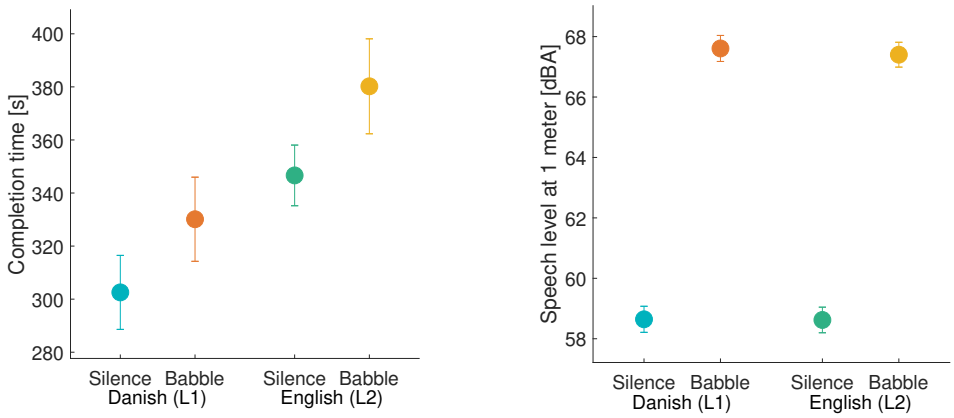


Fig. 2: Average completion time (left panel) and speech level (right panel) produced in the four combinations of language and noise. The bars indicate standard error.

of Fig. 2 shows the average completion time in the four conditions. There was a statistically significant training effect, i.e. the average completion time decreased with replicate [$F(2,216) = 23.1, p <$ 8.35e-10]. A pairwise comparison revealed a significant decrease in completion time between first and second replicate [$t(216) = 4.68, p <$ 5e-6], but only a borderline significant decrease between second and third replicate [$t(216) = 1.92, p < 0.0563$]. The completion time in noise compared to quiet increased significantly by 31 s [$F(1,216) = 12.3, p <$ 5.46e-4]. Similarly, participants took on average 47 s longer to complete the task in L2 compared to L1 [$F(1,216) = 29.2, p <$ 1.71e-7].

In all conversations, the articulation rates of the individual talkers were computed using the Praat script presented in Jong and Wempe (2009) with default parameter settings. The rate is measured as the number of syllables in the portions of the recording containing speech (i.e., periods of silence are excluded from the analysis). The average articulation rates are depicted in Fig. 3, left panel. The decrease in articulation rates in L2 compared to L1 was statistically significant [$F(1,438) = 680.02, p <$ 2.2e-16], and articulation rates increased significantly in noise [$F(1,438) = 60.3, p <$ 5.95e-14].

The average rate of overlaps-within (Fig. 3, right panel) increased both in L2 [$F(1,216) = 10.9, p < 0.00114$] and in noise [$F(1,216) = 77.7, p <$ 4.09e-16], and decreased with replicate: [$F(2,216) = 3.35, p < 0.037$]. A pairwise comparison revealed a significant decrease in rate between first and second replicate: [$t(216) = 2.42, p < 0.0165$], but not between second and third replicate [$t(216) = -0.408, p < 0.684$].

The overall hypothesis was that with increased processing demands, we would see

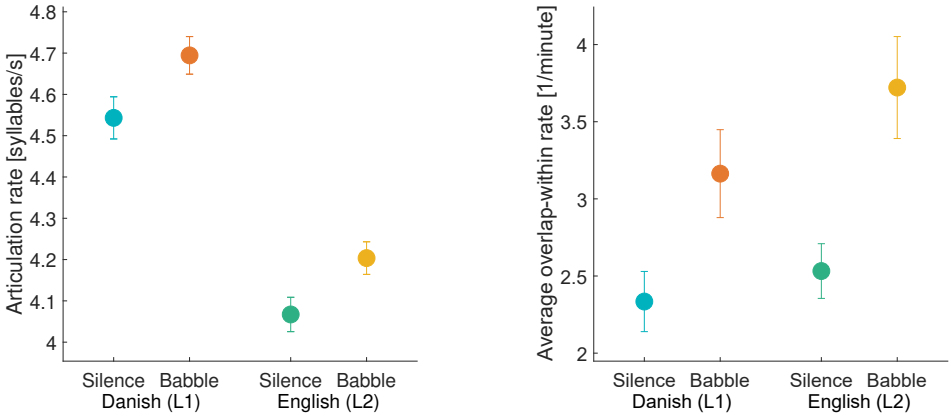


Fig. 3: Average articulation rates (left panel) and rate of occurrence of overlaps-within (i.e., turns from one talker that occur completely within a turn of the other talker) (right panel) produced in the four combinations of language and noise. The bars indicate standard error.

a delay and more variability in the timing of people's turn-taking. As a measure of centrality of the distribution, the median was used rather than the mean as FTO distributions are slightly positively skewed. There was a statistically significant main

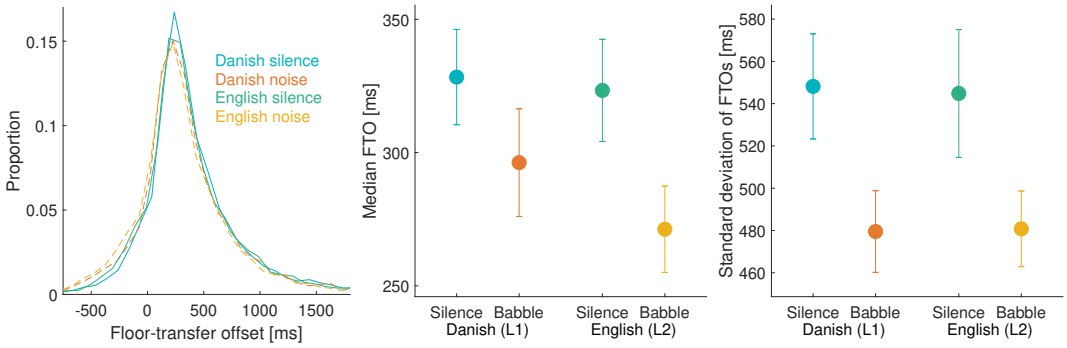


Fig. 4: Normalized distributions (left panel) of floor-transfer offsets (FTOs) along with the median (middle panel) and standard deviation (right panel) for the four combinations of language and noise. The bars indicate standard error.

effect of background [$F(1,218) = 46.6, p <$ 8.67e-11] and language [$F(1,218) = 5.91, p < 0.0159$] on the median FTO. Opposite to our hypothesis, the median decreased in the presence of background noise (by 40 ms, on average) and in L2 (by 15 ms, on average). The standard deviation averaged across pairs is plotted in Fig. 4. The analysis was performed on log-scaled standard deviations as the residuals were

not normally distributed, and there was a statistically significant effect of background [$F(1,219) = 21.5, p <$ 6.05e-6]. On average, the standard deviation decreased by 60 ms in background noise. In noise, interlocutors lengthened their turns substantially as

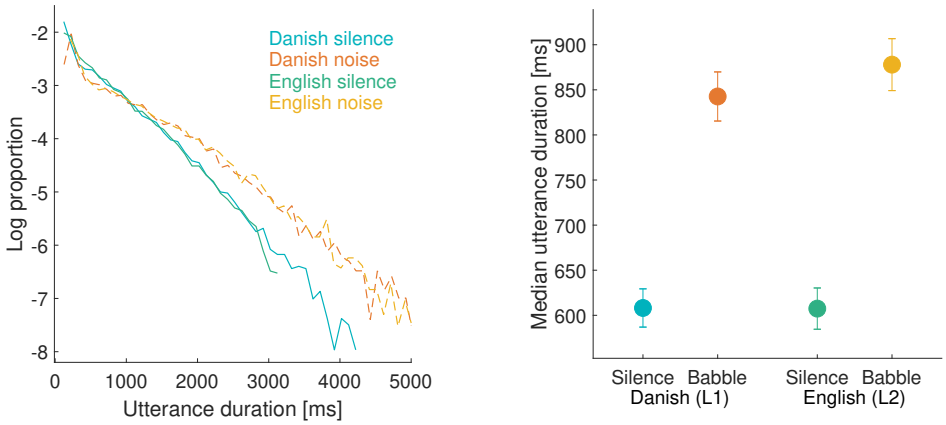


Fig. 5: Normalized distributions (left panel) and median (right panel) of the duration of utterances for the four combinations of task and noise. The bars indicate standard error. Note that the density in the left panel has been log-transformed to more easily compare the slopes.

is shown by a significant increase of about 42% of the median utterance duration in noise: [$F(1,219) = 738.71, p <$ 2.2e-16], depicted in Fig. 5, right panel. The analysis was performed on log-transformed utterance durations. The shallower slope of the pooled utterance durations across pairs as seen in Fig. 5, left panel, indicates a general lengthening of utterances in noise.

DISCUSSION

We hypothesised that communicating in noise and L2 would lead to an FTO distribution that was shifted to the right and broader due to an increase in cognitive processes used to understand and produce speech. We observed the opposite. In both noise and L2 the median FTO decreased, and the standard deviation decreased in noise. However, the changes were very small: the median decreased by about 40 ms in noise and 15 ms in L2, and the standard deviation decreased by about 60 ms in noise. Moreover, the overall shape of the distributions were similar across conditions and are similar to what has been found in other studies (e.g., Levinson and Torreira, 2015; Stivers *et al.*, 2009). This suggests that this turn-taking behaviour is important and that there may be a universal pattern in how we take turns that is more important to maintain than other properties of the speech we produce.

If maintaining rapid turn switches is important for social interaction, people may initiate their turns without having fully planned them yet. For example, they could

start with a filler-word, and then figure out what to say as they proceed with their sentence, leading to longer utterances. We saw, indeed, a substantial lengthening of the participant's turns in background noise. We also observed faster articulation rates in noise. Since speech planning is 3-4 times faster than articulation (Wheeldon and Levelt, 1995), longer utterances give the talker more time to continue speech planning.

In L2, participants decreased their rate of articulation and took longer to complete the task than in L1. However, unlike noise, L2 did not affect utterance duration or speech level. This is somewhat surprising as conversing in L2 should increase the processing loads for both speech perception and production, whereas noise should only affect the processing load for speech perception.

In Sørensen *et al.* (2020), when talking to a hearing-impaired (HI) interlocutor, the median FTO of both NH and HI increased in noise, and the FTO distributions became broader. While increased utterance durations were also found in that study, as well as other adaptations such as talking slower and overlapping less, their strategies may not have been enough to overcome the increased communication difficulty imposed by the hearing loss, leading to longer and more variable response times. In this study, however, the adaptive behaviour of the participants may have been sufficient to maintain "normal" turn switching behaviour.

Different, and sometimes opposite adaptive behaviours in noise have been observed in other conversation studies. For example, while Beechey *et al.* (2018); Sørensen *et al.* (2020) and the present study observed increases in utterance duration in noise, Hadley *et al.* (2019) observed the opposite. While there are several methodological differences across these studies, such as differences in the task or whether the average background noise levels were switched regularly or held fixed for several minutes, it is not yet clear which factors are responsible.

An increase in articulation rate, the amount of overlaps-within and the small decrease in median and standard deviation of FTOs in the noise condition may be an indication that interlocutors in the present study tried to move their operating point towards more rapid interaction. The increase in completion time, however, suggests that those changes did not maintain communication efficiency.

SUMMARY

Normal-hearing participants took longer to solve the Diapix task both in L2 vs. L1 and in noise vs. quiet. The median and standard deviation of FTOs decreased slightly in the presence of noise. An increase in the average utterance duration of about 40% in noise indicated that participants held their turn longer, allowing more time for their own speech planning. This suggests that interlocutors prioritize maintaining turn-taking fluency when adapting their behaviour in challenging acoustic environments.

ACKNOWLEDGEMENTS

A.J.M.S. and a portion of this study was supported by the William Demant Foundation (16-3968).

REFERENCES

Baker, R., and Hazan, V. (**2011**). "DiapixUK: Task Materials for the Elicitation of Multiple Spontaneous Speech Dialogs," Behav. Res. Methods, 43(**3**), 761–70. doi: 10.3758/s13428-011-0075-y

Beechey, T., Buchholz, J.M., and Keidser, G. (**2018**). "Measuring communication difficulty through effortful speech production during conversation," Speech Commun., **100**, 18-29. doi: 10.1016/j.specom.2018.04.007

de Jong, N. and Wempe, T. (**2009**). "Praat script to detect syllable nuclei and measure speech rate automatically,"Behav. Res. Methods, **41**, 385-90. doi: 10.3758/BRM.41.2.385

Dreschler, W., Verschuure, H., Ludvigsen, C. and Westermann, S. (**2001**). "ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment," Int. J. Audiol., 40(**3**), 148–57. doi: 0.3109/00206090109073110

Gravano, A., and Hirschberg, J. (**2011**). "Turn-taking cues in task-oriented dialogue", Comput. Speech Lang., **25**, 601–634. doi: 10.1016/j.csl.2010.10.003

Hadley, L. V., Brimijoin, W. O., and Whitmer, W. M. (**2019**). "Speech, movement, and gaze behaviours during dyadic conversation in noise," Sci. Rep., **9**, 10451. doi: 10.1038/s41598-019-46416-0

Levinson, S. C., and Torreira, F.(**2015**). "Timing in turn-taking and its implications for processing models of language," Front. Psychol. **6**, 731. doi: 10.1038/s41598-019-46416-0

Sørensen, A. J. M, Fereczkowski, M., and MacDonald, E. N. (**2018**). "Task dialog by native-Danish talkers in Danish and English in both quiet and noise," Dataset. doi: 10.5281/zenodo.1204951

Sørensen, A. J. M., Lunner, T., and MacDonald, E. N (**2020**). "Timing of turn taking between normal-hearing and hearing-impaired interlocutors," Proc. ISAAR, **7**, 37-44.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.P., Yoon, K., Levinson, S.C. (**2009**). "Universals and cultural variation in turn-taking in conversation," Proceedings of the National Academy of Sciences Jun 2009, **106**(26), 10587-10592. doi: 10.1073/pnas.0903616106

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (**2010**). "The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles," Lang. Speech, **53**(4), 510–540. doi: 10.1177/0023830910372495

Wheeldon, L. R., and Levelt, W. J. M. (**1995**). "Monitoring the time-course of phonological encoding," J. Mem. Lang., 34, 311–334. doi: 10.1006/jmla.1995.1014

Rapid perceptual learning of time-compressed speech and the perception of natural fast speech in older adults with presbycusis

TALI ROTMAN[1,*], LIMOR LAVIE[1] AND KAREN BANAI[1]

[1]*Department of Communication Sciences and Disorders, University of Haifa, Haifa, Israel*

Older people, especially ones with age-related hearing loss (ARHL), often have difficulties understanding naturally fast speech (NFS). This difficulty has been attributed to both sensory and cognitive factors. We now ask if rapid perceptual learning, assessed with a time-compressed speech (TCS) task, also contributes to the perception of NFS in older adults with ARHL, while accounting for the potential contribution of other cognitive factors. 45 participants with and without experience with hearing-aids completed the study. Significant rapid perceptual learning of TCS occurred within the first 20 sentences. This learning was significantly and positively correlated with NFS perception. Additionally, the perception of NFS was positively associated with vocabulary and memory span and negatively correlated with hearing thresholds. We found no significant differences between experienced-users of hearing-aids and non-users. Findings suggest that declines in rapid perceptual learning may play a role in the perception of NFS in people with ARHL, which is additional to the contribution of other cognitive variables.

INTRODUCTION

Perceiving rapid speech is a major challenge for older listeners (Gordon-Salant, 2001; Schneider *et al*. 2005; Janse, 2009). Many studies evaluated the effect of aging on the perception of rapid speech, in most cases using time-compressed speech (TCS). They attributed the perceptual difficulties of older listeners to both auditory (peripheral and central) and cognitive declines (Gordon-Salant, 2001; Wingfield & Tun; 2001; Janse, 2009; Gordon-Salant & Friedman, 2011).

Another factor that influences the recognition of rapid speech, even among older adults, is short-term perceptual learning (Peelle & Wingfield, 2005; Golomb et al., 2007). Manheim *et al*. (2018) showed that rapid perceptual learning is positively correlated with the perception of natural fast speech (NFS) in young normal hearing adults and in older adults, with either normal hearing or mild to moderate hearing loss. However, to our knowledge, the contribution of rapid perceptual learning to the perception of NFS has not been evaluated in older adults with more severe hearing impairments.

*Corresponding author: trotma01@campus.haifa.ac.il

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Thus, the main goal of the current study was to evaluate the contribution of short-term perceptual learning of TCS to the perception of NFS in older adults with significant hearing impairments. Because previous studies showed that sensory (hearing acuity, speech audibility) and cognitive factors also accounted for variance in speech recognition in elderly listeners (Humes, 2002; Lunner et al., 2009; Humes & Dubno, 2010), we also evaluated the contribution of hearing thresholds, the initial recognition of TCS and cognitive factors (working memory, selective attention, verbal and non-verbal intelligence) to the perception of NFS.We hypothesized that rapid perceptual learning would contribute to the perception of NFS, even after considering the contribution of these other factors.

We also compared the rapid perceptual learning of TCS and perception of NFS between experienced hearing-aid users and non-users (participants who have never used hearing-aids and weren't fitted with hearing-aids during the study period). Despite the fact that all participants were tested through headphones, we hypothesized that long-term experience with hearing-aids may have an impact on perceptual learning or speech perception under more strained conditions such as NFS.

MATERIALS AND METHODS

Participants

Forty-five older adults (24 males, 21 females) aged 65-89 years (mean age: 75, SD: 7) with age-related hearing loss participated in this study. The mean pure tone threshold (at 0.5, 1, 2 & 4 kHz) in both ears was 50dB HL (range: 30-70dB HL, SD: 8) and the mean score in the word recognition test (WRT) in quiet was 90% (SD: 7). All participants had an excellent control of Hebrew, with no history of neurological problems (according to self-report) and achieved a score of 24 and above in the Hebrew Mini-Mental State Examination (MMSE) (Folstein *et al*., 1975). The mean years of education was 14 (range: 8-22, SD: 3).

Twenty-three of the participants had a long-term experience (1-12 years, mean: 4.41, SD: 3.15) with digital hearing-aids in both ears, while the rest of the participants had no experience with amplification devices whatsoever. Both groups were balanced in mean age, gender, mean years of education and mean score in the MMSE.

Perception of NFS and TCS

40 simple Hebrew sentences (taken from Prior & Bentin, 2006) were used. All sentences were five to six words long and had a common subject-verb-object grammatical structure. 20 sentences were semantically plausible (e.g., "The wide pipe fills a round bath"), while the other 20 were semantically implausible (e.g., "The sharp knife hurt the inflated ball"). All sentences were recorded in a sound attenuating booth and their root mean square (RMS) levels were normalized after recording using Audacity audio software. 20 sentences (10 semantically plausible and 10 semantically implausible) were recorded in natural fast speech by a female native Hebrew speaker at a mean rate of 221 words/minute (SD=23). The other 20 sentences were recorded

by another female native Hebrew speaker at a normal speech rate of 113 words/minute (SD: 12). The latter were timed-compressed to 50% of their original length in Matlab, using a WSOLA algorithm (Verhelst & Roelands, 1993). We compressed the normal speech by 50% in order to match it to the mean speech rate of the NFS.

All stimuli were presented bilaterally via headphones (hearing-aids were not used during stimuli presentation) at the most comfortable level (MCL) of each listener (range: 107-115dB sound pressure level (SPL) , mean: 112, SD: 1.4). The individual MCLs were determined in reference to the levels in dB SPL that were measured in the headphones output for a pure tone of 1kHz played from a personal laptop. An increment of 5% on the volume scale in the computer was equivalent to a change of 1dB SPL. First, the participants listened to the 20 sentences in NFS, and then to the 20 time-compressed sentences. Plausible and implausible sentences were presented in a random order (the same order across all participants). The listeners were asked to repeat what they heard after each single trial and the experimenter transcribed their replies. Performance with NFS was quantified as the percentage of words correctly identified across all sentences.

The perception of TCS was used to evaluate the rapid perceptual learning. We followed previous studies (e.g. Dupoux & Green, 1997; Peelle & Wingfield, 2005) in defining the perceptual learning effect as the difference between the percentage of correctly identified words in the final and first five sentences.

In order to determine if rapid learning of TCS accounted for individual differences in NFS, we used a two-stage linear regression model. In the first stage, we created a base model with all potential explanatory variables, except for rapid learning of TCS. In the second stage, we added the rapid-learning slope.

Cognitive measures

Flanker task

A computerized version of the Flanker task (Eriksen & Eriksen, 1974) was used as a measure of selective attention. We created this computerized version in SuperLab software, according to the parameters presented by Scharenborg *et al*. (2015). Participants were asked to respond to the direction of a central arrow flanked by arrows pointing at the same direction (congruent trials), the opposite direction (incongruent) or = signs (neutral trials). The "flanker cost" for each participant was calculated as the mean log reaction time (logRT) in ms of the correct responses in the incongruent trials divided by the mean logRT of the correct responses in the neutral trails.

Subtests from the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III) in Hebrew:

1) *Digit span forward and backward* – This subtest was used to measure auditory working memory capacity. Participants listened (with their personal hearing-aids or with a personal amplifier) to a sequence of digits and were asked to recall the sequence correctly, with increasingly longer sequences being tested in each item.

Each item included two trials. In digit span forward the participants were asked to recall the sequence in the same order of presentation, while in the backward task they were asked to recall it in reverse order.

2) *Vocabulary* – This subtest evaluated the patient's semantic knowledge and verbal concept formation in Hebrew. The participants were asked to give oral definitions of up-to 33 words. Each word was presented both visually (written on a card) and orally (by the experimenter).
3) *Block Design* – The participants were requested to copy a two-colors pattern using colored blocks. The patterns became more and more complex and were time-limited (0.5/1/2 minutes according to the difficulty level of the specific item). The score in each item reflected the accuracy and speed of performance. This subtest assessed the participant's ability to understand complex visual information.

All three subtests were administrated and scored according to the test manual; raw scores were converted to standard scores according to the participant's age group.

RESULTS

Cognitive measures

Flanker task

The mean accuracy of the responses pooled over all participants was 94% (SD: 12). Accuracy was lowest and most variable in the incongruent condition (87%, SD: 24), while accuracy was best and least variable in the congruent condition (98%, SD: 8); accuracy for the neutral condition was close to that of the congruent condition (97%, SD: 9). The mean Flanker cost was 1.01 (SD: 0.007). Higher Flanker cost means poorer selective attention.

Subtests from the WAIS-III

The mean standard scores and standard deviations in digit span, vocabulary and block design subtests across all participants were: 9.17 (SD: 2.4), 9.17 (SD: 2.7) and 9.31 (SD: 2.4), respectively.

Perception of NFS and rapid perceptual learning of TCS

The mean percentage of the correctly identified words across all 20 sentences in NFS was 36% (SD: 18, median: 34%, inter-quartile range (25th-75th): 27%).

For each participant, we calculated the rapid perceptual learning of TCS as the difference between the mean percentage of correctly identified words across the first five sentences to that of the final five sentences. This difference was defined as the rapid-learning delta (Fig. 1A). The mean rapid-learning delta across all participants was 22% (SD: 16). Paired-samples t-test showed that the rapid-learning delta was statistically significant (t (88) = 4.31, p<.0001) with a large effect size (Cohen's d=1.37).

We also calculated the slope of the learning curve over the 20 TCS sentences as follows: mean recognition accuracy was calculated for each mini block of 5 sentences, and the slope of this curve was computed (Fig. 1B). The mean rapid-learning slope across all participants was 7% (SD: 5), equivalent to an improvement rate of 1-2 words/sentence.

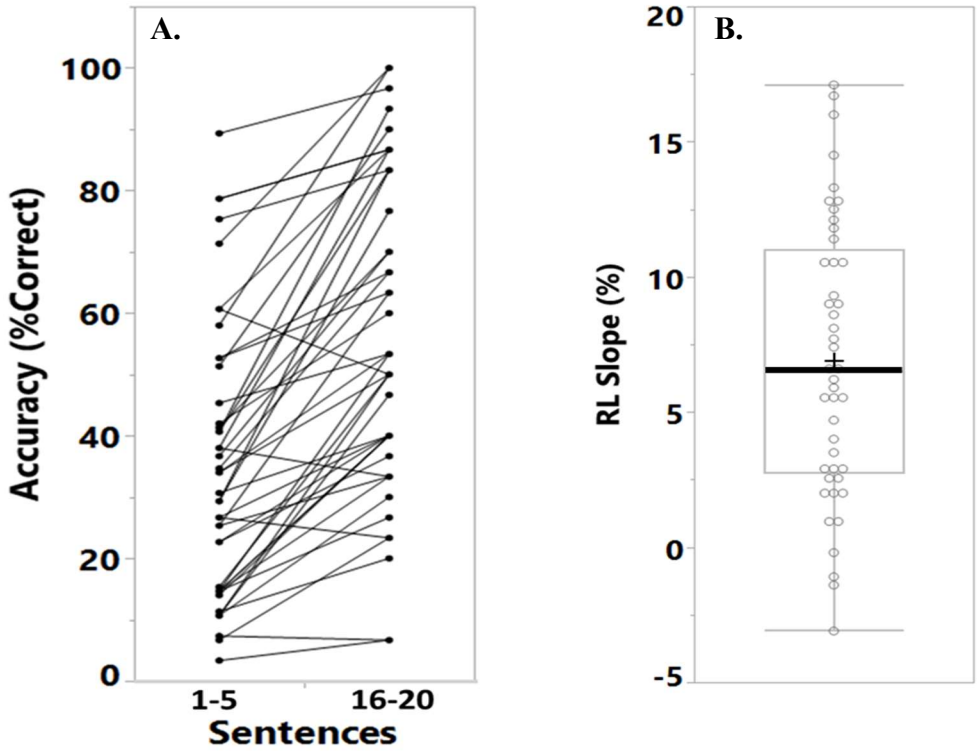


Fig. 1. A. Recognition of time-compressed speech. Line show mean individual performance during the first and the final five-sentences mini blocks. **B. Rapid-learning slope of time-compressed speech.** The box represents the interquartile range (25th-75th percentile). Whiskers are 1.5 the interquartile range. The thick line within the box marks the median and the + sign marks the mean. Each dot represents one participant (45 in total).

No significant differences were found between experienced and non-users in either rapid perceptual learning of TCS nor in perception of NFS ($t < 1$).

Individual differences in NFS perception as a function of hearing, cognition and learning

Recognition of NFS and the recognition accuracy in the first 2 time-compressed sentences were highly correlated ($r = 0.826$, $p < .001$). Likewise, the recognition of

NFS was also significantly or marginally correlated with age (r = 0.272, p=0.071), education (r = 0.399, p=0.007), hearing (PTA) (r = -0.499, p< .001), digit span (r = 0.444, p=0.002) and vocabulary (r = 0.465, p=0.001). The final linear regression model accounted for 85% of the variance in NFS perception, adjusted for the large number of predictors used relative to the sample size (F (7,37) = 37.88, p <.001). Of this, 7% were a unique contribution of rapid learning (F (1,37) = 20.99, p <.001). The full model is shown in Table 1. TCS perception at baseline was the strongest predictor, followed by rapid learning, hearing and vocabulary.

Stage	Predictor	R^2	Adjusted R^2	F	β	t
1	age				-0.08	-1.27
	education				0.14	1.93^{m}
	hearing				***-0.21***	***-3.66****
	digit span				0.04	0.58
	vocabulary				***0.20***	***2.81****
	TCS (baseline)				***0.62***	***8.7****
2	**rapid learning**				***0.27***	***4.58****
Full model		**0.88**	**0.85**	**37.88*** **		

Table 1: Summary of the regression model for natural fast speech perception. R^2 – unadjusted proportion of variance; β – standardized regression. Coefficients; $^{m}p < 0.1$, *p < 0.05, **p < 0.01, ***p < 0.001.

DISCUSSION

Consistent with our initial hypothesis, rapid perceptual learning of TCS and the perception of NFS were positively associated in older adults with significant age-related hearing loss. This extends the outcomes of two previous studies: Karawani *et al*. (2017) showed that older participants who improved less over the course of short-term speech-in-noise (SIN) training had poorer starting performance on the trained task as well as poorer performance on untrained speech in noise (SIN) tasks. Manheim *et al*. (2018) showed the same positive correlation between rapid perceptual learning of TCS and NFS perception in young adults with normal hearing, as well as in older adults with normal and impaired hearing. The results of the current study extend the results of Manheim *et al*.'s study to older adults with more severe hearing loss and account for the potential contributions of additional factors that have not been considered before in the context of perceptual learning for speech. Although correlations do not indicate causation, such correlations should exist if perceptual learning is to support the perception of degraded speech, as suggested by Samuel & Kraljic (2009).

Study participants completed tasks which evaluated the level of vocabulary, working memory, selective attention and non-verbal intelligence. We found that vocabulary and digit spans were positively correlated with the perception of NFS. Vocabulary was also a significant predictor of individual differences in NFS recognition. These findings are consistent with the 'ease of language learning' model (Rönnberg, 2003; Rönnberg *et al*., 2008; Rönnberg & Jerker, 2013). According to this model, perception of speech in adverse conditions (e.g. background noise, accented speech, rapid speech) involves explicit processing in which cognitive resources, such as working memory and semantic knowledge, have a critical role. Significant correlation between working memory capacity and recognition of SIN among older adults with age-related hearing loss was demonstrated by Nagaraj (2017).

To our knowledge, the present study is the first to evaluate the effect of long-term experience with hearing-aids on the perception of NFS and rapid perceptual learning of TCS. We found no differences between experienced hearing-aid users and the non-users. This result may stem from the fact that both groups listened to the stimuli through headphones at their MCLs, and hearing aids were not used during the perception and learning tests. Further research is needed to examine rapid speech perception through hearing aids, both in experienced users and in first-time hearing-aid users.

In conclusion, the results in this study support the idea that the difficulty of older adults to perceive rapid speech is connected to perceptual learning, but further work is required to elucidate the full interplay of sensory (e.g. temporal and spectral processes), cognitive and learning in the perception of challenging speech.

ACKNOWLEDGEMENTS

The Israel Science Foundation (grant 206/18) and Mr. Omri Gavish supported this study.

REFERENCES

Dupoux, E., Green K. (**1997**). "Perceptual adjustment to highly compressed speech: effect of talker and rate changes". J. Exp. Psychol. Hum. Percept. Perform., **23**, 914.

Eriksen, B. A., and Eriksen, C. W. (**1974**). "Effects of noise letters upon the identification of a target letter in a nonsearch task," Atten. Percept. Psycho., **16**, 143-149.

Folstein, M., Folstein, S., and McHugh, P. (**1975**). "Mini-mental state. A practical method for grading the cognitive state of patients for the clinician," J. Psych. Res., 12, 189-198.

Golomb, J. D., Peelle, J. E., and Wingfield, A. (**2007**). "Effects of stimulus variability and adult aging on adaptation to time-compressed speech," J. Acoust. Soc. Am., **121**, 1701-1708.

Gordon-Salant, S. (**2001**). "Sources of age-related recognition difficulty for time-compressed speech," J. Speech. Lang. Hear. Res., **44**, 709-719.

Gordon-Salant, S., and Friedman, S. A. (**2011**). "Recognition of rapid speech by blind and sighted older adults," J. Speech. Lang. Hear. Res., **54**, 622-631.

Humes, L. E. (**2002**). "Factors underlying the speech recognition performance of elderly hearing-aid wearers," J. Acoust. Soc. Am., **112**, 1112-1132.

Humes, L. E., and Dubno, J. R. (**2010**). "Factors affecting speech understanding in older adults," in *The Aging Auditory System*, Edited by S. Gordon-Salant et al. (Springer, New York, NY), pp. 211-257.

Janse, E. (**2009**). "Processing of fast speech by elderly listeners," J. Acoust. Soc. Am., **125**, 2361-2373.

Karawani, H., Lavie, L., and Banai, K. (**2017**). "Short-term auditory learning in older and younger adults," Proc. ISAAR, **6**, 1-8.

Lunner, T., Rudner, M., and Rönnberg, J. (**2009**). "Cognition and hearing-aids," Scan. J. Psycho, **50**, 395-403.

Manheim, M., Lavie, L., and Banai, K. (**2018**). "Age, hearing, and perceptual learning of rapid speech," Trends Hear., **22**, 1-18.

Nagaraj, N. K. (**2017**). "Working memory and speech comprehension in older adults with hearing impairment," J. Speech. Lang. Hear. Res., **60**, 2949-2964.

Peelle, J. E., and Wingfield, A. (**2005**). "Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech," J. Exp. Psychol. Hum. Percept. Perform., **31**, 1315.

Prior, A., and Bentin, S. (**2006**). "Differential integration efforts of mandatory and optional sentence constituents," Psychophys., **43**, 440-449.

Rönnberg, J. (**2003**). "Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: a framework and a model," Int. J. Audiol., **42**, S68-S76.

Rönnberg, J., Rudner, M., Foo, C., and Lunner, T. (**2008**). " Cognition counts: a working memory system for ease of language understanding (ELU)," Int. J. Audiol., **47**, S99-S105.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., …, and Rudner, M. (**2013**). "The ease of language understanding (ELU) model theoretical, empirical and clinical advances," Front. Syst. Neurosci., **7**, 31.

Samuel, A. G., and Kraljic, T. (**2009**). "Perceptual learning for speech," Atten. Percept. Psycho., **71**, 1207-1218.

Scharenborg, O., Weber, A., and Janse, E. (**2015**). "The role of attentional abilities in lexically guided perceptual learning by older listeners," Atten. Percept. Psycho., **77**, 493-507.

Schneider, B. A., Daneman, M., and Murphy, D. R. (**2005**). "Speech comprehension difficulties in older adults: Cognitive slowing or age-related changes in hearing?," Psychol. Aging., **20**, 261.

Verhelst, W. and Roelands, M. (**1993**). "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **2**, 554-557.

Wingfield, A., and Tun, P. A. (**2001**). "Spoken language comprehension in older adults: Interactions between sensory and cognitive change in normal aging," Semin. Hear., **22**, 287-302.

The next generation of audio intelligence: A survey-based perspective on improving audio analysis

BJÖRN SCHULLER[1,2,3], SHAHIN AMIRIPARIAN[2], GIL KEREN[2], ALICE BAIRD[2], MAXIMILIAN SCHMITT[2] AND NICHOLAS CUMMINS[2,*]

[1] *GLAM – Group on Language, Audio & Music, Imperial College London, SW7 2AZ London, UK*

[2] *Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, 86159 Augsburg, Germany*

[3] *audEERING GmbH, 82205 Gilching, Germany*

Computer audition has made major progress over the past decades; however it is still far from achieving human-level hearing abilities. Imagine, for example, the sounds associated with putting a water glass onto a table. As humans, we would be able to roughly "hear" the material of the glass, the table, and perhaps even how full the glass is. Current machine listening approaches, on the other hand, would mainly recognise the event of "glass put onto a table". In this context, this contribution aims to provide key insight into the already made remarkable advances in computer audition. It also identifies deficits in reaching human-like hearing abilities, such as in the given example. We summarise the state-of-the-art in traditional signal-processing-based audio pre-processing and feature representation, as well as automated learning such as by deep neural networks. This concerns, in particular, audio diarisation, source separation, understanding, but also ontologisation. Based on this, concluding avenues are given towards reaching the ambitious goal of "holistic human-parity" machine listening abilities – the next generation of audio intelligence.

INTRODUCTION

Typical real-world audio consists of complex combinations of overlapping events from a variety of sources, creating both clashing and harmonious relationships. Despite this complexity, humans can, with relative ease, decipher across audio through understanding, decomposing, interpreting, and ontologisation an abundance of potentially conveyed messages and their related semantic meanings.

Historically, developments in the field of computational audio understanding (computer audition) were initially driven by speech analysis, in particular, the field of automatic speech recognition (ASR). From its inception at Bell labs in the 1950's with the "Audrey" system, capable of recognising spoken digits (Davis *et al.*, 1952), through the considerable advancements during the 1980's associated with the use of Hidden Markov

*Corresponding author: schuller@ieee.org

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Models (Hansen and Hasan, 2015), and to the recent deep learning revolution (Hinton *et al.*, 2012), ASR technologies have now matured to the point where they are embedded in everyday technologies, e. g., SIRI™, CORTANA™, and ALEXA™. A similar transforming effect has recently occurred through deep learning, in terms of the immense increase in recognition accuracy and robustness in music analysis (e. g., Coutinho *et al.*, 2014), and for the recognition of acoustic scenes and the detection of specific audio events (Mesaros *et al.*, 2018).

Considering the advances in computer audition throughout the last decades, the time is now to unite these domains of audio understanding by creating a fully-fledged across-audio approach, thereby pushing this somewhat overlooked and currently underdeveloped mode of research to the forefront of intelligent machine understanding. To date, computer audition approaches have been typically mono-domain focused, with only consideration for the previously aforementioned domains of speech, music, and in general in an isolated singular manner. The view proposed here would unify these domains to truly understand and interpret audio.

The ground-breaking nature of such an "across-audio analysis" approach is the simultaneous understanding of the entire acoustic scene. Imagine, as an example, an acoustic scene set in a garage with two people working on repairing a car while listening to music. An across-audio analysis will isolate the conversation, the music, and engine noises and then assign relevant state and trait tags to each. For instance, the music genre and individual instrumentation could be recognised, the age and gender of each person and their relationship to one another determined, the car's age, model and condition identified, and finally the repair duration logged.

In the following, we move quickly through the state-of-the-art in audio analysis as related to the needed aspects of such a view on the next generation of audio intelligence: audio diarisation, (audio) source separation, audio understanding, (audio) ontologisation.

STATE-OF-THE-ART IN AUDIO ANALYSIS

Audio diarisation

Audio diarisation is a generalisation of speaker diarisation to general sound sources, e. g., vehicles, musical instruments, animals, or background noise types (Reynolds and Torres-Carrasquillo, 2005). The state-of-the-art is mostly marked by speaker diarisation, as general audio diarisation is only gaining momentum at this time. Speaker diarisation thereby is tagging an audio recording of several individuals with speaker turn information, i. e., to provide information relating to "who is speaking when". The dominating trend of the last few years in speaker diarisation research is to find suitable speaker embeddings which give a reliable multi-dimensional clustering of speech segments according to speakers. In this regard, the *i-vector* and *Gaussian mixture model-based* approaches (Anguera *et al.*, 2012) are being overtaken by deep neural network (DNN) feature representations (Bredin, 2017). Note that DNN-based

speaker embeddings are sometimes called *d-vectors*, as opposed to *i-vectors* (Wang *et al.*, 2017). The advantage of DNNs for speaker diarisation is that they are capable of simultaneously learning the embeddings, i. e., the feature vectors describing speaker characteristics, and the scoring function, which represents the similarity between the embeddings of different segments (Garcia-Romero *et al.*, 2017). Nevertheless, when comparing different scoring functions for i-vector embeddings, DNNs have been shown to outperform conventional scoring functions, such as *cosine similarity* and *probabilistic linear discriminant analysis* (Le Lan *et al.*, 2017).

Source separation

Audio source separation is the decomposition of an arbitrary audio signal into several signals with only a single audio source of interest present in each decomposed part. The audio source could be a speaker, a musical instrument, a sound produced by an animal or a vehicle, or background noise, such as breaking sea waves. In most conventional approaches, a mixture-spectrogram is separated into several source spectrograms. In the past, nonnegative matrix factorization (*NMF*; Nikunen *et al.*, 2018) or *non-negative tensor factorisation* (Ozerov *et al.*, 2011) have been used for single-channel (monaural) source separation (Virtanen, 2007), and *independent component analysis (ICA)* or *multichannel NMF* (Nikunen *et al.*, 2018) used for multi-channel audio.

Well-studied aspects of source separation are speech denoising and speech enhancement. Previous research on speech denoising comprises *NMF* (Weninger *et al.*, 2012), *deep NMF* (Le Roux *et al.*, 2015), recurrent neural network (*RNN*)-based discriminative training (Weninger *et al.*, 2014b), *long short-term memory recurrent neural networks* (*LSTM-RNNs*; Weninger *et al.*, 2015), *memory-enhanced RNNs* (Weninger *et al.*, 2014a), and *deep recurrent autoencoders* (Weninger *et al.*, 2014c). Latest approaches to *speech source separation* also employ different DNN types, such as *feed-forward neural networks* (*FFNNs*; Naithani *et al.*, 2016), RNNs (Huang *et al.*, 2015; Sun *et al.*, 2017) or *end-to-end (E2E) learning* using a CNN- or RNN-autoencoder instead of the usual spectral features (Venkataramani *et al.*, 2017). Recently, *generative adversarial nets (GANs)* were found to be promising in modelling speech (Subakan and Smaragdis, 2018) and singing sources (Fan *et al.*, 2018). For the task of music source separation, it was found that both FFNNs and RNNs are suitable, achieving superior scores in the *Signal Separation Evaluation Campaign (SiSEC)* music task (Uhlich *et al.*, 2017). Latest efforts in music source separation employed *U-nets*, a CNN variant from the image processing domain (Jansson *et al.*, 2017). Moreover, a *weakly labelled data* approach has also been proposed for the task of singing voice separation (Kong *et al.*, 2017). This approach utilised information about the presence or absence of singing as given by the output of a diarisation system. Notably, despite the huge amount of publications in the field of source separation, cross-domain audio signal separation (i. e., seperation of audio sources with distinct variance in character) is still largely unexplored.

Audio understanding

We consider audio understanding to be the task of acquiring a higher level semantic understanding of acoustic scenes, sound events, speech, and music. We consider audio understanding the task of acquiring a higher level semantic understanding of acoustic scenes, sound events, speech, and music. For this task, the aim of understanding the audio goes beyond the simple identification of speech, music, objects or events and their respective attributes. The goal, instead, should be to understand the relations between the elements of an acoustic scene. This understanding includes their relation to each other as well as their contextual meaning to a listener. For example, two individuals speaking loudly, followed by door slam and then a person crying, could be understood as a heated discussion causing emotional implications.

Unlike the field of computer vision, where considerable research has been carried out on higher-levels of semantic understanding of visual tasks (e. g., visual question answering (Agrawal *et al.*, 2017), image captioning (Xu *et al.*, 2015)), only a few works have been realised in the audio domain. One example is the recent work described in (Drossos *et al.*, 2017), in which an *encoder-decoder neural network* is used to process a sequence of Mel-band energies and to compute a sequence of words that describe a given audio segment. The already proved success of encoder-decoder sequence to seqence (S2S) architectures for structured prediction tasks such as more general audio combined with the small number of existing works applying such models to audio understanding tasks (to the best of our knowledge) creates a window of opportunity for conducting successful research in applying encoder-decoder for the above-mentioned tasks.

Audio ontologisation

A core component of an across-audio analysis, for both interpretation and understanding of acoustic scenes, is multi-domain audio ontologisation. An ontology is a formally documented knowledge base, which provides a precise description of the concepts encompassed within a domain, with additional attributes of each concept describing possible features. Within the machine learning community, ontologisation has been widely studied and applied in the text analysis domain (Buitelaar *et al.*, 2005), human activity recognition (Hoelzl *et al.*, 2014), and for "hierarchical" image-understanding domains (Durand *et al.*, 2007). In the audio domain, however, due to the complexities of the everyday life soundscapes, most efforts have been focused on specific domains (Han *et al.*, 2010; Nakatani and Okuno, 1998).

To date, there have been scarce attempts to create complete cross-audio domain ontologisations of everyday life soundscapes. The AudioSet (Gemmeke *et al.*, 2017) by Google has been perhaps the most interesting audio ontologisation attempt to date. It offers an ontologisation of audio events and their relationships within a sub-field, i. e., classes include; music, animals, human sounds, and the corresponding dependent children are; rock, dog, and whistling. AudioSet, however, does not include descriptors of the audio (e. g., the object action, or emotion). This aspect aside, it does provide a

platform for further and deeper ontologisation by the computer audition community. Until the release of AudioSet, the majority of works in ontologisation of acoustic scenes had come from studies focusing on the ontologisation of explicit audio domains, e. g., for music genre classification (Raimond *et al.*, 2007), music emotion perception (Han *et al.*, 2010), and audio features (Allik *et al.*, 2016). Excluding AudioSet, attempts at multi-domain audio ontologisation have mainly focused on the segregation of speech and music (Nakatani and Okuno, 1998), or sound objects retrieval (Hatala *et al.*, 2004).

In order to build a basis for ontologising a domain, previous research has commonly functioned in a manual nature, developing a methodology for collaborative ontology development via data mining based visual user interfaces, such as Orange WorkFlows (known as OWLs; Hilario *et al.*, 2009). These methods create a simple "seed" of basic concepts for the ontology structure (Noy *et al.*, 2006), with further adaptations requiring huge amounts of collaborative labour, using mechanisms for carrying out discussion (e. g., polling, and moderators; Farquhar *et al.*, 1997), something which in the long run can be time-consuming and costly. In an attempt to automate the construction of an ontology ((known as ontology learning; Gotmare, 2017), there have been efforts in the field of natural language processing, for intelligent web crawling (Maedche and Staab, 2001; Ehrig and Maedche, 2003). The web offers a mass of diverse but fragmented data sources, and targets for this can include Wikipedia, YouTube, and WordNet (Gemmeke *et al.*, 2017). Such approaches use relevance computation (Zheng *et al.*, 2008), to prioritise URLs of high relevance to the data which needs to be labelled, and extract metadata from social media, e. g., comments, tags, or titles. This textual data is then clustered into groups which may provide meaning to the associated data. To create these potential clustered groupings, unsupervised learning methods for data classification have been applied in the past (Vicient *et al.*, 2013), as well as semi-supervised and active learning methods, in which categories are assigned based on the most informative instances (Gotmare, 2017).

Until this point, the deep ontologisation of a particular domain has been time-consuming, requiring a mass of human labour (even the state-of-the-art AudioSet ontology required a huge amount of manual human effort; Gemmeke *et al.*, 2017).An across-audio–domain approach will not only improve on the state-of-the-art through the inherent need for additional and more expansive audio event terminology (e. g., body acoustics, animal calls, or automotive functions), but also through more fine-grained event attributes at both the state (e. g., mood) and the trait (e. g., age) level. A starting point can be given by exploiting deep learning-based approaches for web crawling (Amiriparian *et al.*, 2017), and clustering sourced data, as well as intelligent crowdsourcing approaches to reduce the need for manual labour, in which active learning is applied to prioritise the most informative instances (Hantke *et al.*, 2017).

TOWARDS THE NEXT GENERATION OF AUDIO INTELLIGENCE

From the above, we conclude that audio is largely being treated as a single-domain phenomenon, but the ingredients needed for a full-fledged "holistic" and likewise, an

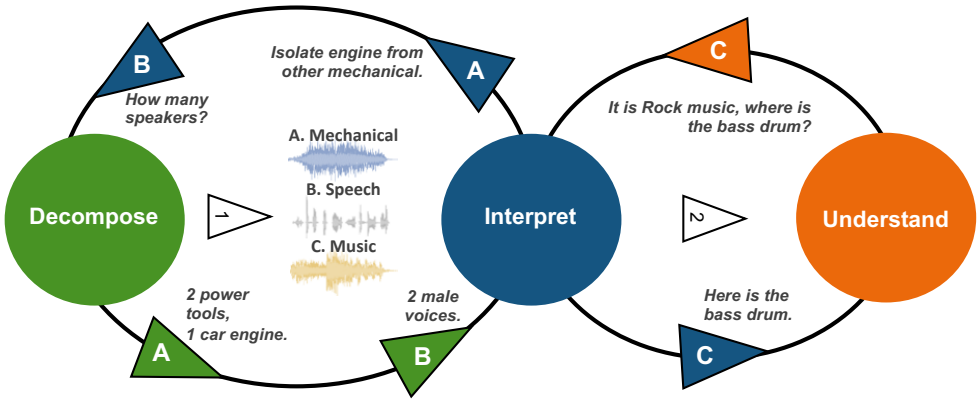


Fig. 1: Example for an iterative approach to decompose audio interpreting on different semantic levels of "understanding" to lead to an optimal "holistic" audio understanding. Imagine a garage with two people working on a car and listening to music as the (acoustic) scene.

arguably more "human-like" audio understanding are primarily available. In other words, one mainly needs to put the pieces of the puzzle together, and then feed a learning system with sufficient audio data. To overcome data sparseness, many approaches described in the literature use auditory and visual information in tandem to improve the understanding of video content. In Aytar *et al.* (2016), a neural network is trained on a corpus of unlabelled videos to match the representation extracted from the audio part with that extracted from the visual information by pretrained networks for object and scene classification. Facilitating such research avenues, there exist a number of video corpora, that can be used for a multimodal video understanding such as Rohrbach *et al.* (2015) and Torabi *et al.* (2015).

Figure 1 visualises a potential concept towards such holistic audio intelligence. It uses an example of an acoustic scene, as described in the introduction. The number and type of sources present in an audio signal are not known beforehand. Hence, decomposition could be modelled as an iterative process in interaction with an interpretation component, which is providing information about the signal and indicating a request for further separation, as illustrated in Figure 1. In the proposed across-audio-domain iterative decomposition solution, the first step would be to decompose speech, music, and sound and send separate signals to the interpretation component. The interpreter would be able to identify the types and then call the source separation again to decompose the signal events further. The source separation is aided by weak labels from the diarisation in this context, to know the temporal occurrences of the fractionally overlapping events. Finally, after the types of the audio have been classified by the interpretation component, these are analysed deeper with respect to states, finding that potentially parts are missing from a semantically higher perspective. This deeper analysis allows for an iterative process. Figure 2 additionally

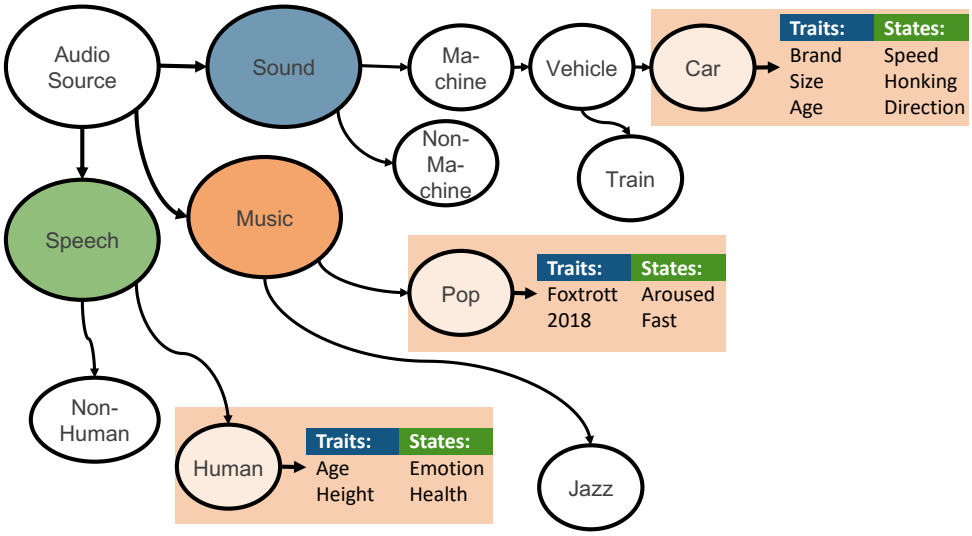


Fig. 2: Example of an ontology that consequently attributes audio sources states and traits – not only for speech as is the current usual state-of-literature. In this depiction we see that the audio source is decomposed into 3 sub-sources; speech, music and sound, which are then each further decomposed. For example, one of the "sound" sources is noted as being mechanical, vehicle, car, and the car is further labelled for its brand, as well as action e. g., Speed.

exemplifies audio ontologies that could suit the need for a complete and "holistic" audio understanding. Note that the concept of state and trait assignment as known from speech analysis is consequently extended to general audio sources such as sound or music – after all, sound always has a source which has certain traits and is in certain states.

CONCLUSION

We discussed the state-of-the-art in audio intelligence focusing on audio understanding when it comes to general audio which often consists of a blend of speech and/or music and/or sound. We surveyed in nutshell components which we believe are crucial to lead to a general audio understanding including audio diarisation, source separation, understanding, and ontologisation. From this, we showed a potential approach on how to combine the pieces to lead to a more advanced form of "cross-domain" audio analysis with a rich ontology unified across the audio domains. To realise such a concept, recent deep learning methods seem well suited, such as learning weakly supervised in an end-to-end manner. Once realised, such an audio intelligence will find an abundance of potential applications from retrieval to robotics, and beyond.

REFERENCES

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. (**2017**), "VQA: Visual Question Answering," Int. J. Comput. Vis., **123**(1), 4–31.

Allik, A., Fazekas, G., and Sandler, M. B. (**2016**), "An Ontology for Audio Features," Proc. International Society for Music Information Retrieval Conference (ISMIR) (ISMIR, New York, NY), 73–79.

Amiriparian, S., Pugachevskiy, S., Cummins, N., Hantke, S., Pohjalainen, J., Keren, G., and Schuller, B. (**2017**), "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," Proc. Biannual Conference on Affective Computing and Intelligent Interaction(ACII) (San Antonio, TX), 340–345.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (**2012**), "Speaker diarization: A review of recent research," IEEE Trans. Audio Speech Lang. Process., **20**(2), 356–370.

Aytar, Y., Vondrick, C., and Torralba, A. (**2016**), "SoundNet: Learning sound representations from unlabeled video," Proc. Advances in Neural Information Processing Systems (NIPS) (MIT Press, Barcelona, Spain), 892–900.

Bredin, H. (**2017**), "TristouNet: Triplet Loss for Speaker Turn Embedding," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, New Orleans, LA), 5430–5434.

Buitelaar, P., Cimiano, P., and Magnini, B. (**2005**), *Ontology learning from text: methods, evaluation and applications* (Impacting the World of Science Press, Amsterdam, The Netherlands).

Coutinho, E., Weninger, F., Schuller, B., and Scherer, K. (**2014**), "The Munich LSTM-RNN approach to the MediaEval 2014 "Emotion in Music" Task," Proc. MediaEval Multimedia Benchmark Workshop (CEUR, Barcelona, Spain), no pagination.

Davis, K., Biddulph, R., and Balashek, S. (**1952**), "Automatic recognition of spoken digits," J. Acoust. Soc. Am., **24**(6), 637–642.

Drossos, K., Adavanne, S., and Virtanen, T. (**2017**), "Automated audio captioning with recurrent neural networks," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (IEEE, New Paltz, NY), 374–378.

Durand, N., Derivaux, S., Forestier, G., Wemmert, C., Gançarski, P., Boussaid, O., and Puissant, A. (**2007**), "Ontology-based object recognition for remote sensing image interpretation," Proc. IEEE International Conference on Tools with Artificial Intelligence (ICTAI) (IEEE, Patras, Greece), 472–479.

Ehrig, M. and Maedche, A. (**2003**), "Ontology-focused Crawling of Web Documents," Proc. ACM Symposium on Applied Computing (SAC) (ACM, Melbourne, Florida), 1174–1178.

Fan, Z., Lai, Y., and Jang, J. R. (**2018**), "SVSGAN: Singing Voice Separation Via Generative Adversarial Network," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Calgary, Canada), 726–730.

Farquhar, A., Fikes, R., and Rise, J. (**1997**), "The Ontolingua Server: A tool for

collaborative ontology construction," Int. J. Hum.-Comput. St., **46**(6), 707–727.

Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., and McCree, A. (**2017**), "Speaker diarization using deep neural network embeddings," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (New Orleans, LA), 4930–4934.

Gemmeke, J., Ellis, D., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (**2017**), "Audio set: An ontology and human-labeled dataset for audio events," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, New Orleans, LA), 776–780.

Gotmare, P. (**2017**), "Methodology for Semi-Automatic Ontology Construction using Ontology learning: A Survey," IJCA Proceedings on Emerging Trends in Computin, volume ETC-2016, 1–3.

Han, B., Rho, S., Jun, S., and Hwang, E. (**2010**), "Music emotion classification and context-based music recommendation," Multimed. Tools Appl., **47**(3), 433–460.

Hansen, J. H. L. and Hasan, T. (**2015**), "Speaker Recognition by Machines and Humans: A tutorial review," IEEE Signal Process. Mag., **32**(6), 74–99.

Hantke, S., Zhang, Z., and Schuller, B. (**2017**), "Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world," Proc. INTERSPEECH (ISCA, Stockholm, Sweden), 3951–3955.

Hatala, M., Kalantari, L., Wakkary, R., and Newby, K. (**2004**), "Ontology and rule based retrieval of sound objects in augmented audio reality system for museum visitors," Proc. ACM Symposium on Applied Computing (SAC) (ACM, Nicosia, Cyprus), 1045–1050.

Hilario, M., Kalousis, A., Nguyen, P., and Woznica, A. (**2009**), "A data mining ontology for algorithm selection and meta-mining," Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) (Bled, Slovenia), 76–87.

Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., and Sainath, T. (**2012**), "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," IEEE Signal Process. Mag., **29**(6), 82–97.

Hoelzl, G., Ferscha, A., Halbmayer, P., and Pereira, W. (**2014**), "Goal oriented smart watches for cyber physical superorganisms," Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (ACM, Seattle, WA), 1071–1076.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (**2015**), "Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation," IEEE/ACM Trans. Audio Speech Lang. Process., **23**(12), 2136–2147.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. (**2017**), "Singing voice separation with deep U-Net convolutional networks," Proc. International Society for Music Information Retrieval Conference (ISMIR) (ISMIR, Suzhou, China), 323–332.

Kong, Q., Xu, Y., Wang, W., and Plumbley, M. D. (**2017**), "Music Source Separation

using Weakly Labelled Data," Proc. International Society for Music Information Retrieval Conference (ISMIR) (Suzhou, China), no pagination.

Le Lan, G., Charlet, D., Larcher, A., and Meignier, S. (**2017**), "A Triplet Ranking-based Neural Network for Speaker Diarization and Linking," Proc. INTERSPEECH (ISCA, Stockholm, Sweden), 3572–3576.

Le Roux, J., Hershey, J. R., and Weninger, F. (**2015**), "Deep NMF for speech separation," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Brisbane, Australia), 66–70.

Maedche, A. and Staab, S. (**2001**), "Ontology learning for the semantic web," IEEE Intell. Syst., **16**(2), 72–79.

Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., and Plumbley, M. D. (**2018**), "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," IEEE/ACM Trans. Audio. Speech Lang. Process., **26**(2), 379–393.

Naithani, G., Parascandolo, G., Barker, T., Pontoppidan, N. H., and Virtanen, T. (**2016**), "Low-latency sound source separation using deep neural networks," Proc. Global Conference on Signal and Information Processing (GlobalSIP) (Washington, DC), 272–276.

Nakatani, T. and Okuno, H. G. (**1998**), "Sound ontology for computational auditory scene analysis," Proc. Conference of the Association for the Advancement of Artificial Intelligence (AAAI) (Madison, WI), 1004–1010.

Nikunen, J., Diment, A., and Virtanen, T. (**2018**), "Separation of Moving Sound Sources Using Multichannel NMF and Acoustic Tracking," IEEE/ACM Trans. Audio Speech Lang. Process., **26**(2), 281–295.

Noy, N. F., Chugh, A., Liu, W., and Musen, M. A. (**2006**), "A Framework for Ontology Evolution in Collaborative Environments," Proc. International Semantic Web Conference (ISWC) (Athens, GA), 544–555.

Ozerov, A., Févotte, C., Blouet, R., and Durrieu, J.-L. (**2011**), "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Prague, Czech Republic), 257–260.

Raimond, Y., Abdallah, S. A., Sandler, M. B., and Giasson, F. (**2007**), "The Music Ontology," Proc. International Society for Music Information Retrieval Conference (ISMIR) (Vienna, Austria), 417–422.

Reynolds, D. A. and Torres-Carrasquillo, P. (**2005**), "Approaches and applications of audio diarization," Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (IEEE, Philadelphia, PA), 953–956.

Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. (**2015**), "A dataset for Movie Description," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Boston, MA), 3202–3212.

Subakan, Y. C. and Smaragdis, P. (**2018**), "Generative Adversarial Source Separation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Calgary, Canada), 26–30.

Sun, Y., Zhu, L., Chambers, J. A., and Naqvi, S. M. (**2017**), "Monaural source separation based on adaptive discriminative criterion in neural networks," Proc. International Conference on Digital Signal Processing (DSP) (London, UK), 1–5.

Torabi, A., Pal, C., Larochelle, H., and Courville, A. (**2015**), "Using descriptive video services to create a large data source for video annotation research," arXiv preprint arXiv:1503.01070.

Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., and Mitsufuji, Y. (**2017**), "Improving music source separation based on deep neural networks through data augmentation and network blending," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (New Orlearns, LA), 261–265.

Venkataramani, S., Casebeer, J., and Smaragdis, P. (**2017**), "Adaptive Front-ends for End-to-end Source Separation," Proc. Conference on Neural Information Processing Systems (NIPS) (Long Beach, CA), no pagination.

Vicient, C., Sánchez, D., and Moreno, A. (**2013**), "An automatic approach for ontology-based feature extraction from heterogeneous textualresources," Eng. Appl. Artif. Intel., **26**(3), 1092–1106.

Virtanen, T. (**2007**), "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," IEEE Trans. Audio Speech Lang. Process., **15**(3), 1066–1074.

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., and Moreno, I. L. (**2017**), "Speaker diarization with LSTM," arXiv preprint arXiv:1609.04301.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., and Schuller, B. (**2015**), "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," Proc. International Conference on Latent Variable Analysis and Signal Separation (Liberec, Czech Republic), 91–99.

Weninger, F., Eyben, F., and Schuller, B. (**2014**a), "Single-channel speech separation with memory-enhanced recurrent neural networks," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Florence, Italy), 3709–3713.

Weninger, F., Hershey, J. R., Le Roux, J., and Schuller, B. (**2014**b), "Discriminatively trained recurrent neural networks for single-channel speech separation," Proc. Global Conference on Signal and Information Processing (GlobalSIP) (Atlanta, GA), 577–581.

Weninger, F., Watanabe, S., Tachioka, Y., and Schuller, B. (**2014**c), "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Florence, Italy), 4623–4627.

Weninger, F., Wöllmer, M., and Schuller, B. (**2012**), "Combining Bottleneck-BLSTM and Semi-Supervised Sparse NMF for Recognition of Conversational Speech in Highly Instationary Noise," Proc. INTERSPEECH (Portland, OR), 302–305.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (**2015**), "Show, Attend and Tell: Neural Image Caption Generation with

Visual Attention," Proc. International Conference on Machine Learning (ICML) (Lille, France), 2048–2057.

Zheng, H.-T., Kang, B.-Y., and Kim, H.-G. (**2008**), "An ontology-based approach to learnable focused crawling," Inf. Sci., **178**(23), 4512–4522.

Prediction of speech intelligibility with DNN-based performance measures

ANGEL MARIO CASTRO MARTINEZ*, CONSTANTIN SPILLE, BIRGER KOLLMEIER AND BERND T. MEYER

Medizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

In this paper, we present a speech intelligibility model based on automatic speech recognition (ASR) that combines phoneme probabilities obtained from a deep neural network and a performance measure that estimates the word error rate from these probabilities. In contrast to previous modeling approaches, this model does not require the clean speech reference or the exact word labels during test time, and therefore, less *a priori* information. The model is evaluated via the root mean squared error between the predicted and observed speech reception thresholds from eight normal-hearing listeners. The recognition task in both cases consists of identifying noisy words from a German matrix sentence test. The speech material was mixed with four noise maskers covering different types of modulation. The prediction performance is compared to four established models as well as to the ASR-model using word labels. The proposed model performs almost as well as the label-based model and produces more accurate predictions than the baseline models on average.

INTRODUCTION

The intelligibility of speech is crucial for our social interaction, and it is an important measure for a diagnosis of hearing deficits through speech audiometry and for the optimization of speech enhancement algorithms in hearing aids or cochlear implants. Accurate models that predict the speech intelligibility (SI) in the presence of different masking noises are desirable since they can quantify the outcome of such an optimization and could, therefore, reduce the requirement of SI measurements that are usually time-consuming and costly.

Several models for SI prediction have been proposed that take into account the signal-processing strategies of the auditory system such as the speech-intelligibility index (SII; ANSI, S3 22-1997, 1997); the extended SII (ESII; Rhebergen and Versfeld, 2005) which extends SII to account for temporal modulations; the short-time objective intelligibility (STOI; Taal *et al.*, 2011), which is based on correlations between original and degraded signal; and the multi-resolution speech envelope power spectrum model (mr-sEPSM; Ewert and Dau, 2000), which incorporates temporal modulation filters in different frequency bands.

*Corresponding author: angel.castro@uni-oldenburg.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Schubotz *et al.* (2016) compared these models in a study to determine how well they can predict the speech reception threshold (SRT), which is the signal-to-noise ratio (SNR) at which 50% of words presented are correctly recognized.

An alternative modeling approach combines signal extraction based on auditory principles with pattern matching algorithms borrowed from automatic speech recognition (ASR). For example, Barker and Cooke (2006) introduced a glimpsing model in which the above-threshold time-frequency patches (glimpses) were used as features for a backend that combines a Gaussian mixture model (GMM) with a hidden Markov model (HMM) to produce a transcript from the input glimpses, which was compared to listener responses. A GMM-HMM approach dubbed Framework for Acoustic Discrimination Experiments (FADE) was proposed in (Schädler *et al.*, 2015). This model produces SRT estimates by retraining a GMM-HMM system at different SNRs, and by selecting the model that produces the lowest SRT when using the same training and test sentences.

All previously mentioned models either require separate clean and degraded speech, or separate speech and noise signals. Motivated by the success of deep learning in ASR, Spille *et al.* (2018) proposed an ASR model that combines a deep neural network (DNN) trained to estimate phoneme probabilities given the acoustic observation with an HMM. The predictive power of this model exceeded the four baseline models mentioned above on the dataset collected by Schubotz and colleagues. The root-mean-square error (RMSE) between measurement and prediction was 1.8 dB on average when using multi-condition training as well as modulation features, which can be compared to the RMSE of baseline models in the range of 5.6 to 9.5 dB. The model is blind with respect to speech because training and test sets are speaker-independent. Therefore, it marks a step towards reference-free SI models, which could serve as models-in-the-loop in assisted hearing. A use case for such a model is the constant monitoring of SI in the current acoustic scene and to identify the speech enhancement algorithm that is optimal for that scene.

However, it requires the correct labels of the words in the utterance used as model input. These labels are compared to the transcript produced by the ASR system from which the recognition accuracy is calculated. For online applications of SI models, this is an essential limitation of the models.

In this paper, we introduce a model of SI prediction that does not require either the speech reference *or* the actual labels of the tested utterances. The model is based on the DNN-based approach introduced in Spille *et al.* (2018), but instead of computing the word error rate (WER), we test a method for *estimating* the WER directly from the phoneme posterior probabilities emitted by the DNN. The method explored here was first proposed for estimating phone error rates (Hermansky *et al.*, 2013) by analyzing the mean temporal distance or M-measure of phoneme vectors obtained from a neural network.

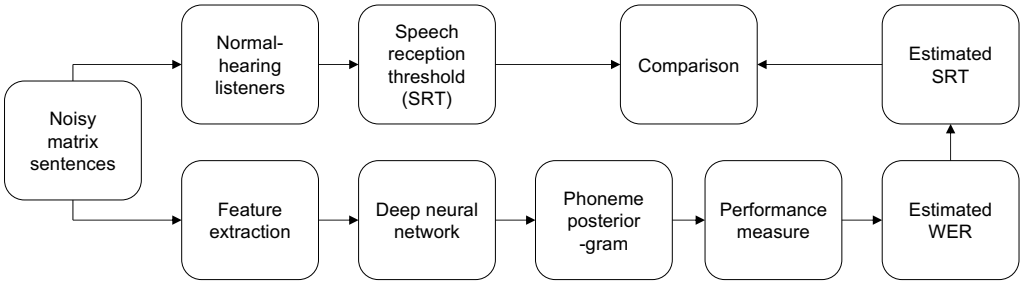


Fig. 1: Building blocks of the modeling approach: Speech intelligibility in noisy sentences is compared to the estimated SRT. To obtain this estimate, a DNN is trained as part of a standard ASR system and subsequently used to measure the degradation of phoneme representations in noise using a performance measure. From this, the WER of the ASR system is estimated, resulting in the predicted SRT.

In the current study, we quantify the relation of the M-measure with the WER and explore if accurate model predictions can be obtained with the DNN alone that operates on a relatively small temporal context window.

MODEL STRUCTURE

Figure 1 illustrates the structure of the proposed model. In previous work, Spille *et al.* (2018) estimated the SRTs from the target speech transcript, and the WER computed from the output of a regular hybrid ASR system. The components of this system are summarized below, together with the modification of using the M-measure to drop the requirement of word labels and estimate SRTs directly from the output of the acoustic model. The characterization of the acoustic model, together with the respective input features, closes this section.

ASR-based model of speech intelligibility

In the label-based ASR approach (Spille *et al.*, 2018), the acoustic model was trained on speech files mixed with different maskers at various SNRs using the Kaldi toolkit[†]. A fully-connected feed-forward DNN was used to map the acoustic features to posterior probabilities of context-dependent triphones. The time sequence of these probabilities was decoded using an HMM (three states for modeling phonemes and five for silence) to obtain a transcript of the utterance. This transcript was compared to the ground truth labels to obtain the word error rate (WER) from this sentence. By using utterances at various SNRs, a broad range of the corresponding WER estimations was obtained. These pairs of points were fitted to a psychometric function, as described by Wagener *et al.* (1999). SRT values served as a mean intelligibility predictor and were compared to SRTs obtained in listening experiments.

[†]The ASR was implemented using the Kaldi speech recognition toolkit (Povey *et al.*, 2011).

The model proposed in this paper differs from this approach by using the HMM during the training procedure only, and omitting the HMM (or any other language model) when predictions were obtained. Instead, a performance measure, as described in the next section, quantifies the degradation of the phoneme posteriorgrams. This measure could be informative about the WER and hence, about the SRT too. We hypothesize that the resulting measure should be sensitive to the SNR, but also show a similar sensitivity to masking noises similar to human listeners as long as the amount of training data is sufficient.

Estimating the word error rate

The WER estimation is based on a measure that quantifies the degradation of phoneme probabilities obtained from a DNN. We chose the mean temporal distance (MTD, also referred to as M-measure) (Hermansky *et al.*, 2013), which takes into account the distance of phoneme vectors and averages this distance. The underlying idea is that acoustically challenging conditions can have a temporal smearing effect on phoneme representations, i.e., the phoneme vectors become more similar. Acoustically optimal conditions produce very distinct phoneme activations, which become more distant in vector space. This distance is captured by the entropy-based divergence averaged over phoneme vectors in the interval from 50 to 800 ms.

The M-measure accumulates the average divergences of two phoneme posterior vectors $\mathbf{p}_{t-\Delta t}$ and $\mathbf{p}_t$ separated by a time interval of Δt and is defined as

$$\mathscr{M}(\Delta t) = \frac{1}{T-\Delta t} \sum_{t=\Delta t}^{T} \mathscr{D}(\mathbf{p}_{t-\Delta t}, \mathbf{p}_t) \qquad \text{(Eq. 1)}$$

where T is the duration of the analyzed representation, in this case, a portion of the posteriorgram. The symmetric Kullback-Leibler divergence is used as distance measure $\mathscr{D}$ between phoneme posterior vectors $\mathbf{p}_{t-\Delta t}$ and $\mathbf{p}_t$.

$$\mathscr{D}(\mathbf{p},\mathbf{q}) = \sum_{k=0}^{K} p^{(k)} \log \frac{p^{(k)}}{q^{(k)}} + \sum_{k=0}^{K} q^{(k)} \log \frac{q^{(k)}}{p^{(k)}} \qquad \text{(Eq. 2)}$$

As defined above, $p^{(k)}$ is the k-th element of the posterior vector $\mathbf{p} \in \mathbb{R}^k$.

We considered 16 values of Δt per utterance; from 50 to 800 ms in steps of 50 ms. For short Δt time spans, divergences are small, indicating neighboring frames often correspond to the same phoneme. The value increases with time up to a point at which both vectors $\mathbf{p}$, and $\mathbf{q}$ come from different coarticulation patterns, and the curve saturates.

As the acoustic model is trained to produce triphone posteriorgrams, to be decoded by the language model when performing ASR, an intermediate step of *grouping* the activations was performed to obtain monophone posteriorgrams. It is possible to

cluster triphones by mapping each transition as a branch of a decision tree, wherein the roots correspond to the central phoneme of the triphone. Monophone posteriorgrams of 42 dimensions were obtained by adding the corresponding activations, thus maintaining the distribution.

Monophone posteriorgrams yield M-measure values comparable to the ones obtained with the triphone equivalents at a lower computational cost without constraining the acoustic model of the temporal context if trained to produce monophones directly.

In our previous study (Castro Martinez *et al.*, 2019), we established the correlation between WER and the M-measure. In this work, we leverage this property, but the estimator ought to be in the same domain as the word recognition accuracy to estimate the SRTs and produce similar psychometric curves as human listeners. Given the non-linearity introduced by the M-measure, a mapping function is required to estimate WERs. The function used to map the M-measure to WER depends on the acoustic model and decays exponentially according to the following equation:

$$WER(\mathscr{M}) = A * e^{k*\mathscr{M}}. \qquad \text{(Eq. 3)}$$

The initial value A and the decay rate k were calculated on a cross-validation set comprised of utterances spoken by a speaker not included in the training set mixed with the same noise maskers described in the following section. Additionally, an upper boundary of 100 (the highest possible error rate) was imposed.

Features and deep neural network

The ASR system is trained with amplitude modulation filterbank (AMFB) features that are based on regular mel spectrograms with 40 frequency channels, which are processed with modulation filters in the range from 5 to 20 Hz (Moritz *et al.*, 2015). They were chosen since the explicit coding of temporal modulations increased model performance, especially for the across-frequency shifted speech-shaped noise (AFS-SSN) masker previously (Spille *et al.*, 2018). AMFB features were used as input to a DNN, which has the purpose of mapping the acoustic observations to phoneme probabilities. A fully-connected network (referred to as DNN) with six hidden layers and 2048 hidden (sigmoid) units was selected to compare this work and the previous SI prediction model from (Spille *et al.*, 2018). The network was trained to classify context-dependent triphones; every phone is modeled with three HMM states except for silence, which uses five states.

The training of the DNN described above was done in up to 20 epochs (stopping when the relative improvement was lower than 0.001). The starting learning rate was 0.008 (halving it every time the relative improvement was lower than 0.01). A soft-max layer of approximately 2000 units was attached to the output to produce the most likely posterior probabilities of each context-dependent HMM state.

An in-house corpus of 10 hours of speech from 20 speakers (10 male, 10 female) with the syntactical structure of Oldenburg Sentence Test (see next section) was selected as

a starting point to train the ASR system; sentences from the original speaker were not contained in the training set.

The training sets comprise of the clean data mixed with random parts of each of the eight different maskers (as described below) at random uniformly distributed SNRs ranging from -10 dB to 20 dB, resulting in 80 h of speech material. Two training sets were created which are based on noises created from a male or female voice. The test set to evaluate the ASR system was created by mixing eight random sentences from the speech material with parts of the respective masker for each of the 400 SNR values uniformly distributed between -30 dB and 20 dB to sample the whole psychometric function.

SPEECH MATERIAL, MASKERS AND SUBJECTIVE DATA

In this section, the speech material for both training and testing is described, the noise signals are introduced, and details about the listening tests from which the human SRTs were calculated by Schubotz *et al.* (2016) are provided.

Matrix test

Both the listening and ASR tests were performed using the *Oldenburger Satztest* (OLSA) (Wagener *et al.*, 1999), which is a matrix sentence test. It consists of 120 utterances produced by one speaker. Target sentences derived from a vocabulary of 50 words equally divided into five categories. For a review of matrix tests in several languages, please refer to (Kollmeier *et al.*, 2015). Each five-word sentence follows the fixed structure: <name><verb><number> <adjective><object>, e.g. *"Peter kauft sechs nasse Tassen"* ("Peter buys six wet cups"). Despite being grammatically correct, these sentences have no semantic context. Moreover, all combinations of words from each of the five categories can occur; therefore, predicting a sentence from previous ones is not possible.

Noise maskers

In the study carried out by Schubotz *et al.* (2016), a set of eight background maskers was created to evaluate the effect of energetic, amplitude modulation and informational masking on SI. We took this benchmark to evaluate our SI prediction model, focusing on four speech maskers described in the following.

First, a stationary SSN with the same long-term spectrum as the International Speech Test Signal (ISTS; Holube *et al.*, 2010) was used. Second, a sinusoidally amplitude-modulated SSN (SAM-SSN) was produced by adding an 8 Hz temporal modulation. The third masker was generated by multiplying the Hilbert envelope of a broadband speech signal with the SSN (BB-SSN). For the fourth, named across-frequency shifted SSN (AFS-SSN), the SSN was filtered in 32 frequency channels; subsequently, every four adjacent channels were multiplied with a different random time section of the

Hilbert envelope used for BB-SSN[‡].

To test the influence of same- or different-gender maskers, all maskers have a male and female version to match the long-term spectrum of the respective gender. Since the original ISTS contains female voices only, Schubotz *et al.* (2016) produced a male version of ISTS via the STRAIGHT algorithm (Kawahara *et al.*, 2008) to match its long-term spectrum.

Listening tests

To benchmark the performance of the proposed SI predictor, we compare the results from the listening tests performed in (Schubotz *et al.*, 2016). These experiments consisted of characterizing SI as a function of the SRT extracted from the adaptive procedure proposed by Brand and Kollmeier (2002). Eight normal-hearing participants (ages between 23-34) participated who were not previously exposed to the speech task; their hearing thresholds for pure tones did not exceed 20 dB at frequencies between 125 Hz and 8 kHz. During testing, the participants attended 20 OLSA sentences with an initial SNR of 0 dB; then, the SNR varied depending on the intelligibility measurement of the previous sentence. The procedure is set to determine the SNRs at which listeners correctly understand 50% (SRT) and 80% (SRT_{80}) of presented words. Each SRT resulted from a different set of sentences; in other words, each participant listened to 40 sentences per noise condition. Finally, the SRTs were averaged across the listeners to obtain the final SRT and SRT_{80} values, which are used to trace the psychometric function of the listeners (described entirely by the SRT and its slope). The slope of the psychometric function was estimated via a maximum-likelihood estimator (Brand and Kollmeier, 2002) with the 40 responses for each listener and masker.

RESULTS

Because the modeling approach presented in this paper is based on estimating the WER (cf. Figure 1), we first analyze if the error rate from the ASR is related to the predicted one based on the M-measure (Figure 2), where each data point corresponds to the error rate for eight matrix sentences.

While the WER with the SSN masker is overestimated and AFS-SSN data is underestimated, we observe a clear relationship between estimated and ASR WER for each masker. Additionally, the mapping is most sensitive at lower word error rates as the mapping function is a decaying exponential constrained to an upper boundary of 100.

To quantify the model performance, we compare the psychometric functions of the listeners to the approach using ASR generated transcripts and the proposed approach (Figure 3).

[‡]resulting in eight different adjacent modulation bands

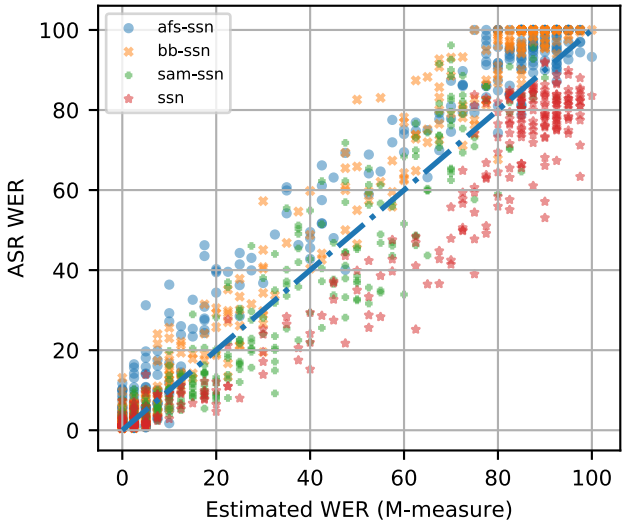


Fig. 2: Relation of estimated WER derived from phoneme probabilities and measured ASR WER for four different maskers.

When comparing both models, the overestimated WER values for the SSN result in a shift to a lower SNR for the new model (top left panel in Figure 3), while the shift to higher SNRs for the AFS-SSN and BB-SSN maskers is a reflection of an underestimation of the WER (noticeable in Figure 2).

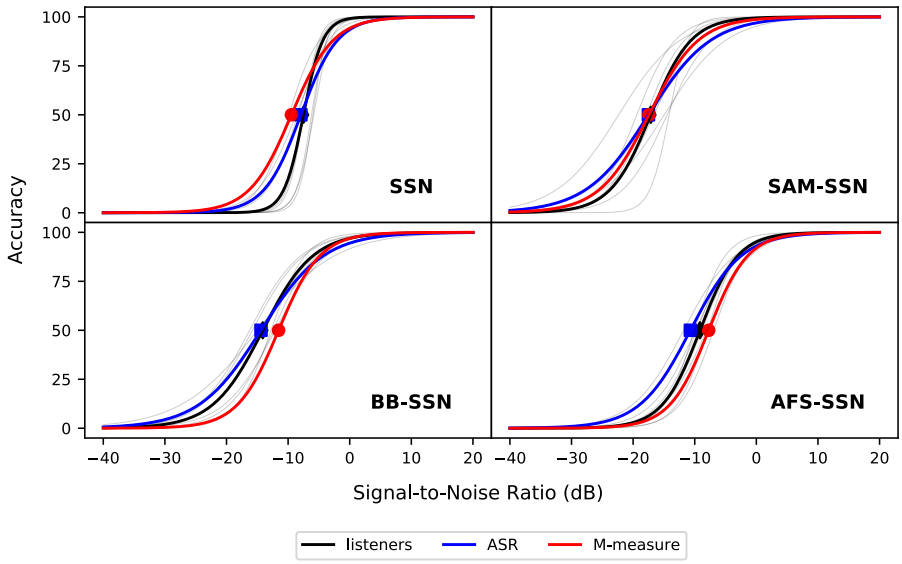


Fig. 3: Psychometric functions or NH listeners (mean in black, individual curves in gray), the ASR-based model that used *a priori* knowledge in the form of transcripts (blue) and the proposed model that estimates the SRT from phoneme probabilities without using the transcript (red).

(A) Average SRTs of listeners and proposed model in dB

	SSN	SAM-SSN	BB-SSN	AFS-SSN
female	-7.5	-17.1	-14.1	-9.2
AMFB-M-Measure	-9.5	-20.0	-14.3	-9.1
male	-8.2	-17.7	-14.9	-9.3
AMFB-M-measure	-10.0	-18.5	-13.0	-7.2

(B) Root-mean-squared error of observed and predicted SRTs

	baseline models				Spille et al	This work
	SII	ESII	STOI	mr-sEPSM	AMFB	AMFB-M-measure
male	6.0	2.2	8.7	4.0	1.6	1.7
female	6.0	2.8	8.5	8.5	0.9	1.7
avg.	6.0	2.5	8.6	6.2	1.3	1.7

Table 1: (A) SRTs of normal-hearing listeners and the corresponding SRT prediction for the proposed ASR-based models. (B) SRT prediction error for baseline models, the label-based previous model and the proposed approach. The rows *female* and *male* correspond to the maskers derived from speech from the female or male speaker, respectively (cf. section on noise types)

These effects, however, seem to be small and provide a good match with the human data (gray lines in the figure); this is also reflected by the RMSE between predicted and observed SRTs, calculated for four baseline models, as shown in Table 1. In Table 1(A), the average SRTs of listeners are compared with the ones yielded by the proposed model, referred to as AMFB-M-measure. The models trained on the female noise maskers are matched to the corresponding female normal-hearing SRTs; likewise, the male results were compared to a model trained on male noise maskers. In both setups, SSN and SAM-SSN, the predicted SNRs were lower than the observed ones, whereas the opposite behavior occurs with the BB-SSN and AFS-SSN noise maskers.

We compute the root-mean-square error (RMSE) between observed and predicted SRTs to measure the precision of the proposed model in all noise conditions shown in Table 1(B). Among the baseline models, ESII yields the lowest error with an average of 2.5. Both DNN-based models show lower RMSE than the previous models. The model from Spille *et al.* (2018), with an average RMSE of 1.3 dB, remains the closest to the human observed SRTs; the female version produces almost half the error as the male counterpart; the same pattern is observed in the mr-sEPSM model. For SII, STOI, and our proposed model, the error difference between the genders is very small.

Note that the DNN-based model was trained as a gender-independent speech with gender-dependent maskers; thus, it produces consistent predictions for both male and female maskers.

DISCUSSION

The proposed model produces accurate predictions while it does not require clean speech reference or the transcript of the utterance that is evaluated. Moreover, because the model was trained on speech data without semantic context, it could potentially generalize to other speech tests. However, in contrast to existing models, it requires a relatively large amount of training data in the range of 80 hours, and it is unclear if predictions can be obtained for SI across languages. It might, therefore, be challenging to apply this approach to low-resource languages without optimizing the training procedure. In related work targeting listening effort, the method of using phoneme probabilities was, however, successful for predicting the listening effort of the German matrix sentence test with fine-tuning using English data (Huber *et al.*, 2018), which indicates that across-language prediction could potentially work if the languages are phonetically not vastly different. An advantage of the proposed model is that it produces absolute predictions for the SRT, again in contrast to the baseline models that are normalized using the prediction for a reference condition, in this case, the stationary SSN.

In the future, the approach could be used as a model-in-the-loop (i.e., it could monitor and estimate the SI resulting from different processing strategies and settings in hearing aids and select the strategy that most likely maximizes SI). However, this would require the prediction of SI for hearing-impaired listeners, while the current model implementation has only been tested for normal-hearing listeners. A simple strategy to take into account the hearing loss that is reflected in a listener's audiogram would be to add frequency-dependent noise to mask the signal properties that are not accessible to the individual listener. Optimally, the corresponding calculations should be carried out on mobile hearing aid hardware. In previous research, we have shown that running at least one feed-forward neural network can be achieved on a hearing aid co-processor (Castro Martinez *et al.*, 2019). However, more efficient net topologies such as time-delay neural networks (Peddinti *et al.*, 2015) need to be considered in the future, as well as taking into account hearing loss, given that a comparison of different processing algorithms requires at least two networks can be used simultaneously.

SUMMARY

This paper explored a modeling approach for SI prediction based on ASR without the requirement of a transcript in the model. It was shown that the model is suitable to predict the SRT of normal-hearing listeners with very similar accuracy to the prediction performance of an ASR-based model that used *a priori* knowledge in the form of transcripts. This achievement was enabled by measuring the degradation of frame-level phoneme representations obtained from a DNN. Our model also

outperforms four established baseline models in four masker types with different types of modulation. Future research should focus on a wider range of maskers and take into account the computational complexity of the approach, which needs to be considered for real-time applications of SI prediction. As the approach was only tested for normal-hearing listeners, we also need to investigate if the model can be extended for predicting SI of (aided) hearing-impaired listeners, which would be a significant step towards using it as model-in-the-loop for real-time optimization in assistive hearing.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177 - Project ID 390895286 and the SFB/TRR 31/3' The active auditory system,' Transfer Project T01).

REFERENCES

ANSI, S3 22-1997 (**1997**), "Methods for calculation of the speech intelligibility index," American National Standard Institute.

Barker, J. and Cooke, M. (**2006**), "Modelling speaker intelligibility in noise," Speech Commun.

Brand, T. and Kollmeier, B. (**2002**), "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am., **111**(6), 2801–2810.

Castro Martinez, A. M., Gerlach, L., Payá-Vayá, G., Hermansky, H., Ooster, J., and Meyer, B. T. (**2019**), "DNN-based performance measures for predicting error rates in automatic speech recognition and optimizing hearing aid parameters," Speech Commun., **106**, 44–56.

Ewert, S. D. and Dau, T. (**2000**), "Characterizing frequency selectivity for envelope fluctuations." J. Acoust. Soc. Am., **108**(3), 1181–96.

Hermansky, H., Variani, E., and Peddinti, V. (**2013**), "Mean temporal distance: Predicting ASR error from temporal properties of speech signal," Proc. IEEE ICASSP, 7423–7426.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**), "Development and analysis of an International Speech Test Signal (ISTS)." Int. J. Audiol., **49**(12), 891–903.

Huber, R., Krüger, M., and Meyer, B. T. (**2018**), "Single-ended prediction of listening effort using deep neural networks," Hearing Res., **359**, 40–49.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (**2008**), "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," Proc. IEEE ICASSP, 3933–3936.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., and Wagener, K. C. (**2015**), "The multilingual matrix test: Principles, applications,

and comparison across languages: A review," Int. J. Audiol., **54**(sup2), 3–16.

Moritz, N., Anemüller, J., and Kollmeier, B. (**2015**), "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," IEEE Trans. Audio Speech Lang. Process., **23**(11), 1926–1937.

Peddinti, V., Povey, D., and Khudanpur, S. (**2015**), "A time delay neural network architecture for efficient modeling of long temporal contexts," Proc. International Speech Communication Association.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (**2011**), "The Kaldi speech recognition toolkit," Proc. IEEE ASRU.

Rhebergen, K. S. and Versfeld, N. J. (**2005**), "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners." J. Acoust. Soc. Am., **117**(4), 2181–2192.

Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. (**2015**), "Matrix sentence intelligibility prediction using an automatic speech recognition system." Int. J. Audiol., 1–8.

Schubotz, W., Brand, T., Kollmeier, B., and Ewert, S. D. (**2016**), "Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features," J. Acoust. Soc. Am., **140**(1), 524–540.

Spille, C., Ewert, S. D., Kollmeier, B., and Meyer, B. T. (**2018**), "Predicting speech intelligibility with deep neural networks," Comput. Speech Lang., **48**, 51–66.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**), "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process., **19**(7), 2125–2136.

Wagener, K., Brand, T., and Kollmeier, B. (**1999**), "Development and evaluation of a German sentence test part III: Evaluation of the Oldenburg sentence test," Zeitschrift Fur Audiologie, **38**, 86–95.

Evaluation of a notched-noise test on a mobile phone

PETTERI HYVÄRINEN[1,*], MICHAL FERECZKOWSKI[1,2] AND EWEN N. MACDONALD[1]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Institute of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, DK-5230 Odense, Denmark*

The ability to conduct hearing tests and estimate auditory function at home or in the workplace can be useful for screening or longitudinal studies and allow the collection of diagnostic data from tests that are too time consuming to be feasible in the clinic. However, moving away from an acoustically controlled environment may influence the results of a test (e.g., through masking due to higher levels of background noise), increasing the uncertainty in the test measurements. In this study, 9 normal-hearing participants completed a notched noise masking experiment with 3 different experimental setups: in a psychoacoustic test booth with a standard laboratory PC; in a psychoacoustic test booth with a mobile device; and in a quiet office room with a mobile device. The accuracy and reliability of the mobile implementation was compared to results obtained with the laboratory setup. The effect of the test environment was investigated by comparing the mobile platform results between booth and office. The mobile device implementation corresponded well with the laboratory results for a notch width of zero, but showed a systematic bias when the width of the notch was increased. The reliability of the mobile implementation was comparable to the laboratory. Moving outside the sound-insulated booth did not affect the mobile platform results.

INTRODUCTION

Mobile phones can potentially be utilized in psychoacoustic field experiments, but it is not known how current methods translate to mobile equipment and acoustically less controlled environments. In this study, we evaluated whether results from a notched-noise test (Weber, 1977; Patterson, 1976; Moore and Glasberg, 1990) are similar when conducted with a mobile phone in a regular office environment, compared to a standard laboratory setting in a listening booth with a desktop computer, soundcard, and high-quality headphones.

METHODS

Notched-noise test

Detection thresholds for a 2 kHz tone, presented simultaneously with a broad-band noise masker, were determined using a two-alternative forced-choice (2-AFC)

*Corresponding author: pehyv@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

paradigm. The masker was a white noise band between 100 Hz and 10 kHz, with a symmetric notch around the target frequency, meaning that the masker spectrum had a gap between $f_T - \Delta f$ and $f_T + \Delta f$, where $f_T = 2$ kHz is the frequency of the target tone. Following the common practice in notched-noise experiments, the gap width is expressed as a normalized value with respect to the target frequency: $g = \frac{\Delta f}{f_T}$. The spectral level of the masker was held constant at 30 dB SPL/Hz. The noise was generated via an inverse Fourier transform, where the frequency components had equal amplitude and random phase in the frequencies containing masker energy, and zero amplitude within the gap frequencies and outside the masker frequency range.

The stimuli consisted of either the masker alone (non-target interval) or the masker and the target tone simultaneously (target interval). The 2-AFC task was to indicate the target interval. The length of the stimuli was 300 ms, and the inter-stimulus-interval was 400 ms. Sounds were presented monaurally to the left ear for each participant.

Expressing the detection threshold as a function of gap width results in a threshold curve which is monotonously decreasing. In other words, as the gap width is increased, the target tone becomes easier to detect and the threshold is lower. This is due to the frequency selectivity of hearing, and reflects the concept of auditory filters.

Grid tracking method

Traditionally, the detection thresholds for the target tone are determined on individual experimental runs for each masker gap width of interest using, for example, a transformed up-down method with a fixed gap width, and varying only the target level. However, in this approach, each experimental run is usually initialized so that the target level is well above the detection threshold. When this procedure is repeated for many gap widths, in the end a considerable proportion of trials is spent far away from the threshold.

To shorten the time needed for estimating a full threshold curve, the grid method by Fereczkowski (2015) takes advantage of the monotonic behavior of the threshold curve, thus increasing the proportion of points sampled close to the threshold curve. In the grid method, the experimental track starts at zero gap width and at a target level well above the threshold, similarly to a transformed up-down track. After the threshold at zero gap width is determined, instead of restarting the track for the next gap width, the grid method simply moves in the positive x-axis direction until the threshold curve is crossed (i.e., increasing the gap width while keeping the target level fixed). When the threshold in the horizontal direction is found, the method continues downwards. The tracking procedure of the grid method is illustrated in Fig. 1. Thus, the grid method alternates between adjusting the target level and the gap width within a single experimental run. Just as in a transformed up-down method, different tracking rules can be implemented, such as 3-down–1-right, and the choice of parameters determines the point along the psychometric function (e.g., the 79.4% detection threshold for a

3-down–1-up track) which the threshold curve approximates.

In the current study, the threshold at zero gap width was first determined with a 3–down-1–up staircase procedure, with four reversals using a 6 dB step size, followed by six reversals using a 3 dB step size. The threshold at zero gap width was determined as the average of the last six reversals. This threshold is later in the text referred to as the *tone–in–noise-threshold*. Then, experimental run continued with a 3–down-1–right grid procedure at the last reversal. The run was terminated when either the maximum gap width of 0.5 or the minimum target level of 30 dB SPL was reached. The information about the shape of the threshold curve was represented by the -10 dB gap width, which refers to the interpolated gap width at which the threshold is 10 dB less than at zero gap width. One experimental block consisted of three repeated experimental runs.

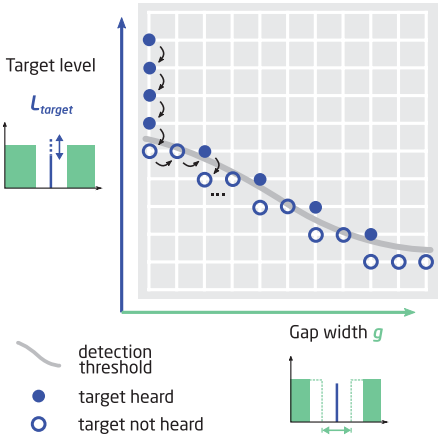


Fig. 1: Illustration of one run of the grid method. The track starts with a moderately high target level at zero gap width. Level is decreased until threshold is reached, after which gap is increased until the target can be heard again.

Hardware platforms

The two hardware platforms compared in the current study were a standard personal computer equipped for psychoacoustic research (referred to as *PC* later in the text), and an Apple iPhone 7 mobile phone (*Phone*).

On the PC, stimuli were generated with Matlab, D/A-converted by an external Fireface UCX soundcard, amplified by a Sound Performance Lab Phonitor mini headphone amplifier, and presented via Sennheiser HD-650 circumaural headphones.

On the Phone, stimuli were generated with iOS's vDSP library, included in the Accelerate framework, and presented via Apple EarPods connected to the phone's Lightning port. The mode of the `AVAudioSession` object used for playback of the stimuli was set to `measurement` in order to avoid any sound processing by the

operating system.

On both systems, a 44.1 kHz sampling rate was used. The frequency responses of the two systems were recorded using a B&K head and torso simulator (HATS, type 4128-C) with a artificial pinna and ear canal (DZ-9769) and a 2-cc coupler. The frequency response of each system was compensated for by digitally inverse filtering the generated stimuli on the device. The spectra of sample stimuli, presented through the two systems after compensation are shown in Fig. 2.

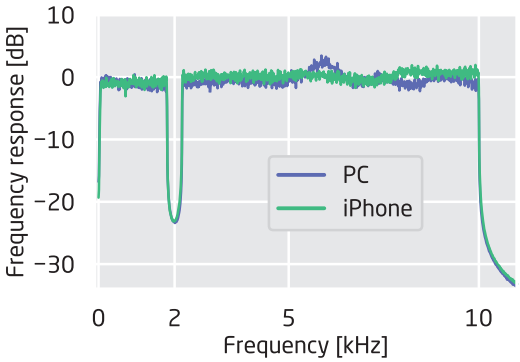


Fig. 2: Two spectra of a noise band (10–10000 Hz) with a spectral gap of 0.1 at 2 kHz, presented via the two different playback systems after compensating for differences in frequency responses by inverse filtering.

Experimental setup

Nine young, normal-hearing subjects (seven male, two female) took part in the study.

The main interest of the current study was to evaluate whether the hardware platform or the environment had any effect on the obtained results. Therefore, the following three conditions were included in the study:

- PC in booth
- Phone in booth
- Phone in room,

where *PC* and *Phone* refer to the hardware platforms (Section 2.3), and *booth* and *room* refer to a double-walled acoustically treated listening booth, and a regular quiet office room, respectively. The *Phone in room*-condition was repeated on another day to get an estimate of the test-retest variability for the same equipment and environment, and prior to the actual experiment, all participants completed one *PC in booth* -training block. Thus, in total all subjects completed one training block and four experimental blocks. Each block consisted of three grid runs, as described in

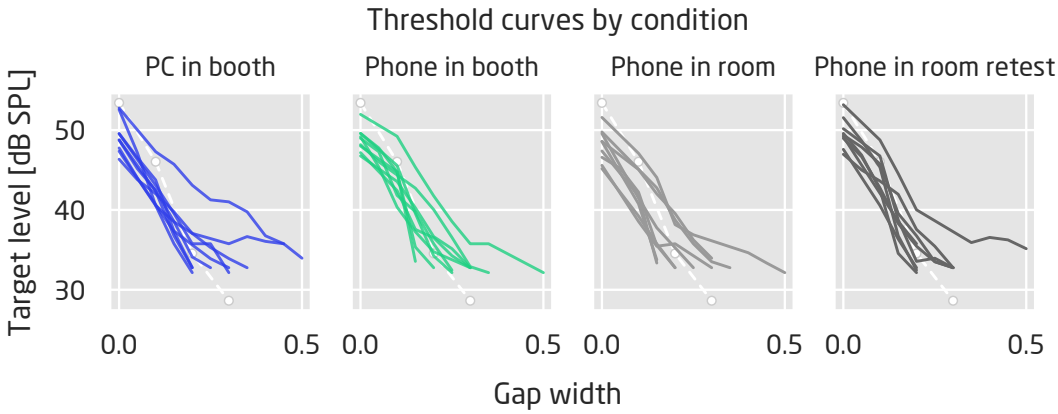


Fig. 3: The lines show the individual threshold curves. The white dashed line is the same in all figures and shows data from Rosen and Baker (1994)

section 2.2. The individual threshold curves from each run were averaged and a single curve was used in the analysis.

RESULTS AND DISCUSSION

Figure 3 shows the threshold curves for each subject, and the grand average taken across all subjects. By visual inspection, it is clear that the curves are very similar in all four cases. For comparison, each panel shows the data from Rosen and Baker (1994), for the same masker type, i.e. symmetric gap and 30 dB SPL/Hz spectral level. The current results are systematically lower than in (Rosen and Baker, 1994), which could be due to differences in the timing between masker and target; in the current study, the masker and target started at the same time, whereas in Rosen and Baker (1994) the masker started slightly before the target. Therefore, it is possible that the lower threshold is caused by an onset cue which was not present in the earlier study. Also, the grand average curves in the current study appear to flatten as the gap is widened, but this is due to limiting the target level to values above 30 dB SPL in the current study. Therefore, at wider gaps there are less datapoints taken into the average, which skews the result.

To investigate the differences between platforms and environments, the four experimental cases were compared in a pairwise manner with a Bland-Altman plot (Bland and Altman, 1986), which is a method for assessing the agreement between two measures. Figure 4 shows the pairwise Bland-Altman plots for the tone-in-noise-thresholds. The largest mean difference between two conditions is 1.4 dB between *Phone in room* and *Phone in room retest*. The 95% limits of agreement ($\pm$ 1.96 $\times$ standard deviation) indicate the estimated range within which 95% of the individual test-retest differences are expected to lie. The widest limits of agreement ($\pm$ 4.6 dB) are between *Phone in booth* and *Phone in room* conditions.

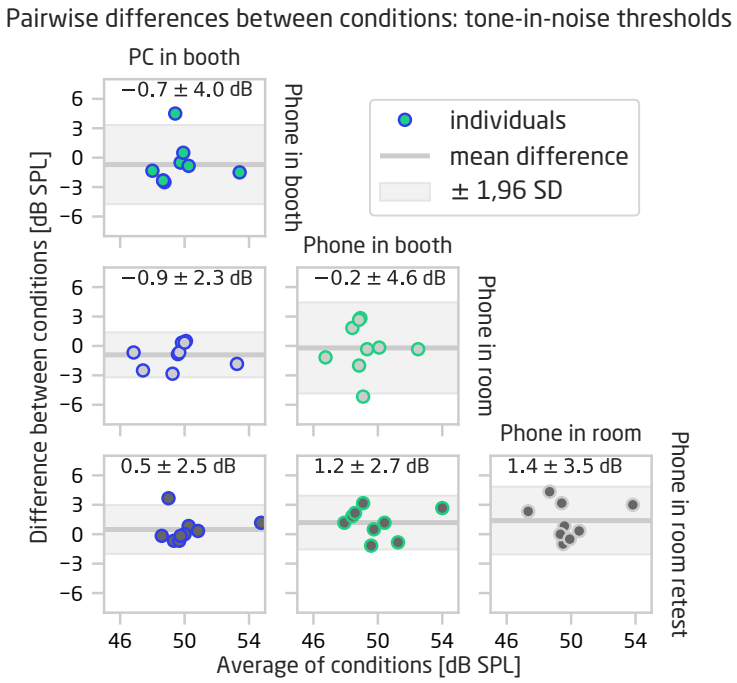


Fig. 4: Pairwise Bland–Altman plots visualizing test–retest repeatability for tone-in–noise thresholds (threshold at zero gap) between two conditions. Horizontal line shows the average difference in thresholds between two conditions, and the shaded area illustrates the 95% limits of agreement (mean ±1.96 SD) between the two conditions.

There are no clear patterns in the data, and in fact the test-retest accuracy is within expected limits for all conditions, as was verified by a Monte-Carlo simulation (1000 rounds) of the tone-in-noise-threshold determination, using the same experimental parameters as in the current study, and reference values for the estimators from Schlauch and Rose (1990). Assuming the threshold estimate to be a normally-distributed variable $X \sim N(\mu, \sigma^2)$, the standard deviation of a single threshold estimate per simulated results is 2 dB. The tone-in-noise-threshold was calculated as an average over three repeated runs:

$$X_{avg} = \frac{1}{3}\sum_{i=1}^{3} X = \sum_{i=1}^{3} \frac{1}{3} X \sim N\left(\sum_{i=1}^{3} \frac{1}{3}\mu, \sum_{i=1}^{3} (\frac{1}{3}\sigma)^2\right) = N\left(\mu, \frac{3}{9}\sigma^2\right), \qquad \text{(Eq. 1)}$$

and so the variance of the tone-in-noise-threshold is expected to be $\sigma^2_{avg} = \frac{3}{9} \cdot 2^2 = \frac{4}{3}$. Looking at the difference between two conditions, the difference in thresholds is expected to be also a normally distributed variable: $X_{diff} = X_1 - X_2 \sim N(0, 2\sigma^2_{avg}) = N(0, 2 \cdot \frac{4}{3}) = N(0, \frac{8}{3})$. If a sample ($n = 9$, the number of participants in the study) is drawn from this distribution, the 95%

confidence interval for the limits of agreement would be 2.16 – 6.13 dB. Thus, it is expected that with the current experimental design, the observed spread is not limited by the equipment or the environment, as in all cases the limits of agreement are smaller than those suggested by the simulations.

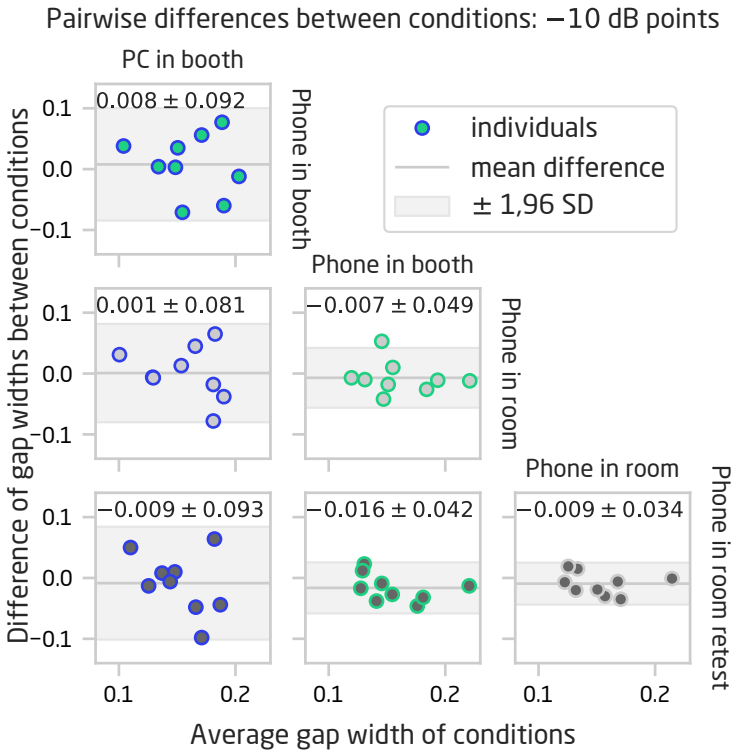


Fig. 5: Pairwise Bland–Altman plots visualizing test–retest repeatability for the −10 dB gap width, i.e. the interpolated gap width at which the threshold is 10 dB lower than at zero gap. Horizontal line shows the average difference in −10 dB points between two conditions, and the shaded area illustrates the 95% limits of agreement (mean ±1.96 SD) between the two conditions.

For the −10 dB gap widths (Fig. 5), the results are more mixed; the phone conditions show clearly better agreement with each other than with the *PC in booth*. The source of this discrepancy between PC and phone conditions is not clear. However, if the poorer agreement would be due to differences in, for example, frequency responses of the two systems, the test-retest differences should show a systematic error. However, since the average difference is still close to zero, it appears that the differences are driven by inter-individual variability. One factor that could play a role in these results is the difference in headphones. On the PC, the headphones were circumaural, whereas the EarPods are in-ear earbuds. Thus, although both systems were calibrated with the same HATS setup, it is plausible that individual differences in outer ear shape could affect the spectral content and thereby the shape of the threshold curve. For

example, changes in the ear canal resonances caused by the partial insertion of the EarPod, or filtering by the pinna, could tentatively explain the observed differences. Further experiments are planned to investigate the effect of the transducer choice.

CONCLUSIONS

Conducting the notched-noise test on a mobile phone, and outside a sound-insulated listening booth did not introduce any bias to the psychoacoustic estimates of hearing. Tone-in-noise-threshold estimation was only limited by the experimental design with no differences in test-retest results between conditions. For the estimates of auditory frequency resolution, the larger spread (but no systematic bias) of test-retest differences between PC and phone conditions may hint at an individual effect of the headphone design on the acoustic coupling with the outer ear.

REFERENCES

Bland, J.M. and Altman, D. (**1986**). "Statistical methods for assessing agreement between two methods of clinical measurement." Lancet, **327**(8476), 307–310.

Fereczkowski, M. (**2015**). *Time-efficient behavioral estimates of cochlear compression*. Ph.D. thesis, Technical University of Denmark.

Moore, B. and Glasberg, B. (**1990**). "Derivation of auditory filter shapes from notched-noise data." Hearing Res., **47**, 103–138.

Patterson, R.D. (**1976**). "Auditory filter shapes derived with noise stimuli." J. Acoust. Soc. Am., **59**(3), 640–654. ISSN 0001-4966. doi:10.1121/1.380914.

Rosen, S. and Baker, R.J. (**1994**). "Characterising auditory filter nonlinearity." Hearing Res., **73**(7), 231–243.

Schlauch, R.S. and Rose, R.M. (**1990**). "Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency." J. Acoust. Soc. Am., **88**(2), 732–740. ISSN 0001-4966. doi:10.1121/1.399776.

Weber, D.L. (**1977**). "Growth of masking and the auditory filter." J. Acoust. Soc. Am., **62**(2), 424–429. ISSN NA. doi:10.1121/1.381542.

Using a deep neural network to speed up a model of loudness for time-varying sounds

JOSEF SCHLITTENLACHER,[1,*] RICHARD E. TURNER[2] AND BRIAN C.J. MOORE[1]

[1] *Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge, CB2 3EB, UK*

[2] *Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK*

The "time-varying loudness (TVL)" model calculates "instantaneous loudness" every 1 ms, and this is used to generate predictions of short-term loudness, the loudness of a short segment of sound such as a word in a sentence, and of long-term loudness, the loudness of a longer segment of sound, such as a whole sentence. The calculation of instantaneous loudness is computationally intensive and real-time implementation of the TVL model is difficult. To speed up the computation, a deep neural network (DNN) has been trained to predict instantaneous loudness using a large database of speech sounds and artificial sounds (tones alone and tones in white or pink noise), with the predictions of the TVL model as a reference (providing the "correct" answer, specifically the loudness level in phons). A multilayer perceptron with three hidden layers was found to be sufficient, with more complex DNN architecture not yielding higher accuracy. After training, the deviations between the predictions of the TVL model and the predictions of the DNN were typically less than 0.5 phons, even for types of sounds that were not used for training (music, rain, animal sounds, washing machine). The DNN calculates instantaneous loudness over 100 times more quickly than the TVL model.

INTRODUCTION

Glasberg and Moore (2002) described a model for predicting the loudness of time-varying sounds: the time-varying loudness (TVL) model. A block diagram of the model is shown in Figure 1. The model includes a sequence of stages to simulate the transmission of sound to the eardrum (Shaw and Vaillancourt, 1985), the transmission of sound through the middle ear (Aibara *et al.*, 2001), the frequency analysis that takes place in the cochlea (resulting in an auditory excitation pattern) (Glasberg and Moore, 1990), the creation of a specific loudness pattern (including the effects of the compression that occurs in the cochlea) (Moore and Oxenham, 1998), and summation of specific loudness across characteristic frequencies (CFs) (Zwicker and Scharf, 1965) to give instantaneous loudness. Within the model, frequency is transformed to the ERB_N-number scale, which has units Cams (Glasberg and Moore, 1990; Moore, 2012). This is a perceptually relevant scale comparable to a scale of distance along the basilar membrane. Instantaneous loudness is assumed to be an intervening

*Corresponding author: js2251@cam.ac.uk; currently at the Department of Neurosciences, University of Cambridge.

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

variable, not available to conscious perception, although it has been shown that certain cortical regions show activity that is correlated with the instantaneous loudness calculated using the model (Thwaites *et al.*, 2016).

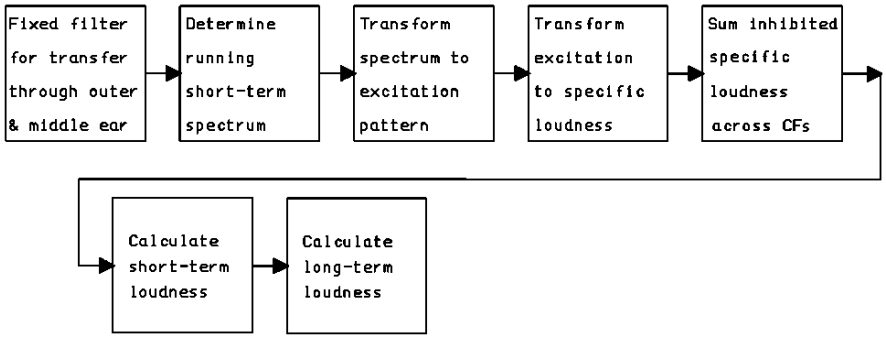


Fig. 1: Block diagram of the TVL model of Glasberg and Moore (2002).

The instantaneous loudness is smoothed over time to calculate short-term loudness, which is meant to represent the loudness of a short piece of sound such as a single word in a sentence or a note in a piece of music. The short-term loudness is itself further smoothed over time to calculate the long-term loudness, which is meant to represent the overall loudness of a longer stretch of sound, such as a whole sentence or a musical phrase. The peak value of the long-term loudness gives good predictions of judged loudness for a variety of sounds, including amplitude compressed speech (Moore *et al.*, 2003), sonic booms and impact sounds (Marshall and Davies, 2007), machinery sounds (Rennies *et al.*, 2015), and speech processed to have increased or decreased dynamic range (Zorila *et al.*, 2016). The model forms the basis of a proposed ISO standard (ISO 532-3, 2019), although the model used in the standard includes additional stages to account for the way that loudness is combined across ears (Moore and Glasberg, 2007).

Computationally, the most time-consuming stage of the TVL model is the calculation of the excitation pattern, which is estimated from the short-term spectrum of the sound and is used to calculate instantaneous loudness at 1-ms intervals. The excitation pattern is defined as the output of the auditory filters as a function of centre frequency (Moore and Glasberg, 1983). It is estimated by calculating the outputs of an array of level-dependent auditory filters in response to each component of the input signal (after outer- and middle-ear filtering) (Glasberg and Moore, 1990; Moore *et al.*, 1997). The time taken to calculate instantaneous loudness makes it difficult to implement the TVL model in real time. This paper describes the development and training of a deep neural network (DNN) to speed up the computation of instantaneous loudness, allowing real-time implementation. The DNN was trained to predict instantaneous loudness using a large database of speech sounds and artificial sounds (tones alone, bandpass filtered and notched noises, and tones in white or pink noise), with the predictions of the TVL model as a reference (providing the "correct" answer, specifically the loudness level in phons).

STRUCTURE AND TRAINING OF THE DNN

The stimuli used for training had a sampling rate of 16 kHz. Spectra were initially calculated using a 1024-point discrete Fourier Transform, with successive windows being shifted by 560 samples. Then bins were grouped to form 61 bands with centre frequencies up to 8 kHz, with one bin per band for centre frequencies up to 0.2 kHz and 1/9th-octave wide bands for higher centre frequencies. The limit of 8 kHz was chosen due to the sampling rate of the training material. The magnitude of the spectrum was expressed in decibels.

Both accuracy and computation speed were important considerations when choosing the design of the DNN. The DNN was a multilayer perceptron that consisted of an input layer with 61 units (corresponding to the 61 frequency bands), three hidden layers with 150 units each, and a single output unit with linear activation. The output of the DNN was a single loudness level in phon. This was chosen because of its similarity to the input scale. Both scales range roughly from 0 to 110, and the just noticeable difference in loudness is roughly constant on these scales. This facilitated the DNN in developing the mapping from input to output without the need for scale transformations. Simple "rectified linear unit" activations (Nair and Hinton, 2010) were used. Alternative architectures were also considered. Convolutional neural networks did not achieve the same accuracy, probably because the input scale used (logarithmic frequency) did not allow the network to simulate filters that were valid over the whole range of the ERB_N-number scale that is used in the TVL model.

The DNN was optimized with regard to the root-mean-square (RMS) error from the predictions of the TVL model. The Adam optimizer (Kingma and Ba, 2014) was used with its default parameters. All weights were initialized randomly. Three sets of training data were used. First, 500,000 spectra were calculated from the LibriSpeech corpus (Panayotov *et al.*, 2015) from the "clean" development set. The sounds were scaled to have an RMS level of 60 dB SPL. Second, about 700,000 pure tones with levels ranging from 15 to 110 dB SPL and various levels of background noise (from inaudible up to 10 dB below the level of the pure tone) were generated. Third, about 500,000 spectra of bandpass filtered noises and noises with spectral notches were generated. They had various overall levels, bandwidths, notch widths and spectral gradients. To check for "over-fitting", the performance of the DNN was assessed after training for 220, 1000 and 5000 epochs, where an epoch is a complete pass over the entire dataset one time.

ASSESSMENT OF THE DNN

Predictions for speech and everyday sounds

Loudness was predicted for two further sets of data from the LibriSpeech corpus, "clean" set and "other" set (not used for training). Each of them consisted of 500,000 spectra and they were scaled to have an RMS level of 60 dB SPL. Loudness was also predicted for 250,000 spectra derived from the ESC-50 corpus (Piczak, 2015). This corpus contains 50 categories of environmental sounds, for example rain, animals,

aircraft, keyboard typing or a washing machine. The sounds were again scaled to have an RMS level of 60 dB SPL. Finally, loudness was predicted for 100,000 spectra from 20 popular songs of the 1960s, which were scaled to have an RMS level of 70 dB SPL. Table 1 shows the RMS error in phons between the predictions of the TVL model and the predictions of the DNN after training for 220, 1000, and 5000 epochs. After 1000 epochs, the RMS error was below 0.5 phons for all classes of sounds. After 5000 epochs the RMS error increased slightly for the LibriSpeech "other" and ESC-50 sounds, which is a sign of "over-fitting". Therefore, in what follows, we focus on the results achieved after training for 1000 epochs.

	Number of epochs		
Test material	220	1000	5000
LibriSpeech "clean"	0.35	0.27	0.28
LibriSpeech "other"	0.55	0.45	0.47
ESC-50	0.56	0.45	0.47
Songs from the 1960s	0.38	0.35	0.31

Table 1: RMS error in phons between the predictions of the TVL model and the predictions of the DNN for sounds not used for training. The error did not vary systematically with the predicted loudness level and the errors had a Gaussian distribution.

Predictions for pure tones

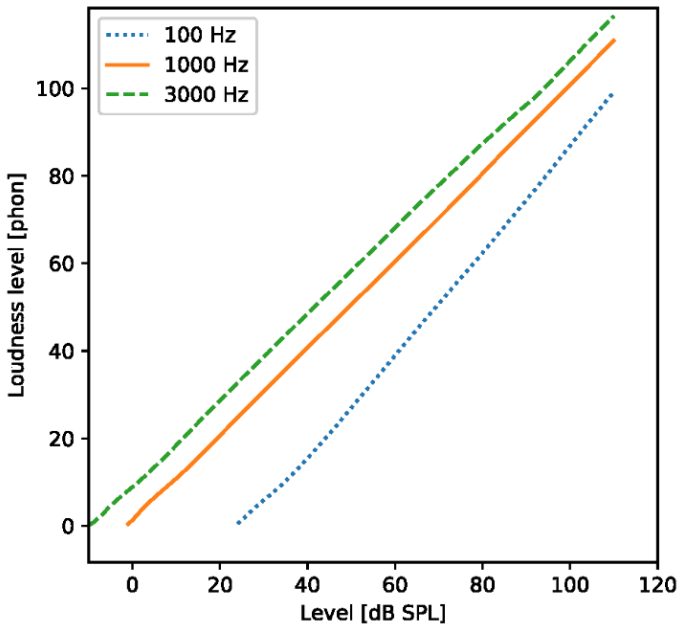


Fig. 2: Loudness level in phons predicted by the DNN as a function of sound level for pure tones with frequencies of 100, 1000, and 3000 Hz.

Figure 2 shows loudness levels predicted by the DNN for pure tones in quiet as a function of input sound level for frequencies of 100 (dotted line), 1000 (solid line) and 3000 (dashed line) Hz, assuming free-field presentation with frontal incidence. The predictions are consistent with empirical data (Hellman, 1976) and are almost identical to the predictions of the TVL model. For the 1000-Hz tone, by definition the loudness level in phons is equal to the physical level in dB SPL. The predictions of the DNN show this relationship almost exactly. The loudness level is greater for the 3000-Hz than for the 1000-Hz tone because 3000 Hz is close to the resonant frequency of the ear canal, so the sound level at the eardrum is boosted relative to that in free field (Shaw and Vaillancourt, 1985). The loudness level is lower at 100 Hz than at 1000 Hz partly because of the attenuation characteristic of the middle ear and partly because less gain is applied by the active mechanism in the cochlea at low frequencies (Cooper, 2004; Moore *et al.*, 1997). Both of these effects are simulated in the TVL model.

Predictions for noises as a function of bandwidth

Figure 3 shows the loudness level of bandpass filtered pink noise centred at 1 kHz, plotted as a function of bandwidth, as predicted by the TVL model and by the DNN. For small bandwidths, the loudness level predicted by the DNN is slightly below that predicted by the TVL model. The predictions of the DNN are actually more consistent with recent empirical data on the loudness of narrowband sounds (Hots *et al.*, 2013), although this is probably just a coincidence.

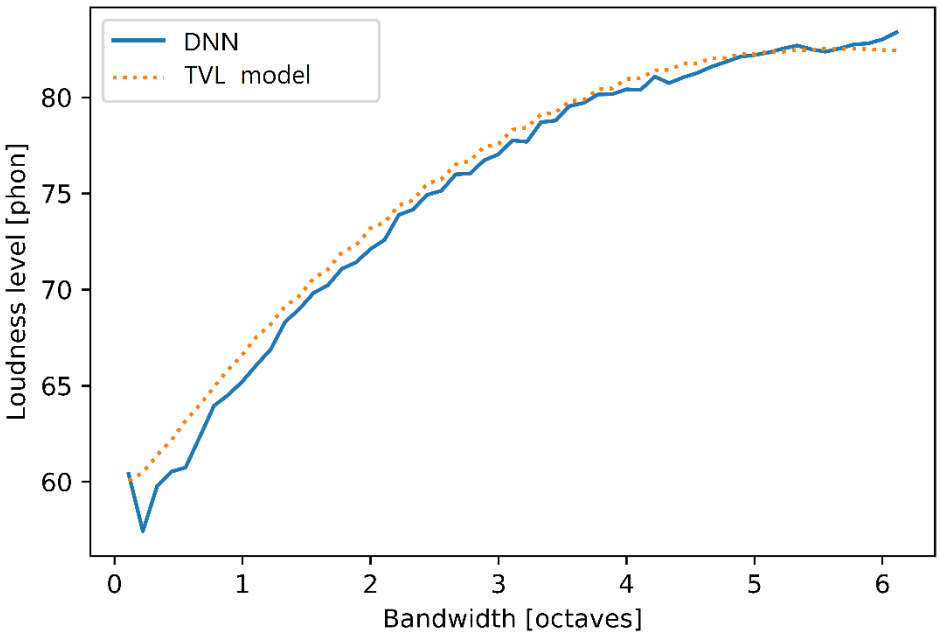


Fig. 3: Loudness level in phons predicted by the DNN (solid line) and by the TVL model (dashed line) as a function of bandwidth.

Note that the sounds whose loudness is predicted in Figures 2 and 3 were included in the sounds used during training, so the accuracy of the predictions is not surprising. Nevertheless, the results show that the inclusion of speech sounds in the training material did not adversely affect the accuracy of the predictions for the artificial tone and noise sounds.

DISCUSSION

The predictions of the DNN for the environmental sounds and music are remarkably accurate. This is noteworthy, since the DNN was trained only using speech and synthetic sounds. This suggests that the DNN generalises well to real-world sounds, and would do so for sounds other than those tested here. The predictions for music were accurate despite the fact that the music test sounds were scaled to have an RMS level of 70 dB SPL, which is higher than the level of 60 dB SPL that was used with the speech sounds used for training. This shows that the DNN works well for sounds with levels that it was not exposed to frequently during training. The good generalisation across level and across types of sounds is probably a consequence of training using tones and noises with a wide range of levels and frequencies as well as with speech sounds. It might be possible to achieve even better generalisation by using an adversarial approach (Szegedy *et al.*, 2013), in which a second DNN tries to find sounds for which the predictions of the first DNN are inaccurate, with the first DNN then adapting in order to achieve more accurate predictions for the problematic sounds. We leave this for a future study.

The gain in computation speed of the DNN relative to the TVL model was a factor greater than 100. This would allow real-time implementation. It is possible with a modern PC (with Intel i7 6th generation central processing unit) to analyse a 24-hour recording at 1-ms intervals in a few minutes.

Potential applications of the DNN include development of a real-time loudness meter and real-time control of levels in broadcasting to ensure (among other things) that the advertisements are not louder than the main programme material (Moore *et al.*, 2003). The DNN could be extended to predict loudness for people with hearing loss (Moore and Glasberg, 1997; 2004). In principle this could be used for on-line control of loudness in hearing aids so as to restore loudness perception more nearly to normal (Launer and Moore, 2003)

CONCLUSIONS

The DNN gave accurate predictions of loudness for environmental sounds and music despite training using speech and synthetic sounds only. This shows good generalisation and suggests that the DNN will give reasonably accurate predictions for a wide variety of everyday sounds. Most predictions were accurate within 0.5 phons, a difference in loudness level that would not be detectable. The DNN calculates instantaneous loudness more than 100 times faster than the TVL model, making real-

time implementation possible. This opens up potential applications in broadcasting and in the on-line control of loudness in hearing aids.

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council (UK, grant number RG78536). We thank Brian Glasberg for his collaboration in the development of the TVL model.

REFERENCES

Aibara, R., Welsh, J. T., Puria, S., and Goode, R. L. (**2001**). "Human middle-ear sound transfer function and cochlear input impedance," Hear. Res. **152**, 100-109. doi: 10.1016/S0378-5955(00)00240-9

Cooper, N. P. (**2004**). "Compression in the peripheral auditory system," in *Compression: From Cochlea to Cochlear Implants*, edited by S. P. Bacon, R. R. Fay, and A. N. Popper (Springer, New York), 18-61.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103-138. doi: doi.org/10.1016/0378-5955(90)90170-T

Glasberg, B. R., and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds," J. Audio Eng. Soc. **50**, 331-342.

Hellman, R. P. (**1976**). "Growth of loudness at 1000 and 3000 Hz," J. Acoust. Soc. Am. **60**, 672-679. doi: 10.1121/1.381138

Hots, J., Rennies, J., and Verhey, J. L. (**2013**). "Loudness of sounds with a subcritical bandwidth: A challenge to current loudness models?," J. Acoust. Soc. Am. **134**, EL334-339. doi: 10.1121/1.4820466

ISO 532-3 (**2019**). *Acoustics - Methods for calculating loudness - Part 3: Moore-Glasberg-Schlittenlacher method for time varying sounds* (International Organization for Standardization, Geneva), (draft).

Kingma, D. P., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980

Launer, S., and Moore, B. C. J. (**2003**). "Use of a loudness model for hearing aid fitting. V. On-line gain control in a digital hearing aid," Int. J. Audiol. **42**, 262-273. doi: 10.3109/14992020309078345

Marshall, A., and Davies, P. (**2007**). "A semantic differential study of low amplitude supersonic aircraft noise and other transient sounds," in *International Congress on Acoustics* (Madrid), pp. 1-6.

Moore, B. C. J. (**2012**). *An Introduction to the Psychology of Hearing, 6th Ed.* (Brill, Leiden, The Netherlands), 1-441.

Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750-753. doi: 10.1121/1.4955005

Moore, B. C. J., and Glasberg, B. R. (**1997**). "A model of loudness perception applied to cochlear hearing loss," Auditory Neurosci. **3**, 289-311.

Moore, B. C. J., and Glasberg, B. R. (**2004**). "A revised model of loudness perception applied to cochlear hearing loss," Hear. Res. **188**, 70-88. doi: 10.1016/S0378-5955(03)00347-2

Moore, B. C. J., and Glasberg, B. R. (**2007**). "Modeling binaural loudness," J. Acoust. Soc. Am. **121**, 1604-1612. doi: 10.1121/1.2431331

Moore, B. C. J., and Oxenham, A. J. (**1998**). "Psychoacoustic consequences of compression in the peripheral auditory system," Psych. Rev. **105**, 108-124. doi: 10.1037/0033-295X.105.1.108

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness and partial loudness," J. Audio Eng. Soc. **45**, 224-240.

Moore, B. C. J., Glasberg, B. R., and Stone, M. A. (**2003**). "Why are commercials so loud? - Perception and modeling of the loudness of amplitude-compressed speech," J. Audio Eng. Soc. **51**, 1123-1132.

Nair, V., and Hinton, G. E. (**2010**). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, edited by J. Fürnkranz, and T. Joachims (Haifa, Israel), pp. 807-814.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (**2015**). "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Brisbane, Australia), pp. 5206-5210.

Piczak, K. J. (**2015**). "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia* (ACM, Brisbane, Australia), pp. 1015-1018.

Rennies, J., Wächtler, M., Hots, J., and Verhey, J. (**2015**). "Spectro-temporal characteristics affecting the loudness of technical sounds: data and model predictions," Acta Acust. united Ac. **101**, 1145-1156. doi: 10.3813/AAA.918907

Shaw, E. A., and Vaillancourt, M. M. (**1985**). "Transformation of sound-pressure level from the free field to the eardrum presented in numerical form," J. Acoust. Soc. Am. **78**, 1120-1123. doi: 10.1121/1.393035

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (**2013**). "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199

Thwaites, A., Glasberg, B. R., Nimmo-Smith, I., Marslen-Wilsen, W. D., and Moore, B. C. J. (**2016**). "Representation of instantaneous and short-term loudness in the human cortex," Front. Neurosci. **10,** article 183, 1-11. doi: 10.3389/fnins.2016.00183

Zorila, T.-C., Stylianou, Y., Flanagan, S., and Moore, B. C. J. (**2016**). "Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing," J. Acoust. Soc. Am. **140**, 402-408. doi: 10.1121/1.4955005

Zwicker, E., and Scharf, B. (**1965**). "A model of loudness summation," Psych. Rev. **72**, 3-26. doi: 10.1037/h0021703

Learning about perception of temporal fine structure by building audio codecs

LARS VILLEMOES[1*], ARIJIT BISWAS[2], HEIKO PURNHAGEN[1], AND HEIDI-MARIA LEHTONEN[1]

[1] *Dolby Sweden AB, Stockholm, Sweden*

[2] *Dolby Germany GmbH, Nürnberg, Germany*

The goal of audio coding is to efficiently describe an auditory experience while enabling a faithful reconstruction to the listener. The subjective quality compared to the original is measured by established psychoacoustic tests (BS.1116, 2015; BS.1534, 2015) and the description cost is measured in number of bits. As it is much cheaper to describe coarse scale signal properties than temporal fine structure (TFS), tools like noise fill, spectral extension, binaural cue coding, and machine learning have increased performance of audio codecs far beyond the first generation based on masking principles (e.g., mp3). In this evolution, implicit knowledge on hearing has been acquired by codec developers, but it has become increasingly difficult to construct tools to predict subjective quality. For example, it is yet unknown which aspects of the TFS that are essential for the listening impression to be preserved. To explore these issues, we study models of auditory representations with the mindset from audio coding. Given a method to solve the inverse problem of creating a signal with a specified representation, evaluating by listening can immediately reveal strengths and weaknesses of a candidate model.

INTRODUCTION

Coarse scale properties of audio signals are cheaper to describe than temporal fine structure (TFS; Moore, 2019). This is exploited in modern audio coding systems. But which aspects of TFS are important to make two signals sound the same to us? In this paper, we walk through current and emerging audio coding methods and suggest an audio coding inspired methodology to improve perceptual modelling. We illustrate this method by an example study regarding tonality which is inspired by research on audio texture synthesis, McDermott *et al.* (2009). For the sake of clarity, we will only discuss mono audio signals.

AUDIO CODING

The goal of audio coding is to convey an auditory experience faithfully while keeping the information rate low (see Fig. 1). A typical source is cinematic content comprising a mix of speech, music, and environmental sounds. Ideally, the decoded content should be perceptually indistinguishable from the original content. This is called transparency. Subjective testing such as BS.1116 (2015) can be used to quantify the

*Corresponding author: lars.villemoes@dolby.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

deviation from this ideal case. For the joint evaluation of several codecs and for larger deviations MUSHRA, BS.1534 (2015), is a better choice. Even in a modern scenario where video coding dominates the bit budget, a lower bitrate for a given perceived quality is preferred.

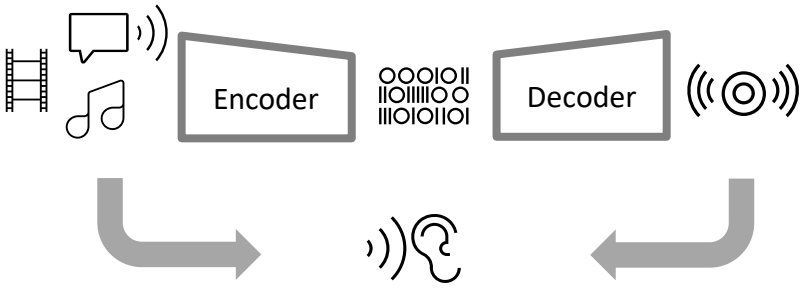


Fig. 1: Audio coding systems encode original sound sources into bits which can be transmitted and decoded into sound again at the receiver end. The combination of an encoder and decoder is called a *codec*.

Currently deployed tools

The concept of a time-frequency (TF) tile is often used to describe a segment of the audio signal in a perceptually motivated frequency band. In practice a filter bank or transform is employed to achieve this. The TF-tiles represent a sufficiently high dimensionality in signals space so that the number of TF-tiles needed to cover the whole signal is much smaller than the number of samples in the signal, see Fig. 2, panel a). Sharing information inside each TF-tile therefore enables bitrate savings. In codecs based on *waveform approximation*, such as mp3, (MPEG-1 layer III), the shared information is a quantization step size which controls the approximation error, and masking principles are employed to make the approximation error inaudible. For high bitrates this method can potentially preserve all aspects of TFS. Significant bitrate savings are obtained by only conveying the energy of the TF-tile, and letting the decoder synthesize a random noise signal in the TF-tile according to this energy target. This method is named *noise fill*. As only a very coarse scale envelope of the signal is preserved, the method rarely offers a high quality. Panel b) of Fig. 2 illustrates the difference between these two methods. *Parametric coding* improves on this situation by adding sinusoids and transients to the repertoire of synthetic signals. Finally, *spectral extension*, illustrated by panel c) of Fig. 2, consists of copying TFS from lower frequencies and adjusting tonal-to-noise ratio with parametric methods. This method is cheap and works surprisingly well. For more details, we refer to the recent tutorials by Brandenburg *et al.* (2013), and Herre and Dick (2019).

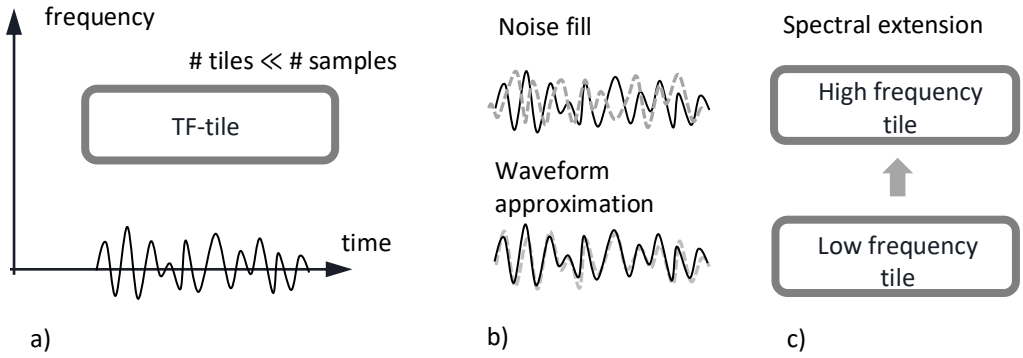


Fig. 2: Currently deployed audio coding tools. Panel a) shows the sharing of information in a TF-tile, panel b) the difference between waveform approximation and noise fill, and panel c) the principle of spectral extension.

Machine learning tools

The application of machine learning to audio synthesis typically consist of training a generative model that maps features or labels into sounds. As depicted in Fig. 3, one can think of these methods as inverse sound classifiers.

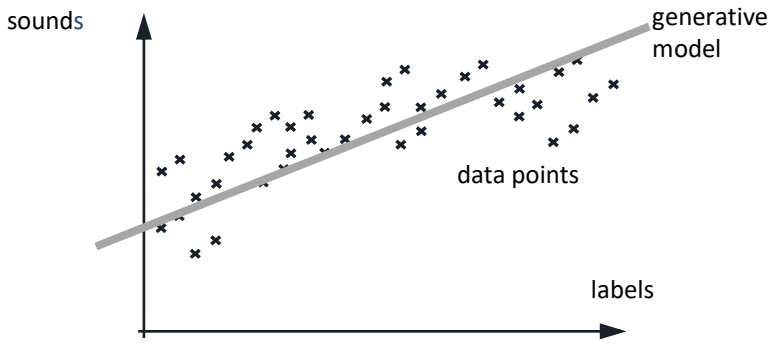


Fig. 3: Conceptual machine learning based sound synthesis.

Recent speech coding examples use vocoder features, such as linear predictive coding (LPC) based spectrum, pitch and degree of voicing (Kleijn *et al.* 2018; Klejsa *et al.* 2019). Autoregressive probability density models are trained on large speech datasets to approximate the distribution of signal samples conditioned on these features. Probabilistic sampling of the resulting model offers a substantial quality improvement over a manually crafted parametric vocoder synthesis. Example results from Klejsa *et al.* (2019) are given in Table 1.

Codec	SILK	SampleRNN	AMR-WB	Vocoder
Bitrate [kb/s]	16	8	23.05	8
MOS-LQO	4.41	3.48	4.39	3.67
MUSHRA	80	78	67	34

Table 1: Bitrates, average of predicted mean opinion scores (MOS-LQO) from the objective tool POLQA, P.863 (2018), and subjective mean MUSHRA scores for four codecs (from Klejsa *et al.*, 2019).

In terms of subjective MUSHRA scores, the machine learning codec based on the autoregressive model SampleRNN performs on par with the waveform codec SILK operating at twice the bitrate, while the parametric vocoder synthesis performs significantly worse. However, the mean opinion scores (MOS-LQO) predicted by POLQA contradict the subjective scores with respect to the comparison between the vocoder and SampleRNN.

PROBLEM

There is a gap in our understanding of auditory perception. The results in Table 1 offer one recent example of the frequently encountered phenomenon that the subjective performance of audio coding is not predicted well by tools such as POLQA and PEAQ (BS.1387, 2001), especially when mechanisms beyond masking are exploited by the codec. Our hypothesis is that TFS aspects are central for explaining this gap. As many model-based predictors of the results of psychoacoustic experiments compare stimuli in an *auditory representation*, (e.g., spectrogram or auditory filter bank, see P.863, 2018; BS.1387, 2001; Dau *et al.* 1996), such auditory representations will be the object of our study.

PROPOSED METHOD

As a complement to the established validation procedures based on targeted psychoacoustic testing, we here propose a method for evaluation and successive improvement of auditory representations based on the idea of building a *mock-up codec*, Fig. 4. For an original sound s having the auditory representation $\theta(s)$, the synthesis process consists of finding a sound u with $\theta(u) \approx \theta(s)$. This "synthesis by analysis" procedure was discussed by Slaney (1995) and is also the basis of spectrogram inversion methods, in which case the representation $\theta(s)$ is a spectrogram (Decorsière *et al.,* 2015).

Synthesis by analysis is an ill-posed inverse problem for which a solution is typically obtained only after many iterations starting from a random noise or manually crafted first guess. Whereas this approach might not be feasible for a deployable codec, off-line synthesis for the purpose of basic research is. Once the synthesis method is constructed, the idea is to run audio signals through the system and evaluate it as a

codec. The machine listener provided by the analysis can then be compared directly to the human listener.

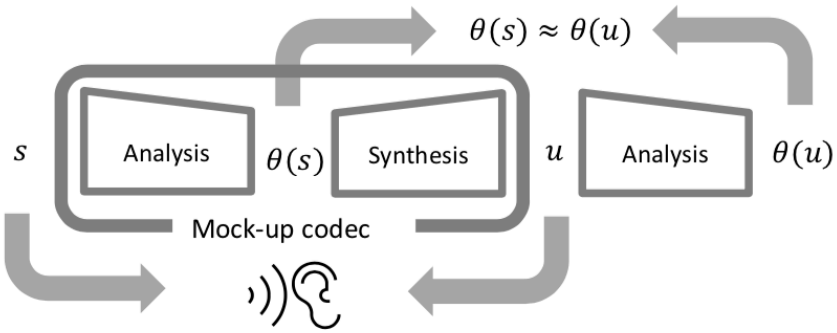


Fig. 4: Synthesis by analysis aims at producing a signal u given the analysis $\theta(s)$ of an original signal s by solving $\theta(u) \approx \theta(s)$.

EXAMPLE STUDY

To illustrate proposed method, we study audio representations derived from a framework which already includes tools for synthesis by analysis and whose signal analysis resembles that of many other models.

McDermott *et al.* (2009) evaluated combinations of time invariant summary statistics for description of stationary audio textures by the method of Fig. 4. For synthesis, an initial white noise signal u_0 was iteratively modified to bring the representation $\theta(u_n)$ closer to $\theta(s)$ than $\theta(u_{n-1})$. Most statistics were derived from envelope values updated every 2.5 ms in a filter bank with 38 bands of perceptually motivated resolution. The quality of textures containing tonal components was not captured well in these experiments. Given our interest in TFS of arbitrary nonstationary signals, and with inspiration from the literature on pitch perception modelling regarding tonality, (Meddis and O'Mard, 1997), we consider two deterministic representations.

A. **Baseline:** measure envelopes every 2.5 ms for all 38 bands as used by McDermott *et al.* (2009).
B. **Extension:** add one lag T and the value $\rho(T)$ of ρ, the normalized autocorrelation function (ACF) for each of the envelope-normalized subband signals every 20 ms.

Fig. 5 depicts the analysis block diagram for both cases. The lag T can be selected in many ways, and the specific steps taken to avoid picking lags related only to the center frequency of the subband are described in Fig. 6. For synthesis, we apply the method of iterative modification of an initial white noise signal. Gradient descent is used for the ACF data.

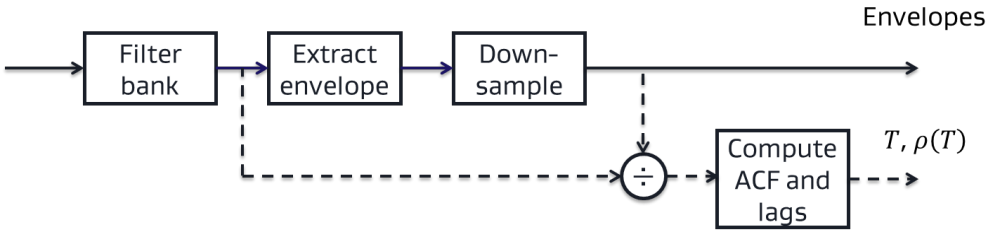


Fig. 5: Analysis algorithm for baseline (solid) and extended (solid and dashed) representations.

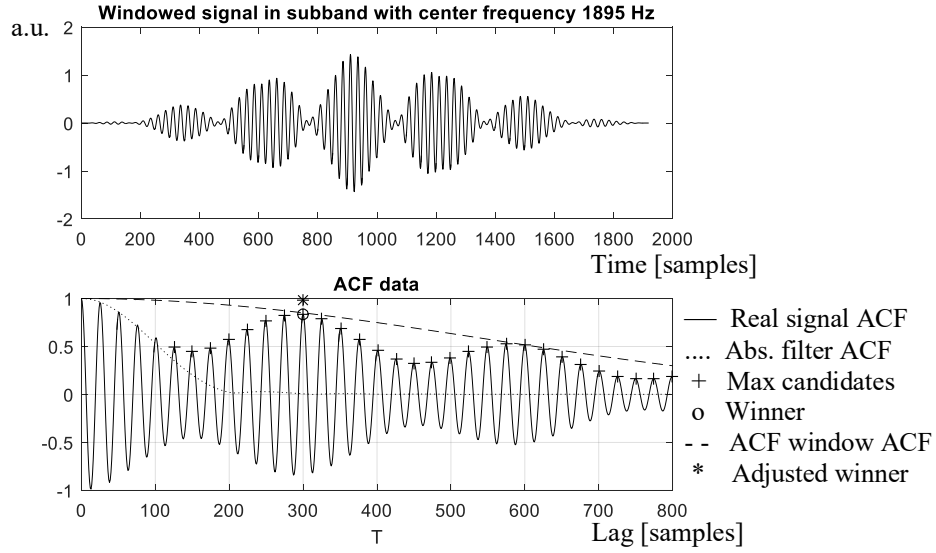


Fig. 6: Example of lag search for sawtooth signal with period 300 samples at 48 kHz sampling. The top panel depicts the windowed subband signal for a band with center frequency 1895 Hz. In the bottom panel, we consider the local maxima (crosses) of the windowed signal's ACF (solid curve), discarding maxima below the ACF of the absolute value of the impulse response of the subband filter (dotted curve). For a better usage of the interval between 0 and 1 where $\rho(T) = 1$ denotes maximum tonality, the maximum value (circle) is divided by the value of the ACF (dashed curve) of the subband ACF window leading to the final selection (star).

RESULTS

Informal listening to inputs and synthesized signals for speech, music, and environmental sounds reveals that the relatively detailed envelope representation (Fig. 7A) alone is not sufficient to capture tonality, while a clear improvement is obtained in voiced parts of speech and tonal parts of music by using the extension (Fig. 7B) including one lag and the corresponding ACF value per band per 20 ms. Spectrograms for an example signal are depicted in Fig. 7.

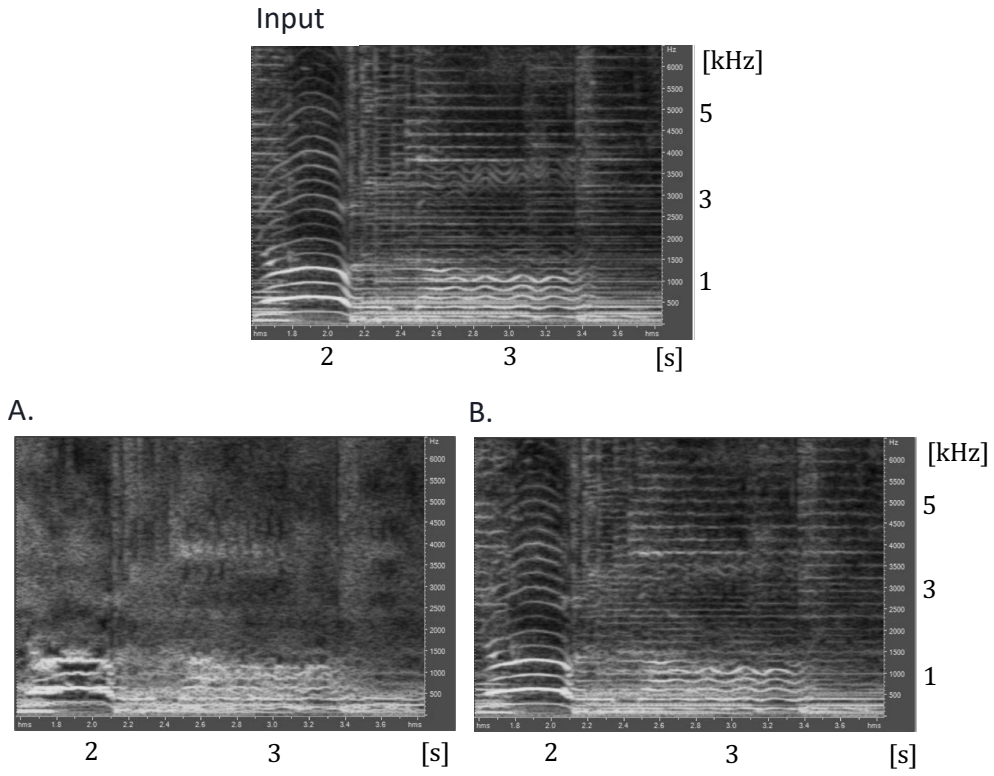


Fig. 7: Spectrograms of input and synthesis from (A) baseline and (B) extended representation for a segment of singing over guitar.

DISCUSSION

The example study shows that informal listening to the outputs of a mock-up codec for general audio provides immediate guidance on the construction of audio representations with the ambition to capture perceptually relevant aspects of audio. A MUSHRA test could be used to verify the shortcomings of the baseline representation (A) relative to the extended representation (B) whose own shortcomings would then also be revealed. For example, we expect both representations to fail for aspects requiring analysis of cross-frequency band coherence of TFS. We believe that the proposed method could be used in the construction and refinement of more well-developed models of TFS aspects of hearing, as well as improved predictors of subjective quality for pairs of perceptually similar signals as those available in the samples link of Klejsa *et al.* (2019).

REFERENCES

Brandenburg, K., Faller, C., Herre, J., Johnston, J. D., and Kleijn, W. B., (**2013**). "Perceptual Coding of High-Quality Digital Audio," Proc. IEEE, **101**, 1905 - 1919, doi: 10.1109/JPROC.2013.2263371

BS.1116 (**2015**). "Methods for the subjective assessment of small impairments in audio systems," Recommendation ITU-R BS.1116-3. Retrieved from: https://www.itu.int/rec/R-REC-BS.1116/en

BS.1387 (**2001**). "Method for objective measurements of perceived audio quality," Recommendation ITU-R BS.1387-1, https://www.itu.int/rec/R-REC-BS.1387

BS.1534 (**2015**). "Method for the subjective assessment of intermediate quality levels of coding systems," Recommendation ITU-R BS.1534-3. Retrieved from: https://www.itu.int/rec/R-REC-BS.1534

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," J. Acous. Soc. of Am., **99**, 3615-3622, doi: 10.1121/1.414959

Decorsière, R., Søndergaard, P. L., MacDonald, E. N., and Dau, T., (**2015**). "Inversion of Auditory Spectrograms, Traditional Spectrograms, and Other Envelope Representations," IEEE-ACM T. Audio Spe., **23**, 46-56, doi: 10.1109/TASLP.2014.2367821

Herre, J., and Dick, S., (**2019**), "Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review," Appl. Sci., **9**, 2854, doi: 10.3390/app9142854

McDermott, J. H., Oxenham, A. J., and Simoncelli, E. P., (**2009**). "Sound texture synthesis via filter statistics," IEEE WASPAA, 297-300 doi:10.1109/aspaa.2009.5346467

Meddis, R., and O'Mard, L. (**1997**). "A unitary model of pitch perception," J. Acoust. Soc. Am., **102**, 1811-20, doi:10.1121/1.420088

Moore, B., (**2019**). "The roles of temporal envelope and fine structure information in auditory perception," Acoust. Sci. Technol., **40**, 61-83, doi: 10.1250/ast.40.61

Kleijn, W. B., Lim, F. S. C., Luebs, A., Skoglund, J., Stimberg, F., Wang Q., and Walters, T. C., (**2018**). "Wavenet Based Low Rate Speech Coding," IEEE ICASSP, 676-680, doi: 10.1109/icassp.2018.8462529

Klejsa, J., Hedelin, P., Zhou, C., Fejgin, R., and Villemoes, L., (**2019**). "High-quality Speech Coding with Sample RNN," IEEE ICASSP, 7155-7159. doi: 10.1109/icassp.2019.8682435 (Samples retrieved from: https://sigport.org/documents/high-quality-speech-coding-sample-rnn)

P.863 (**2018**). "Perceptual objective listening quality prediction," Recommendation ITU-T P.863. Retrieved from: https://www.itu.int/rec/T-REC-P.863

Slaney, M. (**1995**). "Pattern playback from 1950 to 1995," Proc. IEEE Int. Conf. Syst. Man. Cybern., **4**, 3519-3524, doi:10.1109/icsmc.1995.538332

Computational investigation of visually guided learning of spatially aligned auditory maps in the colliculus

TIMO OESS [1,*], MARC O. ERNST[1], HEIKO NEUMANN[2]

[1] *Applied Cognitive Psychology, Ulm University, D-89081 Ulm, Germany*

[2] *Institute of Neural Information Processing, Ulm University, D-89081 Ulm, Germany*

The development of spatially registered auditory maps in the external nucleus of the inferior colliculus in young owls and their maintenance in adult animals is visually guided and evolves dynamically. To investigate the underlying neural mechanisms of this process, we developed a model of stabilized neoHebbian correlative learning which is augmented by an eligibility signal and a temporal trace of activations. This 3-component learning algorithm facilitates stable, yet flexible, formation of spatially registered auditory space maps composed of conductance-based topographically organized neural units. Spatially aligned maps are learned for visual and auditory input stimuli that arrive in temporal and spatial registration. The reliability of visual sensory inputs can be used to regulate the learning rate in the form of an eligibility trace. We show that by shifting visual sensory inputs at the onset of learning the topography of auditory space maps is shifted accordingly. Simulation results explain why a shift of auditory maps in mature animals is possible only if corrections are induced in small steps. We conclude that learning spatially aligned auditory maps is flexibly controlled by reliable visual sensory neurons and can be formalized by a biological plausible unsupervised learning mechanism.

INTRODUCTION

Identifying a location of a visual or auditory event in our environment is advantageous for an organism to orient in space. Localizing a visual stimulus is a rather easy task, since the receptor array (sensorial neurons on a sensory organ) is topographically ordered. Hence, the location of a stimulus can be directly read out by its position on the retina. However, in the auditory domain localizing a stimulus is a complicated task that involves intensive computational steps to overcome two major obstacles. The challenge in sound source localization begins at the stage of the cochlea. The tonotopically organized receptor array represents neighboring frequencies but not adjacent spatial locations. Consequently, the location of a sound source cannot be directly inferred from its array position but needs to be computed using cues created by the head shadow, the distance between the ears or their shape. However, these cues lack associations to absolute locations in space. In order to establish such associations the brain utilizes vision as a guidance signal (Knudsen and Knudsen,

*Corresponding author: timo.oess@uni-ulm.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

1989). Whenever a visual stimulus appears in temporal registration with an auditory stimulus, an association is established between the spatial location of the visual stimulus and computed auditory localization cues to form a topographically ordered representation of auditory sound source locations. Such associations can be found in the external subdivision of the inferior colliculus of barn owls in the form of a topographically aligned map of auditory space (Knudsen and Konishi, 1978).

Here, we present a neural model of auditory map formation in owls utilizing an unsupervised learning rule that evaluates correlations of sensory input streams. Map structures are defined as 1-dimensional arrays of topographically arranged single compartment neurons. Activations of model neurons are defined by rate-based changes of first-order conductance-based membrane dynamics and their transformation into firing rates. The model architecture is based on physiological findings in barn owls (Knudsen and Brainard, 1991; Hyde and Knudsen, 2000; Linkenhoker and Knudsen, 2002) and incorporates parts of external nucleus of the inferior colliculus (ICx) and the optic tectum (OT) to represent auditory and visual inputs, respectively. Unsupervised learning is defined by a neoHebbian 3-factor rule that incorporates an eligibility control signal, a co-activation plasticity mechanism, and a temporal trace of post synaptic activation (Gerstner *et al.*, 2018). Together, these components enable stable map alignment of auditory space instructed by visual guidance signals.

Simulation results demonstrate the ability of the model to resemble behavioral and physiological findings in barn owls and explain the difference in remapping of auditory space in juvenile and adult owls when their visual field is shifted. Specifically, we show that the ability of remapping critically depends on the receptive field size of auditory neurons. This explains why a gradually induced prismatic shift enables remapping of auditory space in mature animals whereas a single large shift only works for juvenile owls.

METHODS

The alignment of auditory space maps in barn owls occurs at the level of the midbrain between the central nucleus of the inferior colliculus (ICC) and ICx. This alignment is guided by retinotopic visual inputs from the OT. The ICC comprises tonotopically ordered neurons that are responsive to certain values of localization cues such as interaural time or level differences. By combining these cues over different frequency channels and associating them with visual information provided by the OT, the ICx forms a topographical map of auditory space.

Inputs to ICx model neurons at location j arise from ICC (audio) and OT (vision) and are denoted s_j^A and s_j^V, respectively. Each input is a one-dimensional vector of $N = 40$ entries that describes the input conductances to the ICx neuron population of size N. This number is chosen to achieve high input resolution while keeping the computational cost in a feasible range. Visual inputs from the OT are topographically structured whereas inputs from the ICC are tonotopical but show a topographical organization of localization cues (Feldman and Knudsen, 1997). Therefore, independent

auditory $s_j^A(x_t)$ and visual $s_j^V(x_t)$ sensory input at location j can be modeled by

$$s_j^{\{A,V\}}(x_t) = \exp(\frac{-(j-x_t)^2}{2 \cdot \sigma_{\{A,V\}}^2}), \quad \text{(Eq. 1)}$$

where x_t is the location at time t and $\sigma_{\{A,V\}}$ the spatial extent of an auditory and visual stimulus on the receptor array, respectively. To ensure similar input strengths the input vectors are min-max normalized. An input location x is randomly chosen between 0 and N and for each such a location a Wiener process (initial setting: $\sigma_{WP} = 1, \mu = x$) is applied that randomly selects five locations in the vicinity of x. Finally, these five locations are consecutively presented, each for 100 time steps (sufficient time to reach a steady state of the membrane potential), before choosing another initial location x. This process is repeated N times.

One model assumption is an increasing energy level in the visual map of space over time due to maturation of the visual system. This incremental increase of energy of visual inputs is modeled by filtering the visual input signal with a temporal high-pass filter according to $\hat{s}_j^V = s_j^V * f$ ($*$: convolution operator) ,where

$$f(t) = \frac{1}{(1+\exp(-(10^{-4} \cdot t - 1)))}. \quad \text{(Eq. 2)}$$

This leads to increased plasticity of the learning for more reliable visual input signals. The auditory input signal is fed to an ICx model neuron r_j, whose membrane potential change is defined by:

$$\tau \dot{r}_j = -\alpha \cdot r_j + (\beta - r_j) \cdot \sum_{i=0}^{N} w_{ji} \cdot s_i^A, \quad \text{(Eq. 3)}$$

where s_i^A describes the auditory input, weighted by connection weight w_{ji} over all input locations. Parameter $\tau = 0.1$ defines the membrane time constant, $\alpha = 1.0$ is a passive membrane leakage rate and $\beta = 1.0$ describes a saturation level of excitatory inputs (standard neuron parameters). To create a firing rate, membrane potential r_j at time t is transformed by an output function $g(r_j(t)) = [r_j(t)]_+ = max(r_j(t), 0)$. Note, that visual inputs do not drive the neuron auditory map neurons in accordance with neurophysiological findings (Knudsen and Knudsen, 1989).

The essential part of the model is a 3-component neoHebbian learning algorithm (Gerstner *et al.*, 2018), that facilitates learning for spatially and temporally aligned inputs. Empirical exploration was used to choose best learning parameter values. Weights are initialized with a large receptive field kernel, $w_{ji} = \exp(\frac{-(i-j)^2}{2 \cdot 20^2})$, so that each receptive field ranges over the entire input space to replicate juvenile owls' receptive fields. The weight adaptation Δw for each learning step is governed by the 3-component structure

$$\Delta w_{ji} = \eta \cdot [post_j(t) \cdot pre_i(t) \cdot fb_j(t) - stabilizer_j(t)], \quad \text{(Eq. 4)}$$

with $\eta = 0.005$ denoting the learning rate, $post_j(t) = \bar{r}_j(t)$ is the temporal trace value of the map activity r_j, $pre_i(t) = s_i^A$ is the activity of the auditory input neuron and $fb_j(t) = F_j^V(t)$ is an eligibility signal. This eligibility signal is driven by the activity of a spatially coincident visual neuron and the overall energy in the visual map, $F_j^V(t) = \hat{s}_j^V(t) \cdot E(t)$ where is $E(t)$ is the normalized total energy in the visual map and is calculated by $E(t) = \sum_j^N \hat{s}_j^V(t) / \sum_j^N s_j^V(t)$. Here, $\hat{s}_j^V(t)$ is the current activation of the visual input after filtering (see Eq. 2) and $s_j^V(t)$ the maximal input strength before filtering. This mechanism guarantees that learning takes place only for reliable visual signals and for spatially aligned auditory and visual events (Brainard and Knudsen, 1993). The temporal trace value $\bar{r}_j$ is calculated by $\bar{r}_j(t+\delta t) = (1-\lambda) \cdot \bar{r}_j(t) + \lambda \cdot r_j(t)$ to achieve more robust map formation. $\lambda = 0.5$ denotes the trace parameter that defines the influence of previous values of r_j on its current state $\bar{r}_j$.

The term $stabilizer_j(t)$ is added to counterbalance the correlative 3-component learning and is defined by $stabilizer_j(t) = \bar{r}_j^2 \cdot w_{ji}$ to pull weights towards constant energy by a rate proportional to the energy of the map neuron activation. By collecting all terms in the weight adaptation mechanism we arrive at a modified Oja learning rule (Oja, 1989) that incorporates an eligibility control signal to regulate the map learning process.

Connections from one neuron to another can degrade over time which is modeled by a decay function that is applied in each time step: $w_{ji}(t+\delta t) = (1-10^{-6}) \cdot w_{ji}(t)$. Here, $w_{ji}(t)$ can become very small but never 0. However, it is possible that connections eventually vanish completely. We model this process by $w_{ji} = 0$, for $w_{ji} < 0.01$. To compensate for this degeneration the model is endowed with a process that allows for reestablishment of already vanished weights. This is achieved by increasing weight values randomly in close vicinity to weights which values exceed a given threshold:

$$\hat{w}_{ji} = exp(\frac{-(i - \text{argmax}_i(j,i))^2}{\sigma_{syn}^2}),$$

$$\Delta w_{ji} = \begin{cases} \hat{w}_{ji} \cdot |\mathcal{N}(0, \hat{w}_{ji})| \cdot \phi_{syn}, & \hat{w}_{ji} \geq t_{syn} \\ 0, & otherwise \end{cases} \qquad \text{(Eq. 5)}$$

where $\sigma_{syn} = 2.8$ defines the range in which new weights can be created, $t_{syn} = 0.5$ is a threshold a weight has to exceed to initiate the process and $\phi_{syn} = 0.1$ defines a scaling factor of the noise. This process is repeated every 1000^{th} time step.

RESULTS

In the following, we present simulation results that, first, show learning abilities of our model for spatially and temporally correctly aligned visual and auditory inputs. This learning of normal responses is used in a second experiment as a reference for comparison with learning for shifted visual inputs. In a third experiment, we show

how the regained ability of shifting an auditory map for incrementally shifted visual inputs in adult owls depends on receptive field size of auditory neurons.

At the beginning of each simulation experiment the auditory system is in its juvenile state (reduced energy in visual map, broad receptive fields of auditory neurons, no auditory map alignment) and develops over the course of $50,000$ time steps to its mature state (maximal energy in visual map, narrow receptive fields of auditory neurons, and supposedly aligned auditory map). To demonstrate correct functionality of the model, we present a learned auditory map for temporally aligned, non-shifted visual and auditory inputs (Fig. 1 **A**). It represents correct map alignment in healthy owls. The abscissa indicates the spatial offset of the alignment from the predicted normal. The predicted normal describes the location offset between auditory and visual signals and is 0° for a perfectly aligned auditory map (spatial coincidence). Data presented here is collected by repeatedly presenting different sound source locations and measuring the response of auditory map neurons.

Temporal coincidence between both inputs determines how well an auditory map can be aligned. For large temporal offsets the map alignment fails due to reduced activity of map neurons at the moment the visual signal would facilitate learning. Temporal coincidence is especially crucial if stimulus locations are randomly sampled from the environment (high σ_{WP} value of Wiener process). However, if consecutive auditory visual events are in spatial vicinity (they are spatially correlated, low σ_{WP} value) correct map alignment is still possible even for large offsets (Fig. 1 **E**).

Experiments with juvenile owls show that if a constant shift of visual inputs is introduced by prismatic goggles, the alignment of the auditory space map is shifted accordingly. This indicates a visually guided learning of auditory space (Brainard and Knudsen, 1993). We replicate this experiment by inducing a shift of the visual inputs by 10° or 20°, respectively, right at the beginning of the learning. Through its role as a guidance signal, the visually shifted input leads to a shifted alignment of the auditory space map (Fig. 1 **B**). However, this shift only occurs when introduced in juvenile owls, but remains ineffective when tested with adult owls (Linkenhoker and Knudsen, 2002). Our model exhibits the same behavior when presented with a non-shifted visual input during development (until time step $100,000$ to simulate sufficient adult experience) followed by a 15° shifted visual input. Map realignment fails since the shifted input is outside the receptive field range of auditory neurons (Fig. 1 **C**). Unlike the ineffective alignment modification in case of a single large shift, incremental small shifts in the visual input can lead to map realignment in adult owls. Simulations with multiple small incremental shifts of the visual signal demonstrate that our model is able to realign the auditory map of space in each consecutive step (Fig. 1 **D**).

Our results indicate that this phenomenon can be ascribed to the receptive field size of auditory map neurons. According to Hebbian learning, new connections and thereby realignment can only be learned if there is a temporally coincident pre- and post-

synaptic activity of neurons (Hebb, 1949). If a large shift is introduced, the activity location of postsynaptic neurons in the auditory map and of visual input neurons do not coincide, since the stimulus location is outside of the spatially aligned auditory neuron's receptive field. In contrast, if small shifts are introduced the stimulus location might still be in the receptive field range of the auditory neuron which leads to an activation and therefore a relearning of connections. We tested this hypothesis by measuring the maximal shift step size depending on the receptive field size of the auditory neurons. For each receptive field size (range $[0.5 \cdot \sigma_A, 7 \cdot \sigma_A] \approx [4px, 40px]$) different shift steps (range $[5°, 25°] = [5px, 25px]$) are tested and the maximal possible shift is measured (a shift is successful if the activity location of the map neurons corresponds to the induced shift). In total, we ran 8 simulations and calculated the

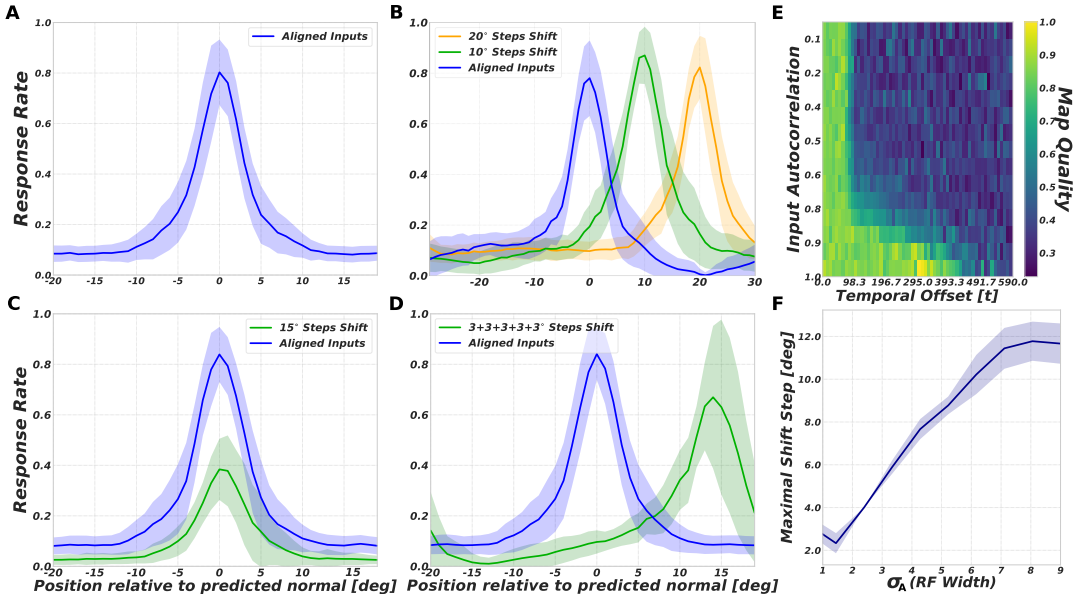


Fig. 1: Left and middle column. Population tuning curves of ICx neurons to different auditory stimulus locations relative to each neuron's predicted normal response are shown (as in Knudsen (1998)). Normal responses of a neuron are derived from the learned map for temporally aligned, non-shifted inputs. Lines depict mean of model neuron responses. Standard deviation in colored area. In all plots the blue line is the normal response plotted for easier comparison. **A** depicts model responses for temporally aligned, non-shifted inputs. **B** shows responses for various shifts of visual inputs. In **C** green line plots model responses for a shifted visual input of 15° in adult animals. **D** shows model responses for incremental shifted inputs in adult animals. Right column. **E** map alignment quality over autocorrelation value of inputs (ordinate) and their temporal offset (abscissa). Perfect map alignment is indicated in bright yellow, no alignment in dark blue. **F** depicts maximal shift over receptive field width when shift is introduced at a mature state.

mean maximal shift for each receptive field size. For increasing receptive field size the maximal shift step becomes proportionally larger (Fig. 1 **F**).

DISCUSSION & CONCLUSION

We presented a neuron model that learns aligned auditory maps of space by applying a 3-component neoHebbian correlative learning rule to its visual and auditory inputs. Previous investigations of auditory map alignment demonstrated that alignment is possible with a *map adaptation cue* in a spiking neuron model (Huo and Murray, 2009) or with a simple, unconstrained Hebbian learning rule for visual and auditory inputs (Witten *et al.*, 2008). Our model differs in that it directly uses visual inputs as a guidance signal for map alignment and is able to explain the unaffected map alignment in adult owls when a large prismatic shift is induced versus the ability to realign for incremental shift step size.

Simulation results demonstrated the model's ability to successfully learn aligned auditory maps of space for temporally and spatially aligned visual and auditory sensory inputs. Due to the introduced eligibility signal, visual inputs do not drive responses of auditory map neurons but merely elicit learning for coincident stimuli, which is important for close replication of biological findings.

The incorporated trace rule can compensate for small temporal offsets between the visual and auditory inputs. However, for large temporal offsets, results indicate that the success of learning strongly depends on the spatial autocorrelation of the input locations. That is, if auditory and visual inputs are not randomly sampled in space but are chosen according to a Wiener process with small σ_{WP}, correct map alignment is still possible. This implies that learning is enhanced for stimulus locations that show strong autocorrelation. When transferring our results to real world scenarios, this enhancement is of special interest since audio-visual events rarely happen to occur at just a single location, followed by other random locations but are likely to happen in spatial vicinity. Therefore, such an enhanced learning capability for strongly autocorrelated input locations seems to facilitate learning of real world stimuli.

It has been shown experimentally that the ability to shift the auditory map alignment is reduced for adult animals (Linkenhoker and Knudsen, 2002). Our model results predict that the receptive field size of the auditory neurons is responsible for this reduction and it can serve as an index of maximal shift step size for visual inputs that can still induce map realignment. This prediction could be tested by varying the prismatic shift step size and determining the receptive field size for adolescents of different ages. If our prediction is correct the two values should correlate.

Despite the presented results, at the moment our model is incapable to maintain already established maps and to quickly readjust to normal vision as it has been demonstrated for owls in (Knudsen, 1998). However, we argue that this could be achieved by adapting the learning algorithm and extend the input by elevation cues. Together with a N-methyl-D-aspartate (NMDA) receptor signal which in addition to

the eligibility signal could control the learning, we expect the model to further extend its capability to resemble neurophysiological and behavioral studies.

REFERENCES

Brainard, M.S. and Knudsen, E.I. (**1993**). "Experience-dependent plasticity in the inferior colliculus: A site for visual calibration of the neural representation of auditory space in the barn owl," J. Neurosci. doi:10.1523/JNEUROSCI.13-11-04589.1993.

Feldman, D.E. and Knudsen, E.I. (**1997**). "An Anatomical Basis for Visual Calibration of the Auditory Space Map in the Barn Owl's Midbrain," J. Neurosci. doi:10.1523/JNEUROSCI.17-17-06820.1997.

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., and Brea, J. (**2018**). "Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules," Frontiers in Neural Circuits. doi:10.3389/fncir.2018.00053.

Hebb, D.O. (**1949**). *The Organization of Behavior: A Neuropsychological Theory* (Psychology Press). ISBN 978-1-135-63191-8.

Huo, J. and Murray, A. (**2009**). "The adaptation of visual and auditory integration in the barn owl superior colliculus with Spike Timing Dependent Plasticity," Neural Netw. doi:10.1016/j.neunet.2008.10.007.

Hyde, P.S. and Knudsen, E.I. (**2000**). "Topographic projection from the optic tectum to the auditory space map in the inferior colliculus of the barn owl," J. Comp. Neurol. doi:10.1002/(SICI)1096-9861(20000529)421:2⟨146::AID-CNE2⟩3.0.CO;2-5.

Knudsen, E.I. (**1998**). "Capacity for Plasticity in the Adult Owl Auditory System Expanded by Juvenile Experience," Science. doi:10.1126/science.279.5356.1531.

Knudsen, E.I. and Brainard, M.S. (**1991**). "Visual instruction of the neural map of auditory space in the developing optic tectum," Science. doi:10.1126/science.2063209.

Knudsen, E.I. and Knudsen, P.F. (**1989**). "Vision calibrates sound localization in developing barn owls," J. Neurosci. doi:10.1523/JNEUROSCI.09-09-03306.1989.

Knudsen, E.I. and Konishi, M. (**1978**). "A neural map of auditory space in the owl," Science. doi:10.1126/science.644324.

Linkenhoker, B.A. and Knudsen, E.I. (**2002**). "Incremental training increases the plasticity of the auditory space map in adult barn owls," Nature. doi:10.1038/nature01002.

Oja, E. (**1989**). "Neural networks, principal components, and subspaces," Int. J. Neural Syst. doi:10.1142/S0129065789000475.

Witten, I.B., Knudsen, E.I., and Sompolinsky, H. (**2008**). "A Hebbian Learning Rule Mediates Asymmetric Plasticity in Aligning Sensory Representations," J. Neurophysiol. doi:10.1152/jn.00013.2008.

“Psychophysical” modulation transfer functions in a deep neural network trained for natural sound recognition

TAKUYA KOUMURA[1,*], HIROKI TERASHIMA[1], AND SHIGETO FURUKAWA[1]

[1] *NTT Communication Science Laboratories 3-1, Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198 Japan*

Representation of amplitude modulation (AM) has been characterized by neurophysiological and psychophysical modulation transfer functions (MTFs). Our recent computational study demonstrated that a deep neural network (DNN) trained for natural sound recognition serves as a good model for explaining the functional significance of neuronal MTFs derived physiologically. The present study addresses the question of whether the DNN can provide insights into AM-related human behaviours such as AM detectability. Specifically, we measured “psychophysical” MTFs in our previously developed DNN model. We presented to the DNN sinusoidally amplitude-modulated white noise with various AM rates, and quantified AM detectability as d' derived from the model’s internal representations of modulated and non-modulated stimuli. The overall d' increased along the layer cascade, with human-level detectability observed in the higher layers. In a given layer, the d' tended to decrease with increasing AM rates and with decreasing AM depth, which is reminiscent of a psychophysical MTF. The results suggest that a DNN trained for natural sound recognition can serve as a model for understanding psychophysical AM detectability. Since our approach is not specific to AM, the present paradigm opens the possibility of exploring a broad range of auditory functions that can be evaluated by psychophysical experiments.

BACKGROUND

Amplitude modulation (AM) is an important physical dimension for natural sound recognition. For example, humans can recognize speech and other natural sounds with a deteriorated temporal fine structure if the amplitude envelopes of the sounds are preserved (Shannon *et al.*, 1995; Gygi *et al.*, 2004).

Numerous neurophysiological studies have sought to reveal how the auditory system represents AM. They have found that the spike synchrony to the stimulus AM and the average spike rate in neurons in the auditory system exhibit tuning to the AM rate. Tuning to the AM rate is often characterized by a modulation transfer function (MTF), which is defined as the spike synchrony or average spike rate as a function of the AM rate. Interestingly, peak AM rates and the upper cutoff frequencies of the MTFs

*Corresponding author: koumura@cycentum.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

systematically transform along the cascade of the brain regions in the auditory system (Joris *et al.*, 2004).

In our previous study, we asked an alternative question: why does the auditory system represent AM in such ways (Koumura *et al.*, 2019)? To explore the functional significance of the systematically transforming MTFs, we built a computational model that can perform a behaviourally meaningful task, namely natural sound recognition. Specifically, we trained a deep neural network (DNN) for the task and analysed the AM representation in it. A DNN is suitable for modelling the auditory system in two ways. First, it can perform natural sound recognition, which is one of the most important functions of the auditory system. Functions such as vocal communication and sound localization are of similar importance, but in this study we only focused on natural sound recognition. Second, it consists of a cascade of layers, which is similar to the cascade of brain regions in the auditory system (see Fig. 30-12 in Kandel *et al*. (2000)).

To directly compare the AM representation in the DNN with that revealed by neurophysiological studies, we performed single-unit recording in the trained DNN. We found that similar transformation of MTFs along the layer cascade in the DNN emerged as a result of the training for natural sound recognition. The similarity gradually increased in the course of the training. The results suggest that AM tuning in the auditory system might also be a result of optimization for natural sound recognition in animals in the course of evolution and development.

While neurological studies have investigated the neural representation of AM, psychophysical studies have sought to characterize behavioural responses to it. They have found the dependency of sensitivity to AM on AM rates, which is characterized by a psychophysical MTF defined as the AM detection threshold as a function of the AM rate. For example, when broadband noise is used as the stimulus carrier, an MTF takes the form of a low-pass filter, with the detection threshold decreasing about 3 dB per octave (Viemeister, 1979).

The present study addresses the question of whether such psychophysical properties also emerge in the DNN trained for natural sound recognition and to what extent they are similar to those observed in humans. We conducted a psychophysical AM detection experiment in our DNN (Fig. 1). To characterize AM sensitivities in the DNN, we calculated sensitivity index d' based on the representation of the stimuli in each layer and each unit, and defined the detection threshold as the minimum AM depth required to yield a certain value of d'. The MTFs in the middle layers were similar to those in humans, whereas an untrained DNN was not as sensitive to AM as humans are. The results suggest that not only neurophysiological MTFs but also psychophysical MTFs can be compared between the auditory system and DNNs to better understand why the MTFs have specific forms.

METHODS

Training of the DNN

As a model, we used the DNN we built in our previous study (Koumura *et al.*, 2019). Here we briefly explain the model and the training procedure. The DNN consists of 13 temporally dilated convolutional layers, and each layer consists of 128 units. All layers operate with 44.1 kHz sampling frequency.

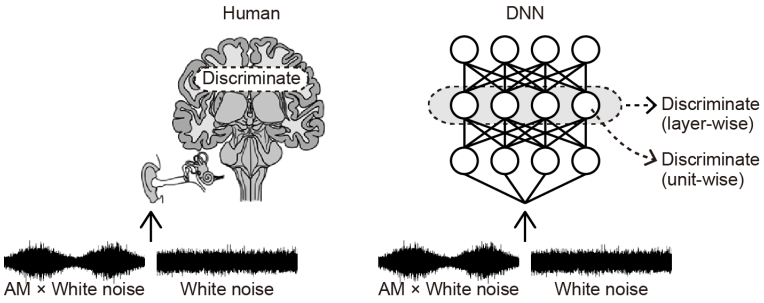


Fig. 1: Our approach. In an AM detection experiment, humans try to discriminate modulated stimuli from unmodulated stimuli (**left**). We simulated this procedure by trying to discriminate modulated and unmodulated stimuli from DNN's representation at the level of single layers or single units (**right**).

The DNN was trained for natural sound recognition. The input was a 0.19 s segment of natural sound, and the DNN's task was to estimate the category of the input sound. The sound data was a subset of ESC-50 (Piczak, 2015). The original dataset is divided into 5 folds. We used folds 1–4 for training and fold 5 for validation. The classification accuracy for the validation set was 45.1%. Due to space limitations, the details of the hyper-parameters and training procedures are not fully described here. They are provided in our previous paper (Koumura *et al.*, 2019).

Stimulus for AM detection experiment

As a stimulus, we used modulated and unmodulated white noise. The duration of the stimulus was 0.5 s, and raised cosine ramps of 50 ms were applied. The starting phase of AM was fixed at 0. The duration of the stimulus and ramps and the starting phase were the same as in Viemeister (1979). The overall amplitude was scaled so that the root mean square (RMS) of the stimulus was equal to the average root mean square (RMS) of the training data. The amplitude was scaled before modulation was applied as in Viemeister (1979). All carrier white noises were independently sampled trial by trial.

RESULTS

Representation of modulated and unmodulated sounds in a single layer

First, we visualized the representation of modulated (AM rate = 32 Hz; depth = −10 dB) and unmodulated white noise in the 7th layer as an example. Response time

courses in the 128 units were recorded in a single layer (Fig. 2). Responses to 32 modulated and 32 unmodulated noises in all units were concatenated, and their dimension was reduced to 4 by principle component analysis (PCA). We confirmed that the results obtained with 2- or 8-dimensional PCA were similar to those with 4-dimensional PCA. Visualizing the first two principle components indicates that in the 7th layer, AM with 32 Hz and −10 dB depth was well discriminated from unmodulated noises (Fig. 3).

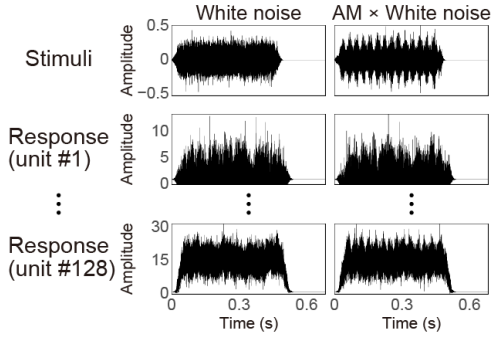


Fig. 2: Examples of stimuli and the DNN's responses. One sample of the noise stimulus is shown for each of the modulated and unmodulated stimuli. Responses in unit #1 and #128 in the 7th layer are shown.

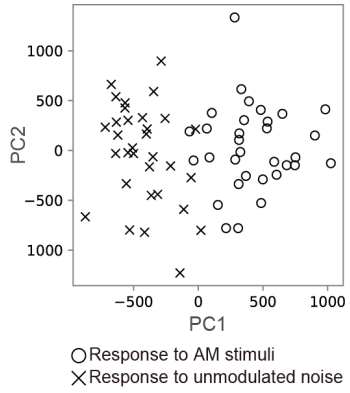


Fig. 3: Two-dimensional visualization of the AM representation in the 7th layer. Responses to 32-Hz modulated and unmodulated noises are shown. Since our model is deterministic, the only cause of the response variability is the variability of input noises. The horizontal and the vertical axes show first- and second-principle components, respectively.

Sensitivity index

As a measure of AM sensitivity, we calculated d' for AM detection based on the representations in each layer. The 4-dimensional representations of the 32 modulated and 32 unmodulated white noises were further projected onto a 1-dimensional axis with maximum detectability in terms of a linear discriminant analysis. The d' was calculated from the means and variances of the 1-dimensional representation (Averbeck and Lee, 2006).

Fig. 4 shows d' in the 7th layer for 32 Hz AM. The d' appeared constant and low with shallow AM, and at a certain AM depth it started to increase linearly on a logarithmic scale. This trend—sensitivity increasing with AM depth—is reasonable when considering the stimulus charactersitics. In theory, the shallower the AM, the more difficult it will be to detect it. Having observed this trend, we fitted a broken line with two segments to the d' on a logarithmic scale. One of the segments for the lower depth was assumed to be constant. The mean squared error of the fitted lines and measured logarithmic d' was 0.024 ± 0.013 (mean ± standard deviation over all AM rates and

all layers). From the fitted lines, we defined the detection threshold as the AM depth at $d' = 1.089$, which corresponds to 70.7% correct, assuming the responses follow a normal distribution. As in the standard psychophysical studies, an MTF was defined as the AM detection threshold as a function of AM rate.

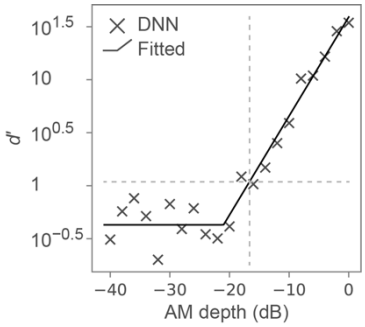


Fig. 4: The d' in the 7th layer for 32-Hz AM rate (crosses) and fitted lines. Horizontal and vertical dashed lines indicate $d' = 1.089$ and the detection threshold, respectively.

The d' was calculated for multiple AM rates and depths in each layer (Fig. 5, upper panels). The calculated values were large for deep and slow AM, and they were also large in the higher layers. The detection threshold in the DNN and humans are compared in the lower panels of Fig. 5. MTFs in the middle layer took the form of a low-pass filter with constant d' up to around 8 Hz and the slope of approximately 3 dB/octave, which is similar to those in humans, although there seems to be a constant discrepancy.

MTFs in a single unit

The above analysis was a comparison between human MTFs and an MTF in a single layer, calculated from the concatenated response timecourses in all units. Next, we calculated an MTF in each unit. MTFs varied among units (Fig. 6, upper panels). Interestingly, in the middle layer, their envelope aligned with the human MTFs. This is more clearly seen by connecting the most sensitive MTFs (Fig. 6, lower panels). In the middle layer, the envelope of the unit MTFs was very similar to that of human MTFs without a constant discrepancy.

AM sensitivity in the untrained DNN

The observed AM sensitivity could be a consequence of the training for natural sound recognition or could be explained by the architecture of the model with cascaded convolutions. To test these possibilities, we calculated the d' in an untrained DNN as a control experiment. The connection weights in the untrained DNN were randomly sampled from the normal distribution, and its activity bias was 0 (He *et al.*, 2015). The d' in the untrained DNN was much lower than those in the trained DNN, indicating that representations in the untrained DNN were not sensitive to the stimulus AM (Fig. 7). The results suggest that parameter optimization is necessary for AM sensitivity. It is worth noting again that our DNN is optimized for natural sound recognition, not for AM sensitivities in humans.

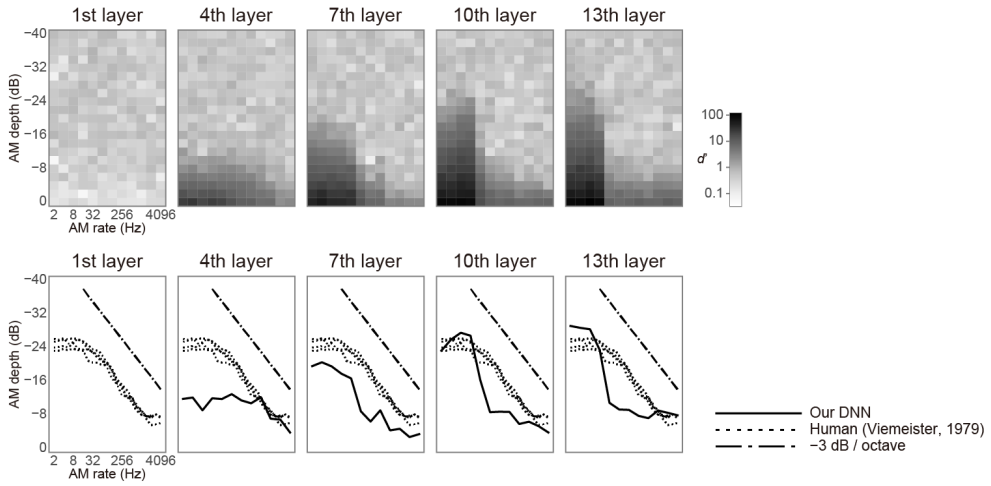


Fig. 5: Sensitivity index and MTF in each layer. The *d′* is colour-coded in the upper panels. The lower panels show MTFs in our DNN (solid lines) and in humans (dotted lines, Viemeister (1979)). We also plotted lines indicating – 3dB/octave (dashed-dotted lines) as in Viemeister (1979).

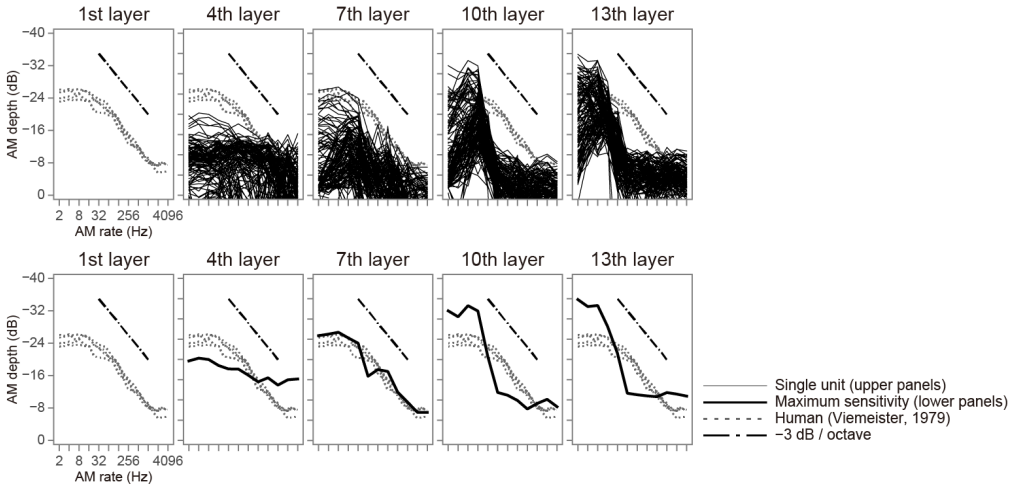


Fig. 6: MTFs in single units (upper) and their envelopes connecting the most sensitive MTFs for each AM rate (lower). Dotted lines and dashed-dotted lines are the same as Fig. 5.

DISCUSSION

We analysed AM sensitivity in the DNN trained for natural sound recognition and found that MTFs similar to those in humans emerged in the middle layers. The untrained DNN did not exhibit high sensitivity. These results, together with the neurophysiological analysis in our previous study, suggest that AM sensitivity in humans might be a result of optimization for natural sound recognition.

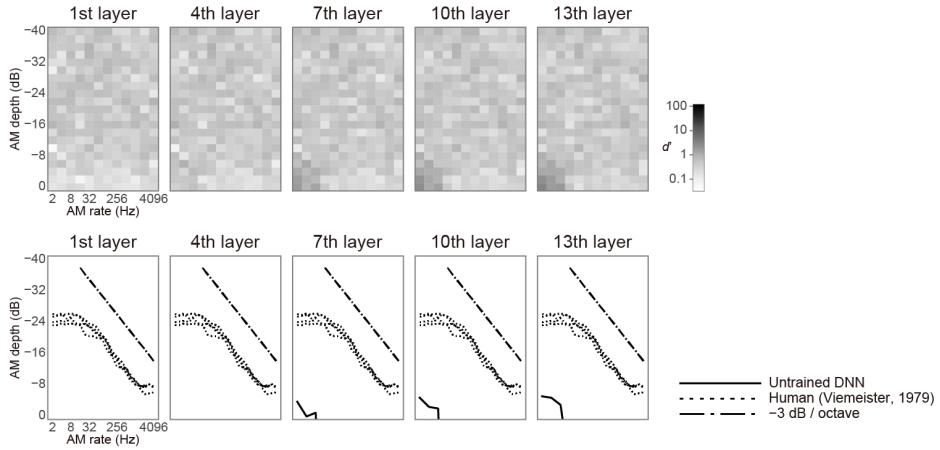


Fig. 7: MTFs in the untrained DNN. Other conventions, including the color scale of *d'*, are the same as in Fig. 5.

In our previous study, DNNs trained with speech signals exhibited neurophysiological MTFs similar to those trained with natural sounds. Therefore, we expect that they would also show psychophysical MTFs similar to those obtained in this study.

Comparison of MTFs in the DNN and in humans

Although the form of the MTF in the middle layer was similar to those in humans, there was a constant discrepancy of around 4 dB, indicating that humans are a little more sensitive to our layer representation. On the other hand, single units can be as sensitive as humans if units with maximum sensitivity are recruited for each AM rate. Thus, it may be possible to model human MTFs by combining MTFs with the most sensitive units. Our previous neurophysiological simulation suggested that unit activities in the middle layers of the DNN may be a model of neural activities in the brainstem. When taken together with the present results, it appears that humans might integrate outputs in the most sensitive neurons in the brainstem to yield responses in an AM detection task.

Future work

The present study only tested an AM stimulus with broad band noise carriers. Psychophysical MTFs have been measured using various carriers, such as those with narrowband noise (Dau *et al.*, 1997), and it has been shown that an MTF depends on the type of carrier. In addition, other types of modulation, such as second-order modulation, has been tested in humans (Lorenzi *et al.*, 2001). Testing other types of stimuli in our model remains as future work.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP15H05915 (Grant-in-Aid for Scientific Research on Innovative Areas "Innovative SHITSUKSAN Science and Technology").

REFERENCES

Averbeck, B.B., Lee, D. (**2006**). "Effects of Noise Correlations on Information Encoding and Decoding," J. Neurophysiol., **95**, 3633–3644. doi:10.1152/jn.00919.2005.

Dau, T., Kollmeier, B., Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation .1. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am., **102**(5), 2892-2905. doi:10.1121/1.420344.

Gygi, B., Kidd, G.R., Watson, C.S. (**2004**). "Spectral-temporal factors in the identification of environmental sounds," J. Acoust. Soc. Am., **115**, 1252–1265. doi:10.1121/1.1635840.

He, K., Zhang, X., Ren, S., Sun, J. (**2015**). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," Proceedings IEEE International Conference on Computer Vision (ICCV), 1026–1034 . doi:10.1109/ICCV.2015.123.

Joris, P.X., Schreiner, C.E., Rees, A. (**2004**). "Neural processing of amplitude-modulated sounds," Physiol. Rev., **84**, 541–577. doi:10.1152/physrev.00029.2003.

Kandel, E.R., Schwartz, J.H., Jessell, T.M. (**2000**). "Principles of Neural Science," Fourth Edition. New York, NY: McGraw-Hill.

Koumura, T., Terashima, H., Furukawa, S. (**2019**). "Cascaded Tuning to Amplitude Modulation for Natural Sound Recognition," J. Neurosci., **39**, 5517–5533. doi:10.1523/JNEUROSCI.2914-18.2019.

Lorenzi, C., Soares, C., Vonner, T. (**2001**). "Second-order temporal modulation transfer functions," J. Acoust. Soc. Am., **110**, 1030–1038. doi:10.1121/1.1383295

Piczak, K.J. (**2015**). "ESC : Dataset for Environmental Sound Classification," In 23rd ACM International Conference on Multimedia, 1015–1018.

Shannon, R.V., Zeng, F-G., Kamath, V., Wygonski, J., Ekelid, M. (**1995**). "Speech Recognition with Primarily Temporal Cues," Science, **270**, 303–304. doi:10.1126/science.270.5234.303.

Viemeister, N.F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am., **66**, 1364–1380. doi:10.1121/1.383531.

Using response times to speech-in-noise to measure the influence of noise reduction on listening effort

ILJA REINTEN[1,*] INGE DE RONDE-BRONS[2], MAJ VAN DEN TILLAART-HAVERKATE[2], ROLPH HOUBEN[2] AND WOUTER DRESCHLER[1]

[1] *Clinical & Experimental Audiology, Amsterdam University Medical Centres location AMC, Meibergdreef 9 1105 AZ Amsterdam, the Netherlands*

[2] *Pento Audiological Centre, Zangvogelweg 150 3815 DP Amersfoort, the Netherlands*

Single microphone noise reduction (NR) can lead to a subjective benefit even when there is no objective improvement in speech intelligibility. A possible explanation lies in a reduction of listening effort. In a previous study, we showed that response times (a proxy for listening effort) to a simple arithmetic task with spoken digits in noise were reduced (i.e., improved) by NR for normal-hearing (NH) listeners. In the current study we complemented the data set with data from twelve hearing-impaired (HI) listeners, the target group for NR. Subjects were asked to add the first and third digit of a digit triplet in noise. Response times to this task were measured, subjective listening effort was rated, and speech intelligibility of the stimuli was tested. Stimuli were presented at three signal-to-noise ratios (SNR; -5, 0, +5 dB) and in quiet. Stimuli were either processed with ideal or non-ideal NR, or unprocessed. In contrast to the previous results with NH listeners, a significant effect of NR on response times was for HI listeners restricted to conditions where speech intelligibility was also affected (-5 dB SNR). We cannot confirm a positive effect on response times to speech-in-noise after applying NR for HI listeners.

INTRODUCTION

It is well known that single microphone noise reduction (NR) in hearing aids can lead to a subjective benefit, in terms of listener preference, even when there is no objective improvement in speech intelligibility (Brons *et al.*, 2014). This suggests that in addition to speech intelligibility there are other factors that determine listener preference for NR, such as a reduction of listening effort. The term listening effort, which is a reflection of the amount of cognitive resources that is required for adequate speech understanding (Hicks and Tharpe, 2002), received increasingly more attention in audiological research over the past few decades. It is closely related to fatigue, and therefore regarded as a fairly subjective measure. In spite of its subjective nature, there has been an ongoing effort to find an objective measure that adequately describes listening effort. Such a measure could be of additional value to describe the non-auditory effects of hearing disabilities and of hearing rehabilitation.

*Corresponding author: i.reinten@amsterdamumc.nl

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Objective measures for listening effort that are described in literature include physiological values such as the pupil dilation response, heart rate variability, and EEG recordings. Another approach is the use of response times in a dual-task paradigm to measure listening effort. A primary listening task is complemented with a secondary task that requires cognitive processing. It is believed that the additional cognitive processing required for the secondary task is slowed down if the primary listening situation is more effortful (Hicks and Tharpe, 2002). The nature of the secondary task can be non-auditory, for instance response times to a visual cue (Sarampalis *et al.*, 2009; Desjardins and Doherty, 2014). Sarampalis *et al.* (2009) tested response times to a visual cue when listening to speech in noise at different signal-to-noise ratios (SNR). They found that at -6 dB SNR, normal-hearing (NH) listeners responded faster when the stimuli were processed with a NR algorithm based on a minimum mean square estimator (MMSE; Epraihm and Malah, 1984). Desjardins and Doherty (2014) tested performance in a secondary visual tracking task in moderate and difficult listening situations. The authors found that in difficult listening situations, hearing-impaired (HI) listeners performed better when NR from a commercially available HA was applied. The secondary task can also be based on the primary auditory-only task where extra processing is required. An auditory-only set-up requires less equipment and is therefore better suited for clinical applications. Additionally, test-results are not influenced by a possible non-auditory sensory impairment of the listener. In an experiment by Houben *et al.* (2013), such an auditory-only dual-task was performed by presenting digit triplets in noise. Participants had to identify the digits as the primary task and add the first and third digit as the secondary task. The authors showed that for NH listeners, response times to this simple arithmetic task reduced with increasing SNR when speech intelligibility was at its maximum. As a follow-up of this work, van den Tillaart-Haverkate *et al.* (2017) measured response times at different SNRs and for different forms of noise reduction processing. They found for a group of 12 NH listeners that noise reduction also caused a reduction (i.e., improvement) in response time.

Although the study of van den Tillaart-Haverkate *et al.* (2017) shows promising results for untangling the possible advantage of applying NR in hearing aids, the study lacks data of HI listeners. HI listeners are the target users of hearing aids, and they can have significantly different opinions regarding sound quality of hearing aid features. HI listeners should therefore be included in experiments that evaluate features such as noise reduction. Therefore, in this study 12 HI listeners participated in a similar experiment. Presently, we are interested whether response times to speech-in-noise are reduced for HI listeners after the application of NR.

METHODS

This study was approved by the Medical Ethics Committee of the Amsterdam UMC (former AMC) in 2013 (MEC2013_082).

Participants

Twelve HI listeners participated in this experiment. HI listeners had a mean age of 60 ± 5.3 years with a mild- to moderate sensorineural sloping hearing loss. The participants were recruited in the Audiological Centre of the Amsterdam UMC, location AMC. The group averaged audiogram of the included participants is shown in Fig. 1. All participants were native Dutch speakers.

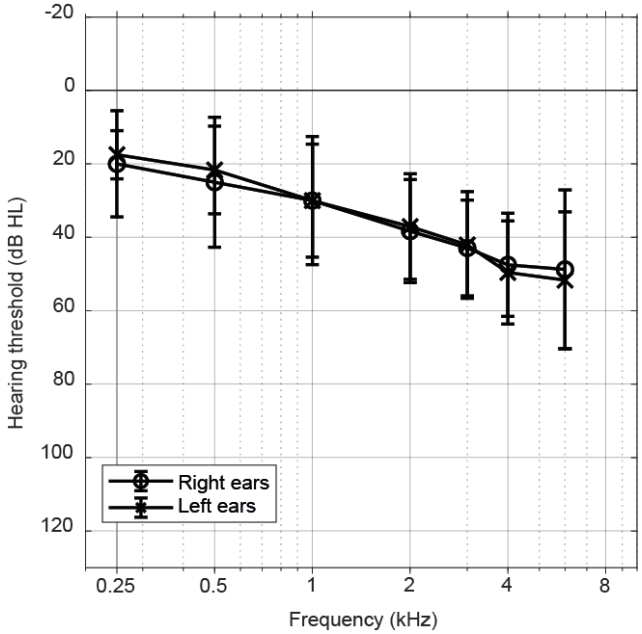


Fig. 1: Group averaged audiograms of the HI subjects with inter-individual standard deviations.

Stimuli and processing

Sixty spoken digit triplets in speech shaped noise were used at four different SNRs: -5, 0 +5 and +∞ (quiet) dB. At each SNR we processed the digit triplets to create three conditions per SNR: one unprocessed condition and two processed conditions with two types of NR algorithms. The two NR algorithms applied in the experiment were the ideal binary mask (IBM; Wang, 2005) and a MMSE (Ephraim and Malah, 1984). The most prominent difference between these two algorithms is that the IBM has a-priori knowledge of the noise and speech material as separate signals and is known to be able to improve speech intelligibility (Wang *et al.*, 2009). Since the IBM requires a-priori knowledge of the noise and speech signals (and the actual SNR), it is not suitable for implementation in real HAs, but it does give insight in the maximum achievable benefit offered by NR. The MMSE on the other hand has to estimate the SNR and is therefore comparable to NR algorithms that are currently implemented in HAs. Implementation of the algorithms was done in MATLAB and is described in

detail in van den Tillaart-Haverkate *et al.* (2017), where a detailed description of the stimuli and equipment used can be found as well. All stimuli were presented diotically through headphones. The average level of the speech was 65 dB(A) with an additional linear amplification for each listener according to the NAL-RP rule (Byrne *et al.*, 2001).

Test procedure and data analysis

The primary outcome measure of this experiment was the response time to an arithmetic task (AR-task). For this task, all participants were presented with digit-triplets in noise and were asked to add the first and third digit. Instructions were given to answer as fast as possible on a numerical keypad. Absolute response times were defined as the time between the end of playing the last digit and the subsequent response key-press. Secondary outcome measures were speech intelligibility (SI) and perceived listening effort rating (LEr). These were tested per condition in the following way: first the participant was asked to correctly identify 20 triplets after which they were asked to rate their perceived listening effort. Listening effort rating was scored on a 9-point scale ranging from 'no effort' (1) to 'extremely high effort' (9) as an answer to the question: "How much effort did it take to understand the last 20 triplets?"

The experiment took place in two visits. The first visit started by measuring hearing thresholds with pure tone audiometry. The AR-task and SI/LEr task were performed in both visits in order to obtain more data points and to allow to investigate the accuracy of the measurement results.

For the AR-task, only correct responses were included in the analysis. For each task, the highest 1.25% of the response times was removed to ensure that unrealistically long response times were not included (Houben *et al.*, 2013). Since absolute response times can have a large inter-individual variation, data analysis was done by using a relative response time. Relative response times were defined by subtracting the response time at $+\infty$ dB SNR from the response time at the other SNRs per processing condition, for each participant.

RESULTS

Fig. 2-A shows the group average absolute response times of the AR-task for all conditions for HI listeners as well as the previously published data (van den Tillaart-Haverkate *et al.*, 2017) of NH listeners. Fig. 2-B shows the mean relative response times of all conditions for HI listeners.

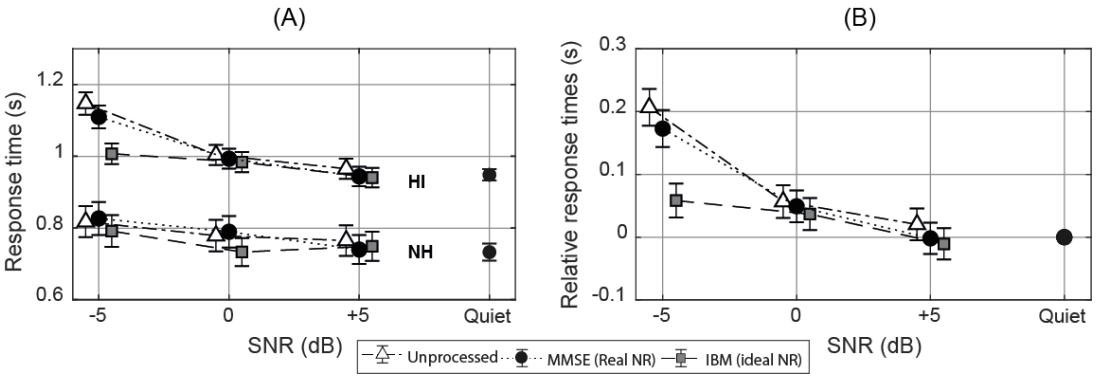


Fig. 2: (**A**) Absolute group-average response times of NH (van den Tillaart-Haverkate *et al.*, 2017) and HI listeners. (**B**) Relative group-average response times of HI listeners. The error bars show 95% confidence intervals.

We analysed the relative response times of the AR-task with a mixed-model ANOVA, with subject and triplet as random effects. Processing condition, SNR and the interaction between processing condition and SNR were considered fixed effects. We found significant effects of processing condition (F = 5.27, p = 0.0184), SNR (F = 107.18, $p < 0.001$), and the interaction between processing condition and SNR (F= 9, $p < 0.001$). Post-hoc pairwise comparisons after Bonferroni corrections (with α = 0.05/27) of the conditions revealed that at -5 dB SNR the IBM condition differed significantly from unprocessed and MMSE. Within the unprocessed and MMSE conditions response times at -5 dB SNR were significantly longer than at all other SNRs and in quiet ($p < 0.001$), and response times at 0 dB SNR were significantly longer than in quiet ($p < 0.001$). Within the IBM condition response times at -5 dB SNR were significantly longer than at +5 dB SNR and in quiet ($p < 0.001$).

Fig. 3-A shows the group average results of the speech intelligibility test in terms of % correct identification of triplets. We analysed the speech intelligibility with a mixed model ANOVA on the rationalized arcsine unit-transformed intelligibility scores, with subject and triplet as random effects. Processing condition, SNR and the interaction between processing condition and SNR were considered fixed effects. We found significant effects of processing condition (F = 12.33, $p < 0.001$), SNR (F = 35.82, $p < 0.001$) and the interaction between processing condition and SNR (F = 5.08, p = 0.001). Post-hoc pairwise comparisons after Bonferroni corrections (with α = 0.05/27) of the conditions revealed that at -5 dB SNR the IBM condition was significantly better intelligible than unprocessed (p = 0.001). Within the unprocessed and MMSE conditions, response times at -5 dB SNR were significantly longer than all other SNRs ($p < 0.001$). Within the IBM condition response times did not differ significantly at all SNRs.

Fig. 3-B shows the group average results of the perceived listening effort rating. We analysed the perceived listening effort rating with an ANOVA with subject as random effect. Processing condition, SNR and the interaction between processing condition

and SNR were considered fixed effects. We found a significant effect of processing condition ($F = 19.3$, $p < 0.001$), SNR ($F = 128.35$, $p < 0.001$), and the interaction between processing condition and SNR ($F = 7.59$, $p < 0.001$). Post-hoc pairwise comparisons after Bonferroni corrections (with $\alpha = 0.05/27$) of the conditions revealed that at -5 dB SNR, the IBM condition differed significantly from the unprocessed and MMSE conditions. Within the unprocessed and MMSE conditions, LEr significantly increased with decreasing SNR ($p < 0.001$), except between +5 dB SNR and 0 dB SNR. Within the IBM condition, LEr at -5 dB SNR was significantly higher than all other SNRs ($p < 0.001$), and LEr at 0 dB SNR was significantly higher than in quiet ($p < 0.001$).

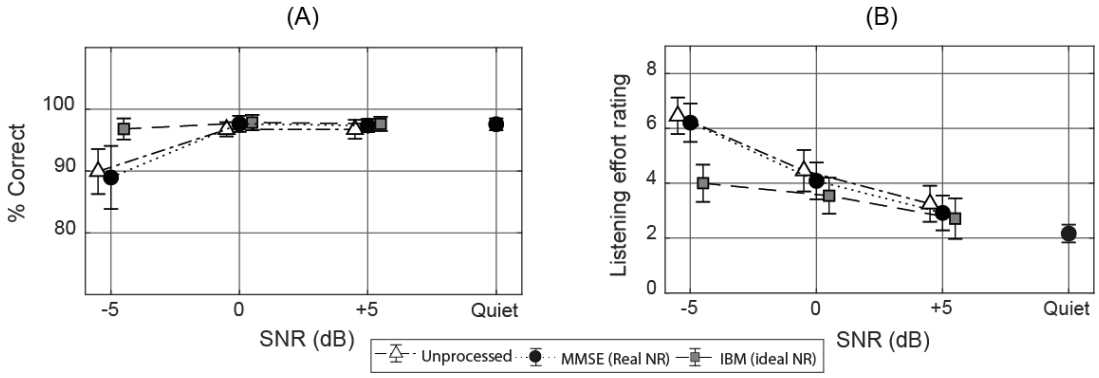


Fig. 3: (**A**) Group averaged speech intelligibility in terms of % correct responses for all SNRs and processing conditions, the error bars show 95% confidence intervals. (**B**) Group averaged perceived listening effort rating for all SNRs and processing conditions, the error bars show 95% confidence intervals.

DISCUSSION

We measured response time of HI listeners on a digit triplet test at different SNRs and for various NR conditions. The results show similarities as well as differences compared with the response time of NH listeners (obtained in a previous study; van den Tillaart-Haverkate *et al.*, 2017). The most obvious difference is that the group of HI listeners needs more time to respond than the previous group of NH listeners. This effect is most likely dominated by the age differences between the two groups. The mean age of the NH listener group (24 ± 4.2 years; van den Tillaart-Haverkate *et al.*, 2017) was on average 36 years younger than the HI listener group of the present study. Response times to tasks are known to increase considerably with age (Verhaeghen and Cerella, 2002; Melzer and Oddsson, 2004). However, we assume that this age effect is negligible in the *relative* response time. Verhaeghen and Cerella (2002) report that reaction times in older adults can commonly be described as a linear transformation to those of younger adults. In the current results we found a significant

reduction of relative response times with increasing SNRs. Within the unprocessed and MMSE condition this reduction was significant even when speech intelligibility was maximal. When comparing Fig. 2-B with Fig. 3-B the same trend is found for response times and perceived listening effort: an increase in SNR is accompanied by a decrease in relative response times or perceived listening effort. These observations, confirmed by statistical analyses, are in agreement with our previous studies and support the hypothesis that response times as such might be used as an objective measure for listening effort (Houben *et al.*, 2013; van den Tillaart-Haverkate *et al.*, 2017).

The main purpose of the current experiment was to test whether response times in a dual-task paradigm are reduced by applying NR for HI listeners, the target group for using NR. Fig. 2-B shows that at all SNRs, the relative response times for the processed conditions are consequently shorter than unprocessed signals. This suggests a positive effect of NR on response times. However, given the significant interaction found between processing condition and SNR, we cannot directly interpret the overall effect of NR on relative response times but instead we have to analyze each SNR separately. This analysis has only revealed a significant reduction of response time between IBM and the other two conditions at -5 dB SNR. This reduction is most likely caused by the large improvement of speech intelligibility for IBM at -5 dB SNR. The decrease of response times with an increase of speech intelligibility is an effect that has been observed before (Gatehouse and Gordon, 1990; Baer *et al.*, 1993), but we are most interested in an effect of NR on response times at SNRs where speech intelligibility is maximal. In this so-called area of interest the reductions in response times by applying NR were not significant, which is in contrast with our previous results in NH listeners (van den Tillaart-Haverkate *et al.*, 2017). A possible explanation for these contrasting results might lie in the prominent reduction in relative response times for IBM at -5 dB SNR, causing the statistical analysis to lose power. The effect of hearing loss also needs to be considered in the test performance. It is well-known that sound perception is different for HI listeners, but little is known how this can affect cognitive processes as measured in dual-task paradigms. Our results are consistent with findings from Sarampalis *et al.* (2009) who also report a reduction in listening effort by applying a similar NR algorithm at a difficult listening situation (-6 dB SNR). Their study did not include HI listeners. Desjardins and Doherty (2014), who did include HI listeners, also measured a reduction of listening effort by NR at a more complex listening condition. Both studies used a visual dual-task paradigm, whereas we used an audiological-only dual task. Our auditory-only dual-task gave similar results to the mentioned visual dual-task paradigms. This finding suggests that an auditory secondary task is suitable for evaluating listening effort. However, the issue remains that the beneficial effects of NR in scenarios where speech intelligibility is maximal may be hard to interpret.

In conclusion, the current dataset of response times to a dual-task paradigm for HI listeners shows a significant and positive effect of increasing SNRs on response times. These results concur with the subjective results of perceived listening effort rating. Nevertheless, in spite of the observed overall effect of NR on response times we

cannot statistically confirm a positive effect on response times to speech-in-noise after applying realistic NR for HI listeners.

REFERENCES

Baer, T., Moore, B. C., and Gatehouse, S. (**1993**). "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," J. Rehabil. Res. Dev., **30**, 49-49.

Brons, I., Houben, R., and Dreschler, W. A. (**2014**). "Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners," Trends Hear., **18**, 1-10.

Byrne, D., Dillon, H., Ching, T., Katsch, R., and Keidser, G. (**2001**). "NAL-NL1 procedure for fitting nonlinear hearing aids: characteristics and comparisons with other procedures," J. Am. Acad. Audiol., **12**(1).

Desjardins, J. L., and Doherty, K. A. (**2014**). "The effect of hearing aid noise reduction on listening effort in hearing-impaired adults," Ear Hearing, **35**(6), 600-610.

Gatehouse, S., and Gordon, J. (**1990**). "Response times to speech stimuli as measures of benefit from amplification," Br. J. Audiol., **24**(1), 63-68.

Hicks, C. B., and Tharpe, A. M. (**2002**). "Listening effort and fatigue in school-age children with and without hearing loss," J. Speech Lang. Hear R., **45**(3), 573-584.

Houben, R., van Doorn-Bierman, M., and Dreschler, W. A. (**2013**). "Using response time to speech as a measure for listening effort," Int. J. Audiol., **52**(11), 753-761.

Melzer, I., and Oddsson, L. I. (**2004**). "The effect of a cognitive task on voluntary step execution in healthy elderly and young individuals," J. Am. Geriatr. Soc., **52**(8), 1255-1262.

Sarampalis, A., Kalluri, S., Edwards, B., and Hafter, E. (**2009**). "Objective measures of listening effort: Effects of background noise and noise reduction," J. Speech Lang. Hear R., **52**(5), 1230-1240.

van den Tillaart-Haverkate, M., de Ronde-Brons, I., Dreschler, W. A., and Houben, R. (**2017**). "The influence of noise reduction on speech intelligibility, response times to speech, and perceived listening effort in normal-hearing listeners," Trends Hear., **21**, 1-13.

Verhaeghen, P., and Cerella, J. (**2002**). "Aging, executive control, and attention: A review of meta-analyses," Neurosci. Biobehav. Rev., **26**(7), 849-857.

Wang, D. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Springer, Boston MA, pp. 181-197.

Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am., **125**(4), 2336-2347.

Hearing examinations in Southern Denmark (HESD): Database description and preprocessing

MANUELLA LECH CANTUARIA[1,2,*], ELLEN RABEN PEDERSEN[3], METTE SØRENSEN[2], FRANS BOCH WALDORFF[4], JESPER HVASS SCHMIDT[1,5,6]

[1] *Institute of Clinical Research, Faculty of Health, University of Southern Denmark, DK-5000 Odense, Denmark*

[2] *Department of Environment and Cancer, Danish Cancer Society Research Center, DK-2100 Copenhagen, Denmark*

[3] *The Mærsk Mc-Kinney Møller Institute, Faculty of Engineering, University of Southern Denmark, DK-5230 Odense, Denmark*

[4] *Research Unit of General Practice, Department of Public Health, University of Southern Denmark, DK-5000 Odense, Denmark*

[5] *Department of Audiology, Odense University Hospital, DK-5000 Odense, Denmark*

[6] *OPEN, Odense Patient data Explorative Network, Odense University Hospital, DK-5000 Odense, Denmark*

All hearing examinations from the public health system of the Region of Southern Denmark have been electronically recorded from 1996 to 2018 and merged into a single database, named the Hearing Examinations of Southern Denmark (HESD) database. This database contains hearing information for more than 143,000 adults, totaling 271,575 valid pure-tone audiograms. The use of this dataset, however, needs to be preceded by an intensive preprocessing procedure in order for the data to be used for research purposes. This study is aimed at describing the HESD database, as well as the preprocessing steps and rules used to classify different types of hearing loss. An initial overview of the different types of hearing profiling and their distribution among our sample is also provided.

INTRODUCTION

The World Health Organization has pointed out hearing impairment as one of the most frequent sensory disabilities worldwide and a leading cause of disease burden. The global prevalence of this disorder for males and females older than 15 years was estimated to be 9.8 % and 12.2 %, respectively (Stevens *et al.*, 2011). Given the aging of the population in many countries, it is likely that the prevalence of hearing loss continues to increase (Cunningham and Tucci, 2017).

Besides aging, there are several other risk factors that can potentially result in hearing loss (HL), such as noise exposure, genetic mutations, cardiovascular diseases (CVD) and ototoxic drugs (Agrawal *et al*., 2009; Cunningham and Tucci, 2017). On the other hand, difficulties in hearing may have critical impacts on the

*Corresponding author: mlca@health.sdu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

individual's ability to navigate in life (e.g., communication), which can increase the risk for other health outcomes. As an example, different studies have consistently found associations between hearing loss and dementia, suggesting hearing loss as an important modifiable risk factor for this disease (Thomson *et al*., 2017).

Even though different hypotheses linking hearing loss and health outcomes, as well as the biological mechanisms behind it, already exist, there is still much to be explored in this regard. Large sample-sized epidemiological data on hearing performance are therefore essential in this context. Within this scope, the HESD (Hearing Examinations in Southern Denmark) database has arisen. The HESD database establishment was based on the data electronically recorded in AuditBase, which is a data capture system used in the public medical system of the Region of Southern Denmark since 1996. However, as this dataset was not originally implemented for research purposes, its use needs to be preceded by an intensive cleaning and preprocessing procedure.

The purpose of this study is to describe the establishment and preprocessing of the HESD database, as well as the information thereby available. We further describe the rules used to classify different types of HL, in order to assess associations between HL characteristics and different diseases. An overview of the different types of hearing profiling and their distribution among our sample is also provided.

METHODS

Database establishment

The HESD database is based on the data electronically recorded in AuditBase from February 1996, March 1998 and June 2003 to 2018 in the public clinics of Vejle, Odense and Sønderborg, respectively. The large majority of the examinations are from patients that had a previous complaint about hearing or any suspected HL (e.g., older patients). The clinics had the software gradually implemented the duration needed for testing and adaptation varied. AuditBase, which was developed by the company Auditdata, is used for collecting and managing auditory clinical data, and therefore consists of a large source of documented data over the years. The clinical data collection is based on the process as defined in ISO 14155 (ISO, 2011) for medical device trials, whereas the audiometric measurements conducted in all of the clinics are based on ISO 8253-1 (ISO, 2010), which addresses the procedure for pure-tone air conduction (AC) and bone conduction (BC) threshold audiometry. The use of the clinical records for research purposes has been authorized by the Danish Patient Safety Authority.

AuditBase contains recorded information of the most important auditory tests, such as: (i) pure-tone audiometry (AC and BC thresholds); (ii) acoustic reflexes (ipsilateral and contralateral stapedius); (iii) speech audiometry (speech reception thresholds and discrimination scores determined using monosyllabic words (Elberling *et al.*, 1989); (iv) tympanometry; and (v) Weber test. Additionally, the system stores person-related information on the patients (e.g. name, date of birth, sex), as well as data on all the people who have performed the testing. The patients

are identified by a unique ID, and each ID is associated to the patient's personal identification number, which can be used to link the HESD data with data from all health registries in Denmark.

Cleaning and preprocessing steps

The raw data extracted from the AuditBase system demands an intensive preprocessing procedure, so that it can be used for research purposes. This is mainly because of the huge size of the dataset, high vulnerability to missing data and lack of consistency between audiograms (e.g. audiograms may vary in terms of the number of frequencies and ears tested, as well as the type of measurements obtained for each ear). Figure 1 shows a simplified flowchart describing the most relevant preprocessing steps used to prepare and transform the data to a suitable form.

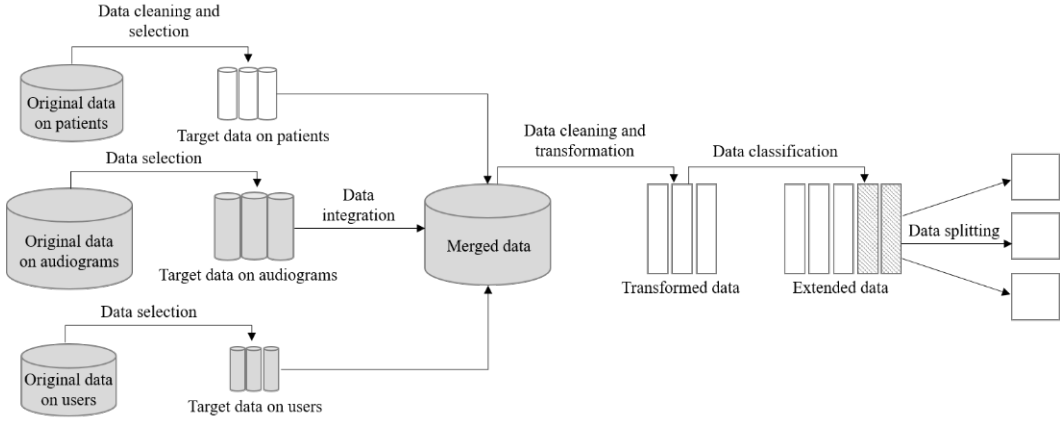


Fig. 1: Steps involved in the HESD database preprocessing.

The preprocessing procedure starts with the original datasets for patients and users' information, as well as the original dataset containing hearing thresholds (measured by, for example, AC, BC and soundfield) and acoustic reflexes thresholds. The latter dataset is organized in a stacked (i.e. narrow) format, meaning that the thresholds obtained for each of the curves present in one specific audiogram were disposed in separate rows. For all the original datasets, the relevant variables were selected and renamed. Additionally, the patients' dataset was cleaned, so that all nonexistent patients (i.e., patients with invalid identification numbers) were removed from the data. In the data integration step, all datasets were merged by the patient's ID.

The merged dataset was further cleaned. In this stage, which was the most demanding one, we have looked for, for example, blank curves and curves with large amount of missing, audiograms obtained for the same patient and at the same examination date, identical audiograms obtained at different dates and audiograms that seem to have been obtained for testing purposes. After this step, we made sure that, for each patient and examination date, we had data for only one audiogram. The transformation stage consisted in: (i) selecting solely the data for the most relevant and frequent audiogram-related measurements (i.e., AC, BC and acoustic reflexes);

and (ii) unstacking the data, so that there is only one row of data for each patient and examination date. It is worth mentioning that a fraction of patients has had their hearing measured more than once along the years in which the data were collected. The data from all these audiograms (i.e. visits) were kept in the dataset.

With the dataset transformed accordingly, we have created categorical variables that indicate:

1. Threshold status, which captures whether the thresholds were masked, out-of-range (i.e., exceeded the maximum output level), crossed-over (i.e., when the sound presented to the test ear was heard by the non-test ear) or uncertain.
2. Data missingness, which captures whether data are available for only one ear and whether BC thresholds were measured. A numerical variable indicating the amount of missing data (for both the total frequency range and the most relevant frequencies) was also created.
3. Data reliability, which captures whether BC thresholds are substantially higher (>10 dB) than the AC threshold measured at the same frequency and masking was done correctly.

Database organization

As the HESD database assembles audiogram data for more than 20 years, differences in the amount of data collected may exist. The lack of data for some specific measurements can also be due to variations on the patients' hearing and symptoms reported, as well as time limitations during the examination. Among all measurements available in the AuditBase system, AC thresholds are certainly the most frequent and the most applicable for research purposes. Given that, the HESD datasets organization is centralized on the AC data. Therefore, all audiograms with AC thresholds available at the most relevant octave frequencies (0.5-4 kHz) for at least one of the ears and obtained for patients older than 18 years were included in the final HESD database.

Hearing loss assessment

To extract the most important information from the pure-tone audiograms, the AC thresholds were used to describe the HL indicated by each audiogram present in the HESD database. We have created a set of rules (Table 1) to classify HL in terms of:

1. Severity, as defined by the pure-tone average (PTA) of 0.5, 1, 2 and 4 kHz.
2. Asymmetry based on interaural AC thresholds differences at octave frequencies between 0.25 to 8 kHz (Margolis and Saly, 2007).
3. Audiogram configuration based on the methods proposed by Demeester *et al*. (2009) and Hannula *et al*. (2011) involving the means of the thresholds at consecutive octave frequencies (i.e., 0.25/0.5, 1/2 and 4/8 kHz) and measurements of the poorer and better thresholds for low, mid and high frequencies.

When BC thresholds were also available, HL was also categorized in terms of:

4. Type of lesion, as defined by the number of air-bone gaps at octave frequencies between 0.25 and 2 kHz (Margolis and Saly, 2007). Air-bone gaps at 4 kHz were

not considered given uncertainties in the measurements for that specific frequency in cases of sensorineural HL (Margolis *et al*., 2013).

	Categories	**Rules**
Severity	Low or no hearing loss	PTA < 20 dB HL
	Mild	20 ≤ PTA < 40 dB HL
	Moderate	40 ≤ PTA < 70 dB HL
	Severe	70 ≤ PTA < 95 dB HL
	Profound	PTA ≥ 95 dB HL
	Not classified	PTA could not be calculated for that ear
Asymmetry	Asymmetric HL	Asymmetry is considered when there are three or more interaural differences (ID) ≥ 10 dB, two or more ID ≥ 15 dB, or one ID ≥ 20 dB (Margolis and Saly, 2007)
	Symmetric HL	
	Not classified *	
Audiogram configuration	Flat	Based on Demeester *et al*. (2009) and Hannula *et al*. (2011)
	High freq. gently sloping (HFGS)	
	High freq. steeply sloping (HFSS)	
	Low frequency ascending (LFA)	
	Mild frequency U-shape (MFU)	
	Mild freq. reverse U-shape (MFRU)	
	Unspecified *	
Type of lesion	No hearing loss	A conductive component is considered when there is a 10-dB air-bone gap (ABG) at three or more frequencies (within 0.25 – 2 kHz), or a 15-dB ABG at any one frequency (within 0.5 – 2 kHz)
	Conductive	
	Sensorineural	
	Mixed	
	Unspecified *	

* Asymmetry, type of lesion or audiogram configuration could not be defined due to data limitation.

Table 1: Hearing loss classification scheme used for the HESD database.

RESULTS

The final number of pure-tone audiograms available in the HESD dataset is 271,575 (Figure 2), which corresponds to hearing data available for 143,794 adults. The data cleaning step was characterized by a drop of 260,894 observations. This is explained by the elimination of blank audiograms, audiograms with a large amount of missing data, repeated curves and invalid measurement due to, for example, testing. The largest drop, however, was in the data transformation stage, where the number of observations was reduced by 417,142. This drop is due to the unstacking of the data, meaning that the information that was previously arranged in several rows (i.e., observations) is now organized in a single row.

Out of the total number of audiograms, 77% presents data for BC thresholds, 38% presents data for CL acoustic reflexes, 29% presents data for IL acoustic reflexes and 84% presents data for speech audiometry. The results displayed in Table 2 reveal that 68% of the audiograms were obtained for older adults (≥ 60 years) and 54% were obtained for male patients. Out of the 143,794 patients, 80,069 (i.e., 55.7%) presented data for only one audiogram in the final dataset.

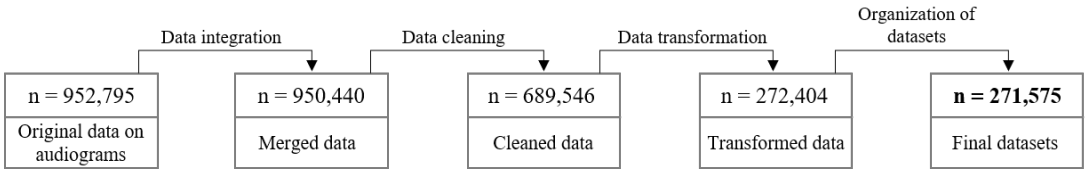


Fig. 2: Number of observations at each of the main data preprocessing stages.

Descriptive statistics on the distribution of hearing loss characteristics (Table 2) for all audiograms available in the HESD database showed asymmetric (53%) and moderate HL (47% left ear and 46% right ear) as the most frequent cases observed in the database. In terms of the type of lesion, the highest prevalence found was for sensorineural (38% left, 37% right), followed by mixed HL (12% left, 13% right).

The most prevalent configuration among audiograms was HFSS (48% left, 46% right), followed by HFGS (23% for both ears) and flat (14% left, 15% right). A chi-squared test of proportion revealed that the distribution of audiogram configurations was significantly different between the left and the right ear (p-value < 0.01).

DISCUSSION

We presented a recently established database that assembles hearing examination data for 143,794 adults (i.e., 271,575 records) who have undergone audiometric testing in the public system of Southern Denmark. The raw data gathered by the AuditBase system required an intensive preprocessing procedure, which also included the development of variables able to classify hearing, more specifically in terms of severity, asymmetry, audiogram configuration and type of lesion.

A variety of definitions for audiometric categorization can be found in literature; however, there is a lack of consistency and standardization among them (Margolis and Saly, 2008). In our study, some of the derived rules were based on the classification system developed by Margolis and Saly (2007), which has been previously validated and defined in order to maximize the agreement among the expert judges involved. This was the case for the definition of asymmetric HL in our study. Our results have shown that the majority of the audiograms were correspondent to asymmetric HL. This high rate of asymmetry may be explained by the fact that these definitions were considerably broad as they, for example, do not require interaural AC thresholds differences at consecutive frequencies.

Given the high number of missing BC thresholds, we were unfortunately unable to categorize the type of lesion for 43% of the cases. Nevertheless, our results showed sensorineural HL as the most predominant type of lesion, followed by mixed HL. Similar results were also found in a previous study from Margolis and Saly (2008), after analyzing audiometric records for a large sample of 16,818 patients.

In terms of the audiogram configuration, we found steep sloping AC curves (i.e., HFSS) to be the most prevalent shape for audiograms, followed by gently sloping

(i.e., HFGS). This result is in agreement with Hannula *et al*. (2011), who have assessed the prevalence of audiogram configurations among 850 adults between 54 and 66 years old, using similar configuration categories as in our study. On the other hand, Demeester *et al*. (2009), who have also used similar methodology, have found flat audiograms as the most prevalent configuration. However, the prevalence for HFGS and HFSS configuration was also shown to be high for their study sample. It is important to point out the limitations of the comparison between results found in the HESD and these studies, as there are fundamental sampling differences (i.e., the HESD is based on clinical data of adults of all ages, whereas the other studies are based on population data of adults between 54-66 years old).

Age at exam		
< 60 years	87,039 (32)	
≥ 60 years	184,536 (68)	
Sex *		
Male	145,565 (54)	
Female	124,211 (46)	
Asymmetry		
Asymmetric HL	144,953 (53)	
Symmetric HL	112,505 (42)	
Not classified	14,117 (5)	
Severity	**Left ear**	**Right ear**
Low or no hearing loss	31,017 (11)	33,247 (12)
Mild	75,181 (28)	76,555 (28)
Moderate	128,232 (47)	124,520 (46)
Severe	22,432 (8)	22,553 (8)
Profound	7,035 (3)	7,030 (3)
Not classified	7,678 (3)	7,670 (3)
Audiogram configuration	**Left ear**	**Right ear**
Flat	38,935 (14)	42,413 (15)
High freq. gently sloping (HFGS)	62,595 (23)	63,579 (23)
High freq. steeply sloping (HFSS)	131,145 (48)	125,250 (46)
Low frequency ascending (LFA)	4,639 (2)	4,964 (2)
Mild frequency U-shape (MFU)	1,436 (1)	1,520 (1)
Mild freq. reverse U-shape (MFRU)	4,362 (2)	4,133 (2)
Unspecified	28,463 (10)	29,716 (11)
Type of lesion	**Left ear**	**Right ear**
No hearing loss	15,816 (6)	17,625 (6)
Conductive	3,209 (1)	3,346 (1)
Sensorineural	102,122 (38)	101,076 (37)
Mixed	33,458 (12)	34,158 (13)
Unspecified	116,970 (43)	115,370 (43)

* Sex data was not available for 1799 audiograms.

Table 2: Demographics and prevalence of HL characteristics in terms of asymmetry, severity, type of lesion and configuration. Results are given in number (%).

Even though the HESD database demanded intensive preprocessing steps, it is remarkable the amount of hearing data that has been merged into a single database. The insights obtained in the study highlight the potential of the HESD database as a promising source of audiology-related epidemiological data, not just to evaluate hearing profiling among adults, but to further explore the effects of hearing impairment on a range of health outcomes.

REFERENCES

Agrawal, Y., Platz, E.A., and Niparko, J.K. (**2009**). "Risk Factors for Hearing Loss in US Adults: Data from the National Health and Nutrition Examination Survey, 1999 to 2002," Otol. Neurotol., **30**, 139-145.

Cunningham, L.L., and Tucci, D.L. (**2017**). "Hearing Loss in Adults," New Engl. J. Med., **377**(25), 2465-2473. doi:10.1056/NEJMra1616601.

Demeester, K., Wieringen, A., Hendrickx, J., Topsakal, V., Fransen, E., Laer, L., Camp, G.V., and Heyning, P.V. (**2009**). "Audiometric shape and presbycusis," Int. J. Audiol., **48**, 222-232. doi: 10.1080/14992020802441799

Elberling, C., Ludvigsen, C.W., and Lyregaard, P.E. (**1989**). "DANTALE: a new Danish speech material," Scand. Audiol., **18**(3), 169-175.

Hannula, S., Bloigu, R., Majamaa, K., Sorri, M., and Mäki-Torkko, E. (**2011**). "Audiogram configurations among older adults: Prevalence and relation to self-reported hearing problems," Int. J. Audiol., **50**, 793-801. doi: 10.3109/14992027.2011.593562

ISO. (**2011**) ISO 14155:2011 "Clinical investigation of medical devices for human subjects – Good clinical practice." Geneva.

ISO. (**2010**) ISO 8253-1:2010 "Acoustics — Audiometric test methods — Part 1: Pure-tone air and bone conduction audiometry." Geneva.

Margolis, R.H., and Saly, G.L. (**2007**). "Toward a standard description of hearing loss," Int. J. Audiol., **46**, 746-758. DOI: 10.1080/14992020701572652.

Margolis, R.H., and Saly, G.L. (**2008**). "Distribution of Hearing Loss Characteristics in a Clinical Population," Ear Hearing, **29**(4), 524-532. doi: 10.1097/AUD.0b013e3181731e2e

Margolis, R.H., Eikelboom, R.H., Johnson, C., Ginter, S.M., Swanepoel, D.W., and Moore, B.C.J. (**2013**). "False air-bone gaps at 4 kHz in listeners with normal hearing and sensorineural hearing loss," Int. J. Audiol., **52**(8), 526–532. doi:10.3109/14992027.2013.792437.

Stevens, G., Flaxman, S., Brunskill, E., Mascarenhas, M., Mathers, C. D., and Finucane, M. (**2011**). "Global and regional hearing impairment prevalence: an analysis of 42 studies in 29 countries," Eur. J. Public Health, **23**(1), 146-152. doi:10.1093/eurpub/ckr176

Thomson, R.S., Auduong, P., Miller, A.T., and Gurgel, R.K. (**2017**). "Hearing loss as a risk factor for dementia: A systematic review," Laryngoscope Invest. Otolaryngol., **2**, 69-79. doi: 10.1002/lio2.65

Investigating the relationship between spectro-temporal modulation detection, aided speech perception, and directional noise reduction preference in hearing-impaired listeners

JOHANNES ZAAR[1,*], LISBETH BIRKELUND SIMONSEN[2], THOMAS BEHRENS[3], TORSTEN DAU[1] AND SØREN LAUGESEN[2]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Interacoustics Research Unit, DK-2800 Lyngby, Denmark*

[3] *Oticon A/S, DK-2765 Smørum, Denmark*

In analogy to the restoration of reduced audibility via hearing-aid amplification, supra-threshold speech processing deficits may be partially compensated for by using state-of-the-art directional noise reduction (NR) techniques. However, while amplification is usually prescribed based on classical audiometry, a clinical test that represents supra-threshold speech processing and is thus useful for prescribing NR settings is yet to be established. The present study explored the potential of a suitably adapted spectro-temporal modulation detection (STMD) test for this purpose by means of laboratory-based tests and field tests with 30 hearing-impaired participants. In particular, it was investigated whether STMD performance (i) predicts aided speech intelligibility measured in a spatial multi-talker set up with different degrees of NR and (ii) predicts preference for moderate vs. aggressive NR. STMD thresholds were strongly correlated with (i) speech scores measured without NR, (ii) speech intelligibility benefit induced by aggressive NR, and (iii) the individual participants' NR preference. The latter relationship was mediated by performance in a reverse digit span task, which measures working memory capacity. Overall, the results suggest that a clinical test that assesses STMD sensitivity may be useful for prescribing NR settings in hearing-aid fitting.

INTRODUCTION

Hearing-aid amplification is typically tailored to the individual's hearing loss based on the pure-tone audiogram to restore audibility. However, some individuals experience severe supra-threshold difficulties with speech understanding in adverse conditions that cannot be resolved by audibility compensation. For those, additional help may be provided in the form of minimum variance distortionless response beamforming combined with single-channel noise reduction (MVDR-NR, in the following simply termed NR). Recent advances have yielded the possibility to substantially improve speech intelligibility (SI) and reduce listening effort – at least

*Corresponding author: jzaar@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

in laboratory-based scenarios – when using aggressively parametrized NR. However, these improvements may come at the cost of an impaired perceived naturalness of the sound scene and its acceptance may therefore be highly listener-specific. To use its full potential, NR thus needs to be carefully tailored to the individual, such that aggressive NR settings are only prescribed to those who truly need and therefore tolerate them.

To this end, a clinically viable measure that represents supra-threshold speech processing is required to identify listeners with severe supra-threshold deficits, who might benefit from aggressive NR. Bernstein *et al.* (2013) employed a spectro-temporal modulation detection (STMD) paradigm to assess such deficits in normal-hearing (NH) and HI listeners. They used broadband (354-5656 Hz) noise carrier signals modulated with various STM patterns (specified by spectral modulation rate in cycles/octave, c/o, and temporal modulation rate in Hz), generating spectral ripples that move upward or downward as a function of time. Bernstein *et al*. (2013) found a significant NH vs. HI difference for the combination of 2 c/o and 4 Hz, which has henceforth been widely used. Furthermore, Bernstein *et al*. (2013) and Mehraei *et al.* (2014) demonstrated that STMD performance was strongly correlated with speech-in-stationary-noise performance in HI listeners, measured at very high presentation levels but without individualized amplification. Bernstein *et al*. (2016) measured STMD performance in HI listeners using a bandlimited noise carrier (354-2000 Hz) with 2 c/o and 4 Hz. STMD thresholds were compared to speech reception thresholds (SRTs) measured in stationary noise and multi-talker babble with simulated hearing-aid processing (i.e., aided). While they found a significant correlation between STMD thresholds and SRTs, the relationship was not as strong as in previous studies, possibly because many listeners could not reach the required detection accuracy, such that thresholds could not be directly measured and instead had to be extrapolated. Introducing various changes to the measurement procedure, such as listener-specific frequency-dependent amplification, increased stimulus duration, and bilateral presentation mode, the authors of the current study proposed an STMD measurement procedure that all tested HI listeners were sensitive to (Zaar *et al*., 2018). Furthermore, the study demonstrated that STMD performance was strongly associated with speech-in-noise performance, measured in co-located stationary noise as well as in a spatial multi-talker set-up with audibility compensation.

Based on the observations and findings described above, the goal of the present study was to explore whether STMD performance, as measured in Zaar *et al*. (2018), is indicative of supra-threshold speech processing deficits in HI individuals and thus predictive of the individuals' preference in terms of NR settings. The following three research questions (RQs) were addressed:

[RQ1] Does STMD performance predict aided SI without NR?

[RQ2] Does STMD performance predict SI benefit offered by aggressive NR?

[RQ3] Does STMD performance predict NR preference?

METHODS

Participants, hearing-aid fitting and NR settings

30 HI participants (mean age: 70.2 years, standard deviation: 9.1 years) were recruited, all of whom were native speakers of Danish and regular hearing-aid users. All participants underwent audiometric screening. Hearing aids were fitted based on the individuals' audiograms using the standard prescription offered by the fitting software. Three NR settings were defined: "*Off*" (NR algorithm inactive, hearing-aid directivity pattern in omni-directional mode), "*Default*" (mode-rate parametrisation of the NR algorithm); "*FullThrottle*" (customized aggressive NR setting).

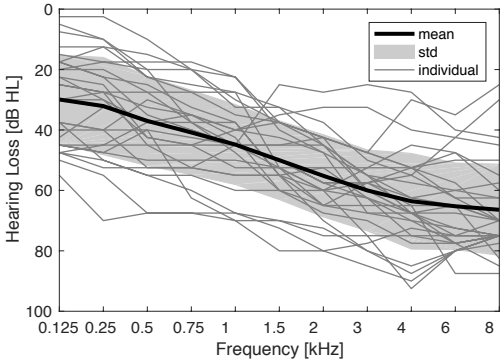


Fig. 1: Pure-tone thresholds (averaged across ears) for all participants (thin grey lines) and on average (thick black line).

Reverse digit span (RDS) test

A measure of working memory capacity was obtained using the reverse digit span (RDS) test, where randomly selected Danish digits from "1" to "9" were presented to the participants at their self-adjusted most comfortable level over Sennheiser HDA200 headphones in a sound attenuated booth. Initially, two digits were presented in each trial. The number of digits per trial was then increased by one after each second trial. The procedure ended after two incorrect responses or after 14 trials (with maximally 8 digits per trial). The participants were required to type the digits they had heard in reverse order on a computer keyboard. In each trial, two points could be obtained, one for the correct number of repeated digits and one for the correct placement of the digits. The maximum possible cumulative score was thus 28.

Spectro-temporal modulation detection (STMD) test

STMD thresholds were adaptively measured using a three-alternative forced choice procedure with a one-up/two-down tracking rule (approaching 70.7 percent correct). The STM stimulus was generated by modulating a bandlimited noise carrier (354-2000 Hz, 1000 log-spaced random-phase sinusoidal components per octave) with an upward moving ripple pattern defined by spectral and temporal modulation rates of 2 c/o and 4 Hz, respectively. In each trial, the two noise-only intervals contained an identical realization of the carrier noise and the signal interval contained the same carrier noise with the modulation imposed on it. The STM starting phase in the modulated stimulus was randomized across trials. The stimulus duration was 1 s with 500 ms inter-stimulus intervals. The modulation depth was considered as the tracking variable, which started at 0 dB (full modulation). The initial step size was 4 dB, which was halved after the first and again after the second upper reversal. After 8 reversals at a step size of 1 dB the procedure terminated. The threshold was calculated as the

mean across the tracking variable at the 8 last reversals. The participants were seated in a sound attenuating booth in front of a computer screen and bilaterally presented with the stimuli using Sennheiser HDA200 headphones. The nominal presentation level was set at a sound pressure level (SPL) of 65 dB and ear-specific linear amplification was applied where necessary to ensure at least 15 dB sensation level in each 3rd-octave band within the stimulus frequency range. The participants provided their responses using a touch screen, a computer keyboard, or a computer mouse, according to their preference. They received visual feedback after each response (correct/incorrect). A short training run was provided by means of a simple amplitude modulation detection task (using broadband noise with a 4-Hz modulation) in order to familiarize the participants with the procedure. Three adaptive measurements were conducted and the median of the resulting three thresholds was considered as the final threshold.

Speech-in-noise test

Speech intelligibility was measured using the Danish hearing in noise test (HINT, Nielsen and Dau, 2011) using a spatial loudspeaker set up in a quiet but slightly reverberant room. Target sentences spoken by a male talker were presented from a frontal location (0º azimuth angle) at 65 dB SPL(C). Running speech interferers spoken by two different male talkers, mixed with low-level speech-shaped noise (-6 dB relative to the running speech level), were played from two loudspeakers positioned at ±100º azimuth. The participants were seated in the middle of the loudspeaker arrangement wearing hearing aids and instructed to use a headrest to maintain a static head position. They were asked to verbally repeat the target-sentence words they had understood, which were then manually scored by an audiologist. SRTs were tracked by adjusting the level of the interferers (i.e., the signal-to-noise ratio, SNR) according to sentence correct scoring (see Nielsen and Dau, 2011). The resulting data were analysed using the method suggested by Rønne et al. (2017) to obtain SRTs relating to 50% sentences correct. Two trainings runs were conducted with NR *Off* using one HINT list (20 sentences) for the first and two lists (40 sentences) for the second run. SRTs were then measured for each NR setting (*Off*, *Default*, *FullThrottle*) using two HINT lists (40 sentences). The presentation order was balanced across participants.

Field testing and questionnaires

The participants were provided with the test hearing aids for two successive field trial periods (3-5 weeks each), with *Default* NR in one period and *FullThrottle* NR in the other one. The order was balanced across participants who were unaware of the difference between the two settings. After the first trial period, participants were asked to fill out the SSQ12 questionnaire (Noble *et al*., 2013); after the second trial period, they were asked to fill out the comparative version of the questionnaire (SSQ12-C; Jensen *et al*., 2009). The participants were asked to rate their preference for the first or second setting on a 5-point scale (-2, -1, 0, 1, 2), where the extremes indicated strong preference for either of the two settings and the midpoint indicated no

preference. In addition, the participants were asked to indicate their level of certainty regarding their preference on a scale from 0 (very uncertain) to 10 (very certain). The preference ratings were multiplied with the certainty ratings (normalized by 10) to obtain the final preference score and then processed such that positive values reflect preference for *FullThrottle* NR and negative values preference for *Default* NR.

RESULTS & ANALYSIS

Effect of NR settings on speech intelligibility

Fig. 2 shows the average SRTs and across-participant standard deviations measured for the three NR settings. As can be seen, *Default* NR yielded about 2 dB and *FullThrottle* NR about 4 dB SRT benefit as compared to NR *Off*. The large standard deviations indicate substantial performance differences across participants. A two-way ANOVA with NR setting as a fixed factor and participant as a random factor showed highly significant ($p<<0.001$) main effects of NR and participant. A post-hoc analysis revealed that the different NR settings were all significantly ($p<0.001$) different from each other.

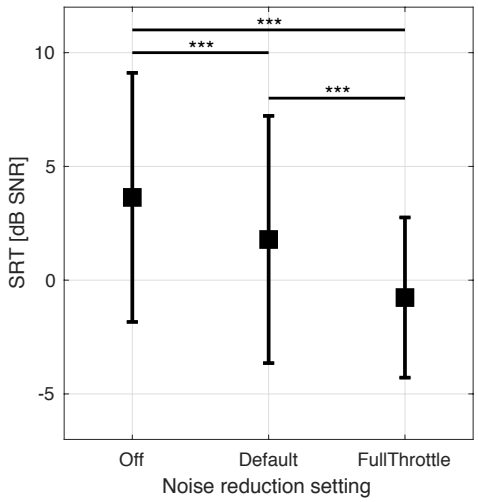


Fig. 2: Mean and standard deviations of SRTs across participants. ***: $p<0.001$.

RQ1: Does STMD performance predict aided SI without NR?

The left panel of Fig. 3 shows the SRT_{Off} (SRTs measured with NR *Off*) as a function of the STMD thresholds. A highly significant positive correlation between the two measures can be observed, with an R-squared of 0.63 and $p<0.001$. The answer to RQ1 is thus "yes". However, other measures may also yield good predictions of SRTs. The middle panel of Fig. 3 shows the percentage of variance explained (i.e., the R-squared in percent) for various predictors. While STMD thresholds accounted for the largest amount of variance in SRT_{Off} (63%), the average pure-tone thresholds between 0.125 and 8 kHz (PTA) also accounted for a substantial 58% ($p<0.001$). Performance on the RDS test accounted for a much smaller yet significant ($p<0.05$) 26% of the variance in SRT_{Off}, whereas age showed no effect. Finally, the right panel of Fig. 3 addresses the question of how much additional predictive power can be provided by the individual predictors beyond that provided by the PTA. Linear regression models were employed with PTA as the first predictor and STMD thresholds, RDS scores, or age as the second predictor. It can be seen that the STMD thresholds accounted for an additional 15% ($p<0.01$) of SRT_{Off} variance explained, amounting to 73% overall. The RDS scores also added a significant amount ($p<0.05$) of 10% SRT_{Off} variance

explained. The contribution of age was again not significant. All reported significance levels were Bonferroni-corrected for multiple comparisons.

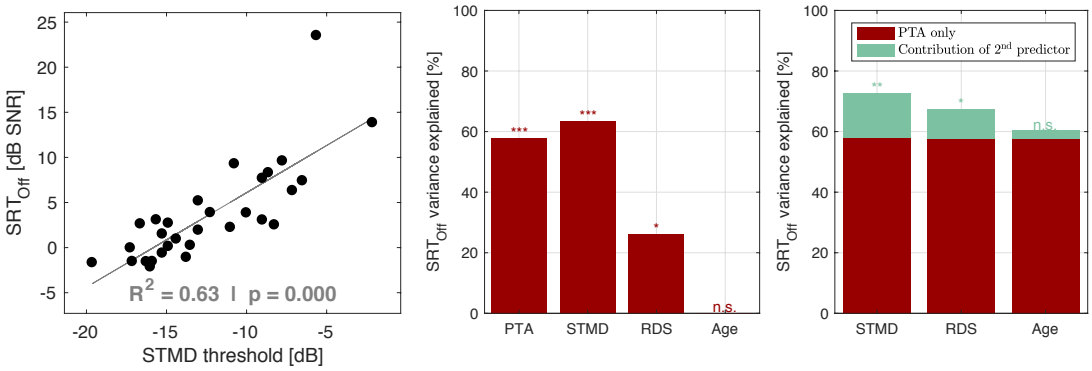


Fig. 3: Left: SRT_{Off} as a function of STMD thresholds along with regression line fit. Middle: percent of SRT_{Off} variance explained by various predictors. Right: percent of SRT_{Off} variance explained by various predictors in addition to PTA in two-predictor linear regression model.

RQ2: Does STMD performance predict SI benefit offered by aggressive NR?

A measure of the SRT benefit induced by *FullThrottle* as compared to *Default* NR is $\Delta SRT = SRT_{Default} - SRT_{FullThrottle}$. Positive ΔSRT values thus indicate an increase in SI induced by *FullThrottle* NR. Fig. 4 shows ΔSRT as a function of the STMD thresholds. A substantial and highly significant positive correlation ($p<0.001$) can be observed, with 51% of the variance in ΔSRT explained by STMD performance. The answer to RQ2 is therefore "yes".

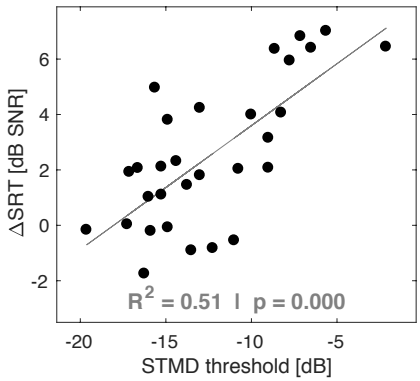


Fig. 4: ΔSRT as a function of STMD thresholds along with regression line.

RQ3: Does STMD performance predict NR preference?

Real-world benefit was evaluated using both the SSQ12-C questionnaire and the participants' preference ratings. Tab. 1 shows the correlations between the preference ratings and the average SSQ12-C score as well as the speech, spatial, and quality subsets SSQ12-C. The preference ratings were almost perfectly correlated with the SSQ12-C scores, except for the spatial subset of SSQ12-C, indicating that the preference ratings were mainly driven by advantages/disadvantages in terms of speech understanding and sound quality. The left panel of Fig. 5 shows the NR preference ratings as a function of STMD thresholds, indicating no obvious relationship between the two measures. However, six participants for whom the hearing-aid logging data indicated little exposure to non-quiet acoustical scenarios were discarded, as well as

another two participants with technical problems related to the hearing-aid fitting. For the remaining 22 participants, a significant correlation was found ($p<0.05$, R-squared of 0.23; middle panel of Fig. 5). Additionally, two parallel "correlated patterns" can be observed in the middle panel of Fig. 5, which were connected to the RDS scores (i.e., working memory capacity), as indicated by diamonds ("good" RDS>11.8) and squares ("poor" RDS<11.8). The working memory capacity thus appeared to influence the STMD performance but not the NR preference ratings. The right panel of Fig. 5 shows the preference ratings as a function of RDS-corrected STMD thresholds (obtained by subtracting the STMD thresholds predicted by RDS scores using linear regression from the actual STMD thresholds), indicating a highly significant ($p<0.001$, R-Squared of 0.5) correlation with the preference ratings.

	SSQ12-C_{All}	SSQ12-C_{Speech}	SSQ12-$C_{Spatial}$	SSQ12-$C_{Quality}$
NR Pref.	0.88***	0.88***	0.48**	0.83***

Table 1: Pearson's correlation between NR preference ratings and SSQ12-C scores averaged across all 12 questions and across the respective subsets related to speech, spatial, and quality aspects. ***: $p<0.001$; **: $p<0.01$.

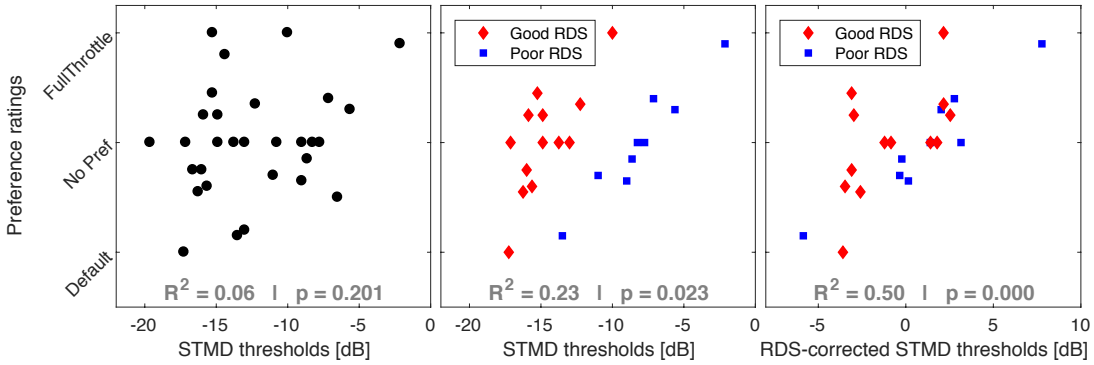


Fig. 5: Left: preference ratings as a function of STMD thresholds. Middle: same as left with 8 participants excluded, symbols according to RDS scores. Right: same as middle with RDS scores factored out of STMD thresholds.

The RDS scores were significantly negatively correlated both with STMD thresholds ($r = -0.53$, $p<0.01$) and SRT_{Off} ($r = -0.51$, $p<0.01$) to the same extent, indicating that working memory capacity positively affected performance both in STMD and speech-in-noise. However, the preference ratings showed no correlation with the RDS scores ($r = 0.01$, $p>0.05$), which merely acted as a mediator variable between the STMD thresholds and the preference ratings. The answer to RQ3 is thus "yes, for most listeners and best in combination with RDS".

CONCLUSIONS

The present study demonstrated that STMD performance as measured with the proposed paradigm (i) can serve as a highly reliable proxy for aided speech-in-noise

perception in a spatial multi-talker set up (RQ1), (ii) is strongly correlated with SRT improvement offered by the aggressive NR considered here (RQ2), and (iii) appears to be associated with NR preference reported by participants of a field study (mediated by working memory capacity, RQ3). These findings suggest that a clinical measure of STMD sensitivity may yield a powerful predictor of supra-threshold speech processing, which could be translated to prescriptions of NR settings in hearing aids.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support from the William Demant Foundation, as well as contributions to this study from Bue Kristensen, James Harte, Valentina Campagnaro, Thomas Lunner, Fares El-Azm, Jacob Aderholt, Erengül Sabedin, and Sébastien Santurette.

REFERENCES

Bernstein, J.G.W., Mehraei, G., Shamma, S., Gallun, F.J., Theodoroff, S.M., and Leek, M.R. (**2013**). "Spectro-temporal modulation sensitivity as a predictor of speech intelligibility for hearing-impaired listeners," J. Am. Acad. Audiol., **24**(4), 293-306. doi:10.3766/jaaa.24.4.5

Bernstein, J.G.W., Danielsson, H., Hällgren, M., Stenfelt, S., Rönnberg, J., and Lunner, T. (**2016**). "Spectro-temporal modulation sensitivity as a predictor of speech-reception performance in noise with hearing aids," Trends Hear., **20**, 1-17. doi: 10.1177/2331216516670387

Jensen, N.S., Akeroyd, M.A., Noble, W., and Naylor, G. (**2009**). "The Speech Spatial and Qualities of Hearing scale (SSQ) as a benefit measure," Poster presented at 4th NCRAR international conference, Portland, US.

Nielsen, J.B. and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol., **50**, 202-208. doi: 10.3109/14992027.2010.524254

Mehraei, G., Gallun, F.J., Leek, M.R., and Bernstein, J.G.W. (**2014**). "Spectro-temporal modulation sensitivity for hearing-impaired listeners: dependence on carrier center frequency and the relationship to speech intelligibility," J. Acoust. Soc. Am., **136**(1), 301-316. doi: 10.1121/1.4881918

Noble, W., Jensen, N.S., Naylor, G., Bhullar, N., and Akeroyd, M.A. (**2013**). "A short form of the Speech, Spatial and Qualities of Hearing scale suitable for clinical use: The SSQ12," Int. J. Audiol., **52**, 409-412. doi: 10.3109/14992027.2013.781278

Rønne, F.M., Laugesen, S., and Jensen, N.S. (**2017**). "Selection of test-setup parameters to target specific signal-to-noise regions in speech-on-speech intelligibility testing," Int. J. Audiol., **56**, 559-567. doi: 10.1080/14992027.2017.1300349

Zaar, J., Simonsen, L.B., Behrens, T., Dau, T., and Laugesen, S. (**2018**). "Towards a clinically viable spectro-temporal modulation test," Poster presented at the international hearing aid research conference (IHCON), Lake Tahoe, US.

Effects of directional hearing aid processing and motivation on EEG responses to continuous noisy speech

TOBIAS NEHER[1,*], BOJANA MIRKOVIC[2,3] AND STEFAN DEBENER[2,3]

[1] *Institute of Clinical Research, University of Southern Denmark, Denmark*

[2] *Department of Psychology, University of Oldenburg, Germany*

[3] *Cluster of Excellence "Hearing4all", Oldenburg, Germany*

Arguably, the next frontier in hearing aid (HA) development are devices that can infer (or learn) the needs of the user via non-invasive physiological measurements such as electroencephalography (EEG) and adjust themselves accordingly. A promising approach to translating EEG signals into HA control signals is the analysis of EEG impulse responses to running speech, as obtained by cross-correlating the audio stimulus with the concurrently recorded EEG signal. Here, we used this method for examining neural correlates of the effects of directional HA processing and listener motivation on speech comprehension in noise. Groups of older participants with normal or impaired hearing listened to audiobook material embedded in realistic cafeteria noise while their EEG was recorded using mobile hardware. A HA simulator was used for (dis)engaging a directional microphone setting and for providing amplification. Motivation was manipulated by offering a monetary reward for good speech comprehension in half of the trials. Motivation influenced the participants' listening performance but not their EEG responses. Directional HA benefit, however, was reflected in both the behavioural and EEG data, thereby illustrating the potential of the tested approach with respect to enabling online HA control.

INTRODUCTION

In clinical practice, hearing aids (HAs) are fitted based on the pure-tone audiogram and feedback from the user (Dillon, 2012). Both types of responses are subjective in nature and therefore prone to bias. Physiological measures of the ability to process speech in noise, on the other hand, could provide an objective basis for both hearing assessment and HA adjustment. Recently, a number of studies investigated the potential of electroencephalography (EEG) measurements with respect to HA adjustments (e.g., Bernarding *et al*., 2017; Van Eyndhoven *et al*., 2017). A potential advantage of EEG-controlled HAs would be direct access to the cognitive state of the user, which could enable automatic HA adjustments in response to changes in the acoustic environment or the user's intent. However, a prerequisite for controlling a HA based on ongoing EEG signals is a robust neural marker that indexes the relevant cognitive processes reliably. In challenging situations, listeners modulate their attention based on the physical properties of the auditory scene (e.g., changes in the signal-to-noise ratio; SNR) and their interest in specific parts of the scene

*Corresponding author: tneher@health.sdu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

(e.g., their intent or motivation). Thus, when investigating EEG markers for HA control, the influence of both bottom-up (acoustic) and top-down (listener-driven) factors needs to be considered.

The current study investigated EEG correlates of both types of factors with a view towards enabling online HA control. More specifically, it focused on the effects of SNR improvement as brought about by directional HA processing as well as listener motivation with respect to speech comprehension in noise. Below, a summary of the methods and results is provided. Parts of them are based on Mirkovic *et al*. (2019).

METHODS

Participants

Out of 38 recruited participants, 16 normal-hearing (NH; mean age = 67 yrs, range: 62-75 yrs) and 15 hearing-impaired (HI; mean age = 74 yrs, range: 63-88 yrs) participants completed the entire study. The normal-hearing participants were required to have pure-tone average hearing losses as calculated across 0.5, 1, 2 and 4 kHz (PTA4) of less than 25 dB HL in both ears. The hearing-impaired participants were required to have PTA4s of at least 35 dB HL in both ears. Figure 1 shows the average audiograms of these two groups. The HI participants were all bilateral HA users with at least six months of HA experience.

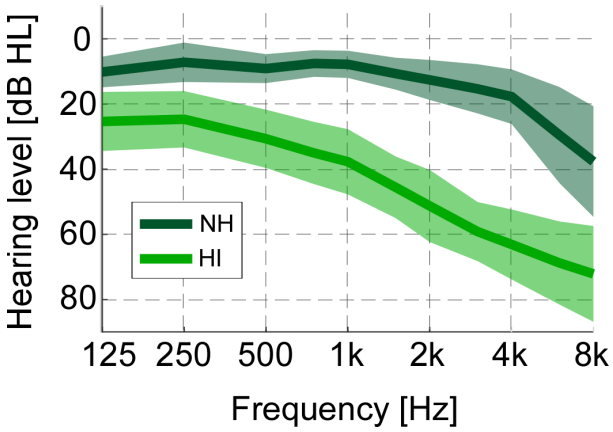


Fig. 1 (colour version online): Mean audiograms of the NH (dark green) and HI (light green) participants. Shaded areas represent ±1 standard deviation (SD).

Experimental paradigm and stimuli

Following the pure-tone audiometry, individual speech reception thresholds (SRTs) were measured using the procedures of Neher *et al*. (2017). The speech material from the Oldenburg sentence test (Wagener *et al*., 1999) was used as target speech. To create the perception of a spatial auditory scene, the target sentences were convolved with head-related impulse responses recorded in an empty cafeteria with

a frontal (0°) source and HA dummies placed on a head-and-torso simulator (Kayser *et al*., 2009). As background noise, a recording of the fully occupied cafeteria was used. To compensate for the raised hearing thresholds of the HI participants, individual amplification according to the "National Acoustic Laboratories-Revised Profound" prescription rule (Dillon, 2012) was applied to the stimuli using a HA simulator (Grimm *et al*., 2006). All stimuli were presented binaurally to the participants via insert earphones.

The main task of the participants was to listen to a continuous speech stream masked by the aforementioned cafeteria noise in twelve 10-min long recording sessions. The continuous speech stream consisted of concatenated audiobooks. The cafeteria noise was presented at a nominal level of 65 dB SPL. The target speech was adjusted in level to result in the individual SRTs at the input of the HA simulator. After each recording session, the participants had to answer questions about the contents of the audiobooks. Responses to these questions were used to assess listening performance.

To investigate the effect of motivation, a monetary reward was offered in half of the trials. Below, the two resultant conditions will be referred to as 'motivated' and 'unmotivated'. To investigate the effect of SNR changes, a directional microphone setting was (dis)engaged after each minute of listening. The directional microphone setting corresponded to two (left and right) static forward-facing cardioid microphones. The other setting corresponded to two (left and right) omnidirectional microphones. The directional setting resulted in a speech-weighted SNR improvement of 3.5 dB relative to the omnidirectional setting. Below, these two conditions will be referred to as 'low SNR' and 'high SNR'.

Recording and analysis of EEG signals

While listening to the audiobooks, the participants' EEG signals were recorded using 24 Ag-AgCl electrodes distributed according to the 10-20 system in customised caps. All cap channels were referenced to the FCz channel. The recordings were made with a sampling frequency of 500 Hz using a lightweight, wireless mobile EEG amplifier placed on the back of the EEG cap (see Figure 2).

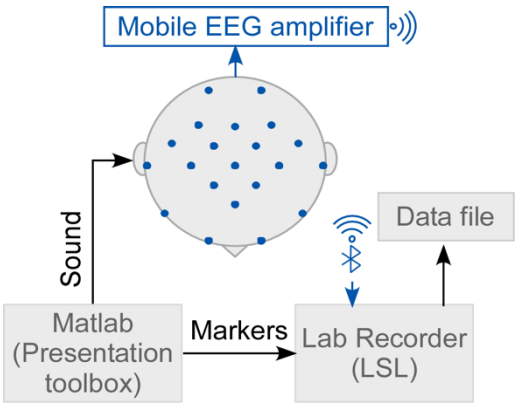


Fig. 2: Illustration of the experimental test setup.

All EEG signal processing was carried out using customized MATLAB scripts as well as the EEGLAB toolbox (Delorme and Makeig, 2004). The raw EEG data were re-referenced offline from the FCz electrode to the algebraic average of the TP9 and TP10 channels to obtain the equivalent of a (commonly used) linked-mastoid reference. Furthermore, the data were band-pass filtered from 0.5 to 45 Hz using a 6th-order Butterworth filter, followed by automatic correction for artefacts such as eye blinks and eye movements using artefact subspace reconstruction (ASR; Mullen *et al*., 2015). Finally, the EEG data were downsampled to 125 Hz.

The EEG impulse responses were estimated based on the analysis pipeline of Petersen *et al*. (2016). First, the Hilbert transform was applied to the speech (or noise) envelope. For each participant, the speech (or noise) was extracted from the signal mixture at the output of the HA simulator and low-pass filtered using a 3rd-order Butterworth filter with a cut-off frequency of 25 Hz. To emphasise speech onsets, the first derivative of the speech envelope was calculated and half-wave rectified. The resultant signal was downsampled to 125 Hz and segmented into 1-min long trials. The analysis pipeline is illustrated in Figure 3. The analysis was performed on the left ear-input signals of each participant.

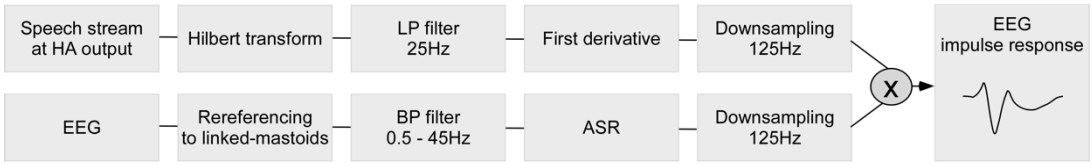


Fig. 3: Illustration of the EEG impulse response analysis pipeline.

Neural oscillations are known to synchronise to the envelope of speech signals – a behaviour that can be modelled by superimposing neural responses to individual speech sounds. The cross-correlation pattern between the resultant EEG envelope and the actual speech envelope reveals the neural impulse response to speech, which is modulated by attention (O'Sullivan *et al*., 2014; Petersen *et al*., 2016). Here, EEG impulse responses were estimated by cross-correlating trials of the extracted speech envelope with corresponding trials of the pre-processed EEG signal, taking into account latencies from -100 to 600 ms. To estimate chance-level synchronisation, EEG trials were also cross-correlated with randomly chosen non-corresponding trials of the speech envelope and used for calculating chance-level EEG impulse responses.

Previous research has found that EEG impulse responses to speech can be physiologically interpreted and that their topography bears similarity to that of auditory evoked potentials (AEPs; Crosse *et al*., 2016). Averages of individual responses were therefore obtained and visualised for the Cz channel where AEPs are traditionally observed. In the resultant grand-average time series, local extremes (or peaks) were identified for each participant, resulting in sets of peak latency and amplitude values that were analysed statistically.

RESULTS

Listening performance

The listening performance of the two participant groups is shown in Figure 4. On average, the NH and HI groups answered, respectively, 68.3% (SD: 12.9%) and 55.3% (SD: 16.7%) of the questions posed correctly. An analysis of variance revealed significant effects of motivation, SNR, group, motivation × SNR and group × motivation (all $p < 0.05$). Closer inspection revealed an influence of motivation in the 'low SNR' ($p < 0.0001$) but not the 'high SNR' ($p > 0.3$) condition. Furthermore, the HI participants performed better in the motivated compared to the unmotivated condition ($p < 0.001$), whereas the NH participants did not ($p > 0.1$).

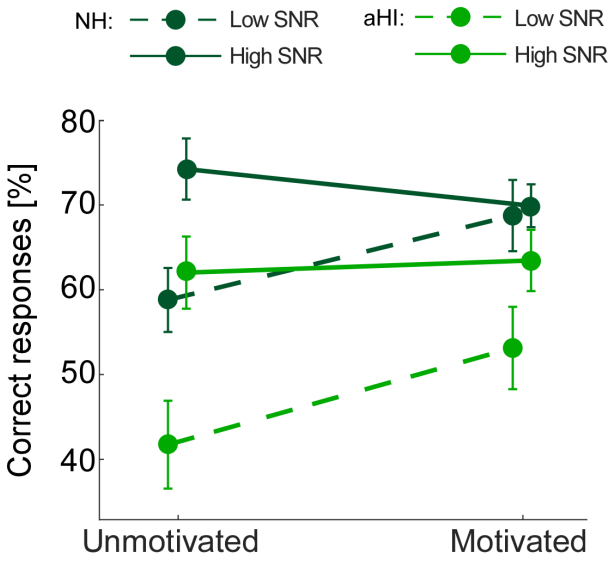


Fig. 4 (colour version online): Mean listening performance of the NH (dark green) and aided HI (light green) groups based on their answers to questions about the contents of the audiobook. Error bars represent ±1 standard error of the mean.

EEG impulse responses

The grand-average EEG impulse responses calculated across all participants are shown in Figure 5. Whereas the chance-level impulse response did not show clear variations in temporal pattern, four peaks were evident in the EEG impulse response for the target speech. The first prominent peak had a latency of 60 ms and was reminiscent of the P100 peak of the traditional AEP response. Responses at this latency are usually characterised as bottom-up-driven. As the focus of this work was on top-down attentional processes, the first peak was not analysed any further. The following analyses were conducted for the other three peaks with latencies of 136 ms ('$N1_{crosscorr}$'), 240 ms ('$P2_{crosscorr}$') and 376 ms ('$N2_{crosscorr}$').

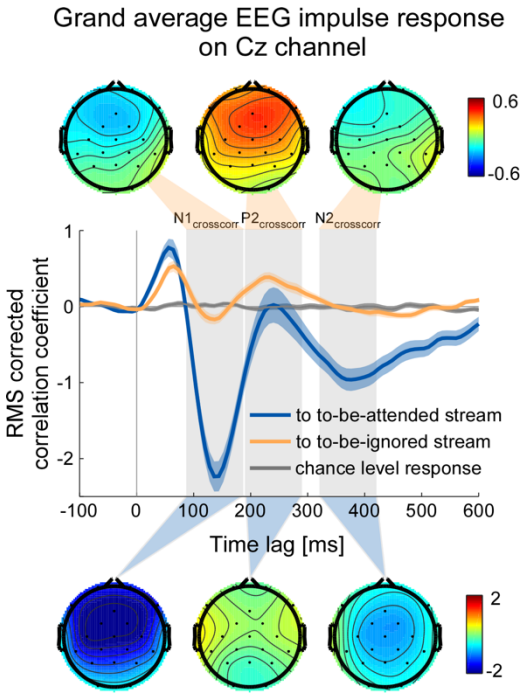


Fig. 5 (colour version online): Grand-average EEG impulse responses for target speech (blue), cafeteria noise (orange) and chance level (grey) across all participants. Shaded areas represent ±1 SD. Topographies corresponding to the grand-average component peaks for the target speech (bottom) and cafeteria noise (top) are also shown.

Analyses of variance carried out on the peak latency and amplitude values revealed significant effects of SNR on $N1_{crosscorr}$ and $P2_{crosscorr}$ latency (both $p < 0.001$) and a significant interaction between group and SNR for the $N2_{crosscorr}$ amplitude ($p < 0.01$). Closer inspection showed faster $N1_{crosscorr}$ and $P2_{crosscorr}$ responses in the 'high SNR' compared to the 'low SNR' condition (see Figure 6). Furthermore, there was a significant influence of SNR on the $N2_{crosscorr}$ amplitude for the HI ($p < 0.01$) but not the NH ($p > 0.1$) participants. Regarding a possible influence of motivation, no effects were observable in any of the $N1_{crosscorr}$, $P2_{crosscorr}$ or $N2_{crosscorr}$ data.

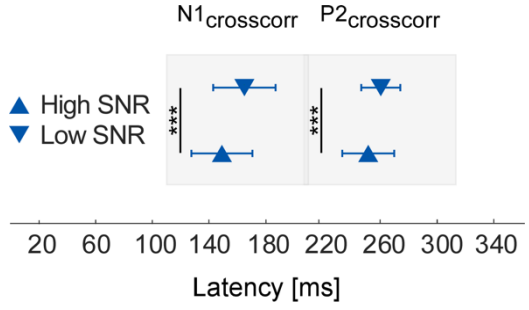


Fig. 6: $N1_{crosscorr}$ and $P2_{crosscorr}$ latencies for the target speech EEG impulse responses when listening with ('high SNR') or without ('low SNR') directional processing. Error bars represent ±1 standard error of the mean.

DISCUSSION

The current study investigated the influence of HA-induced SNR improvement in a realistic auditory scene on both behaviour and neurophysiology. In addition, it manipulated listening motivation using a monetary reward condition. Data were collected using a low-density mobile EEG system and a realistic task, that is, listening to continuous target speech in a spatially complex cafeteria environment. Groups of older NH and HI listeners participated.

As expected, directional HA processing improved listening performance. Furthermore, motivation improved the performance of the HI participants when the directional HA processing was disengaged and listening therefore was demanding. It is currently unclear why the same was not true for the NH listeners who were tested at the same performance level.

The analysis of the EEG impulse responses for the target speech revealed three prominent peaks that are attributable to attentional processes. The topographies and latencies of these peaks corresponded well to those of well-known AEP components (N100 and P200). The latency of these components proved sensitive to the applied SNR changes, with slower responses being evident when the directional HA processing was disengaged. The observed latency effect thus appears to be a physiological measure of the observed behavioural differences.

Regarding the motivational manipulation, no physiological correlates were observable in the EEG data. This was despite the fact that the monetary reward condition improved listening performance, particularly at the low SNR and for the HI group. Thus, further research is needed to investigate top-down influences on EEG responses to continuous speech, perhaps using a different experimental paradigm with a more effective motivational manipulation.

ACKNOWLEDGEMENTS

This research was funded by the DFG Cluster of Excellence EXC 1077/1 “Hearing4all”. The authors thank Julia Schmidt and Manuela Jaeger (University of Oldenburg) for their help with the data collection and data analyses.

REFERENCES

Bernarding, C., Strauss, D. J., Hannemann, R., Seidler, H., and Corona-Strauss, F. I. (**2017**). “Neurodynamic evaluation of hearing aid features using EEG correlates of listening effort,” Cogn. Neurodyn., **11**, 203-215, doi: 10.1007/s11571-017-9425-5.

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (**2016**). “The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli,” Front. Human Neurosci., **10**, 604, doi: 10.3389/fnhum.2016.00604.

Delorme, A., and Makeig, S. (**2004**). "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," J. Neurosci. Methods, **134**, 9-21, doi: 10.1016/j.jneumeth.2003.10.009.

Dillon, H. (**2012**). *Hearing Aids*, 2nd ed., Thieme.

Grimm, G., Herzke, T., Berg, D., and Hohmann, V. (**2006**). "The master hearing aid: A PC-based platform for algorithm development and evaluation," Acta Acust. Acust., **92**, 618-628.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann. V., and Kollmeier, B. (**2009**). "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," EURASIP J. Advances Sig. Proc., 298605, doi: 10.1155/2009/298605.

Mirkovic, B., Debener, S., Schmidt, J., Jaeger, M., and Neher, T. (**2019**). "Effects of directional sound processing and listener's motivation on EEG responses to continuous noisy speech: Do normal-hearing and aided hearing-impaired listeners differ?," Hear. Res., **377**, 260-270, doi: 10.1016/j.heares.2019.04.005.

Mullen, T. R., Kothe, C. A., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., Jung, T.P., and Cauwenberghs, G. (**2015**). "Real-time neuroimaging and cognitive monitoring using wearable dry EEG," IEEE Trans. Biomed. Eng. **62**, 2553-2567, doi: 10.1109/TBME.2015.2481482.

Neher, T., Wagener, K. C., and Latzel, M. (**2017**). "Speech reception with different bilateral directional processing schemes: Influence of binaural hearing, audiometric asymmetry, and acoustic scenario," Hear. Res., **353**, 36-48. doi: 10.1016/j.heares.2017.07.014.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., and Lalor, E.C., (**2014**). "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," Cerebral Cortex, **25**, 1697-1706, doi: 10.1093/cercor/bht355.

Petersen, E. B., Wöstmann, M., Obleser, J., and Lunner, T. (**2016**). "Neural tracking of attended versus ignored speech is differentially affected by hearing loss," J. Neurophysiol., **117**, 18-27, doi: 10.1152/jn.00527.2016.

Van Eyndhoven, S., Francart, T., and Bertrand, A. (**2017**). "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," IEEE Trans. Biomed. Eng., **64**, 1045-1056, doi: 10.1109/TBME.2016.2587382.

Wagener, K. C., Brand, T., and Kollmeier, B. (**1999**). "Development and evaluation of a sentence test for the German language. I-III: Design, optimization and evaluation of the Oldenburg sentence test," Z. Audiol. (Audiol. Acoustics), **38**, 4-15, 44-56, 86-95.

Adaptation to hearing-aid microphone modes in a dynamic localisation task

WILLIAM M. WHITMER[1,*] NADJA SCHINKEL-BIELEFELD[2], DAVID MCSHEFFERTY[1], CECIL WILSON[2], AND GRAHAM NAYLOR[1]

[1] *Hearing Sciences – Scottish Section, Division of Clinical Neuroscience, University of Nottingham, Glasgow, UK*

[2] *Sivantos GmbH, Erlangen, DE*

New technology can foster new ways of listening. A new hearing-aid programme can alter how we hear not only sources of sound but also their locations. While previous research has established how different hearing aid types and microphone modes affect static localisation ability, the current study explored the effects of introducing unfamiliar devices and microphone modes on dynamic localisation ability. Twelve experienced users of bilateral behind-the-ear (BTE) hearing aids oriented themselves to a target sound. Each trial consisted of 5-s segments of a target talker in a continuous background of far-field babble at the same overall level as the target. Targets were presented at either ±30, ±75 or ±120°. Head-orientation trajectories were measured with infra-red cameras. Participants first wore their own hearing aids for one block of 60 trials, then wore a new hearing aid and completed five more blocks in three different directional-microphone modes. In general, results showed trajectory differences between modes, and a modest influence of the preceding mode (i.e., adaptation). Three additional participants experienced with in-the-ear hearing aids oriented poorly with the new BTE device for the first two blocks, then returned to their baseline performance. This suggests that such a form-factor change requires additional time for spatial adaptation.

INTRODUCTION

For the sake of comfortable audibility, hearing aids can alter the spatial information of an acoustic environment. In previous studies of aided localisation, however, the ability to locate static sounds along the azimuth has been only modestly affected by wearing hearing aids in their basic setting, an average increase in error of 1° (Akeroyd and Whitmer, 2016). For directional microphones in hearing aids, which attenuate off-axis sounds to varying extent, the ability to locate a desired sound is only a precursor to the primary task of re-orienting to it. Brimijoin *et al.* (2014) demonstrated in a small group of bilateral hearing-aid users that while orienting accuracy was not affected by conventional cardioid directionality, the duration of and delay to start orientation to the talker was affected by directional microphones. While the effect on static localisation has been shown to be generally negligible, traditional hearing-aid directionality altered the dynamic behaviours that depend on the maintenance of cues

*Corresponding author: bill.whitmer@nottingham.ac.uk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

over time.

Directional technology has seen a recent step change with the proliferation of bilateral beamforming (BiBF), linking the microphones across the ears to create a highly directive "lobar" pattern. In its basic form, BiBF eliminates interaural cues, but continuing advances have improved localisation with BiBF (Neher *et al.*, 2017), though errors are still greater than with traditional unlinked directionality (Picou *et al.*, 2019). With its highly directive pattern in the on-axis "look" direction, the ability to orient to a new desired source is vital to BiBF benefit in the real world. Further, the ways in which BiBFs affect spatial perception may induce orientation behaviours that are not globally beneficial. As part of an adaptive directional scheme, it is important to know how the varying levels of directionality within a scheme affect the behaviours in another scheme. The current study explores these questions by looking at orientation behaviour and its adaptation across BiBF and other directional modes in a realistic conversation-monitoring task.

METHODS

Participants

Twelve adults (7 female) with a median age of 69 years (range 52-72 years) participated. Better-ear four-frequency pure-tone threshold averages ranged from 11-54 dB HL with across-ear asymmetries of 0-14 dB HL (see Figure 1a). All participants were experienced (> 2 years) bilateral behind-the-ear (BTE) hearing-aid users. Three additional participants (2 female; age range 56-70 years) also took part who were experienced (> 3 years) users of completely-in-the-canal (CIC) custom hearing-aid users. Their data is treated separately.

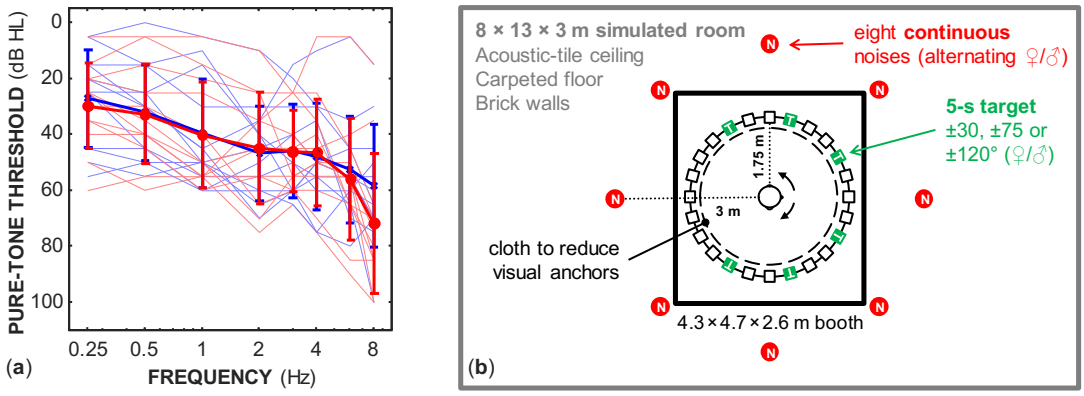


Fig. 1: Panel (a) shows left (blue) and right ear (red) pure-tone thresholds as a function of frequency, both individually (shaded) and means (solid). Error bars show ±1 standard deviation. Panel (b) shows the simulated noise/room configuration, stimuli and actual test apparatus used.

Apparatus

Participants were seated in a freely rotatable chair in the centre of a circular 1.75-m

radius 24-loudspeaker array in a sound-attenuated 4.3 × 4.7 × 2.6-m chamber (see Figure 1b). Head yaw was tracked at the plane of the ears and nose at a sampling rate of 100 Hz using infra-red cameras and reflective crown. The visual location of the loudspeakers was obscured by a black cloth; all stimuli were equalised to offset the frequency dependent 0-2 dB attenuation of the cloth.

Stimuli

Continuous babble noise consisted of eight talkers (alternating male/female) placed in a simulated 8 × 13 × 3-m reverberant room at a distance of 3 m and 45° spacing from the centre (see Figure 1b). The 24-channel impulse response for each noise source was generated using ODEON with nearest-loudspeaker rendering. Noise level (all talkers simultaneous) was calibrated to be 66 dB A at the centre of the participant's head.

Target signals were consecutive five-second segments (25-ms onset and 100-ms linear offset gating) from a Sherlock Holmes story (44.1 kHz sampling rate) spoken by either a man or woman, gender randomised across trials (Macpherson and Akeroyd, 2013). Target signals were presented from the loudspeaker nearest to ±30, ±75 or ±120° from the participant's midsagittal plane on each trial. This created a punctate signal for participants to locate, and due to participants' error distributions, a normal distribution of sources up to ±7.5° from each target location to avoid learning of excursions.

Hearing aids

For the first block of sixty trials, participants wore their own hearing aids. All 12 BTE and 3 CIC wearers were wearing digital hearing aids that were tested in their basic omnidirectional programme. Participants then switched to two Signia 7Nx M hearing aids with receivers in the ear coupled with double domes that were fit to each participant's audiogram using Connexx fitting software. The real-ear insertion gains for their own and the newly fitted devices were measured. The new devices were fit with three customised, non-commercial programmes that were set and monitored during the experiment by the tester: (i) a pseudo-omnidirectional mode (OMNI) that mimics generic pinna directivity; (ii) a fixed unlinked hypercardioid directional mode (DIR); and (iii) a bilaterally linked directional beamformer (BiBF).

Procedure

Participants first performed standard audiometry to establish pure-tone air and bone-conduction thresholds. They were then instructed to imagine an ongoing story being told in a lively room by a series of conversation partners. When they heard a new talker, they were to turn as quickly and comfortably to the new talker, and remain oriented directly towards them until the next talker. Participants then completed 12 practice trials (with their own hearing aids). All participants completed six blocks of sixty trials.

Each block consisted of each target angle repeated ten times in randomised order that was fixated after the first block. That is, to compare trials across blocks, the target angle order for each individual was repeated across all blocks. The first block was

always with subjects' own hearing aids to allow familiarisation and procedural learning. For the subsequent blocks, participants were randomly allocated into one of two hearing-aid programme orders shown in Table 1. The between-group block design was chosen to examine how exposure to one programme affected behaviour in another while adhering to a naturally occurring sequence in a commercial device.

	Block					
	change of HA					
Group	**1**	**2**	**3**	**4**	**5**	**6**
1	Own Has	DIR	BiBF	DIR	OMNI	DIR
2	Own Has	DIR	OMNI	DIR	BiBF	DIR

Table 1: Block order design.

Each block started with 10 s of babble to stabilise the hearing-aid programmes. A 600-ms pause was inserted between target presentations in every block, but the babble was continuous throughout the block.

RESULTS

Analysis

Head yaw trajectories were recorded for each trial; trajectories for trials where the participant did not move were discarded. From each trajectory, eight measures were calculated (cf. Brimijoin *et al.*, 2014). *Error* is the absolute difference between the end angle and the actual target angle. The *start time* of each orientation was defined as the earliest time point at which the angle exceeded ±5° from that trial's starting position. The *end time* of each orientation was when the angle last exceeded ±5° from the end position. The *duration* is the difference between start and end times. *Orientation velocity* was the derivative of angular position. Kinematic studies have shown that the *peak velocity* and *peak velocity time* are indicators of motor-control decisions (Maurer *et al.*, 2017). *Complexity* was calculated as the minimum polynomial fit to the velocity-time function minus two, as any simple movement should have a ballistic velocity fit with a 2nd order polynomial. *Misorientations* were counted when the sign of the initial movement greated than 5° was not equal to the sign of the target (e.g., initially moved to the left for a right hemifield target). *Reversals* were how many changes in direction occurred over each orientation (after 10-sample smoothing).

General results

Mean results (across all 12 BTE participants) for each orientation measure as a function of target angle, device and microphone mode are shown in Figure 2. Results for the DIR mode were averaged across all three DIR blocks. Repeated-measures analyses of variance revealed clear main effects of angle: increased duration, peak velocity and complexity with larger angles, and increased number of reversals with decreasing angle due to overshoot (all $p < 0.001$). There was, however, no main effect

of microphone mode, nor significant interactions across measures.

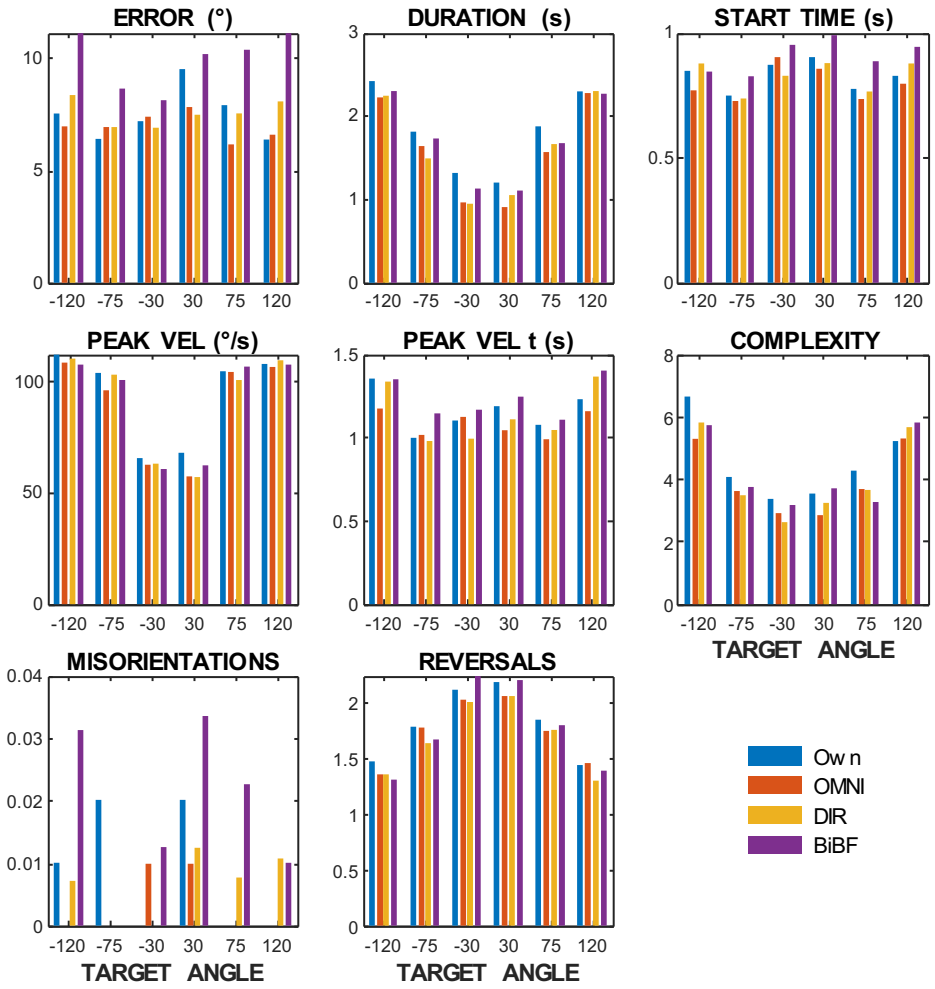


Fig. 2: Mean orientation measures (see text) as a function of angle for own device (blue), omnidirectional mode (OMNI; red), directional (DIR; yellow) and bilateral beamformer (BiBF; purple).

Adaptation results

The block design beginning with participants' own devices (Table 1) allows an examination of adaptation in orientation behaviour during and after exposure to different directional processing. For a simple analysis, measure means were calculated for each block as a function of angle for each participant in each of the block-order groups; the group mean results are shown in Figure 3. In addition, the same means are shown for the three CIC users. What is visually apparent is the dramatic increase in error, duration, start time, misorientations and peak velocity time (coupled with a decrease in peak velocity) for the CIC users, especially for orientations to ±120° targets for the first two blocks after the switch from their CIC devices to the new BTE devices (i.e., block numbers 2-3). As performance in the remaining blocks returns to

that of the BTE groups, the CIC data shows clear signs of feedback-based learning: cautious, contemplative movement, accepting high errors as an adaptation process. Relative to these differences, changes across blocks in any measure between BTE groups 1 and 2 are much smaller.

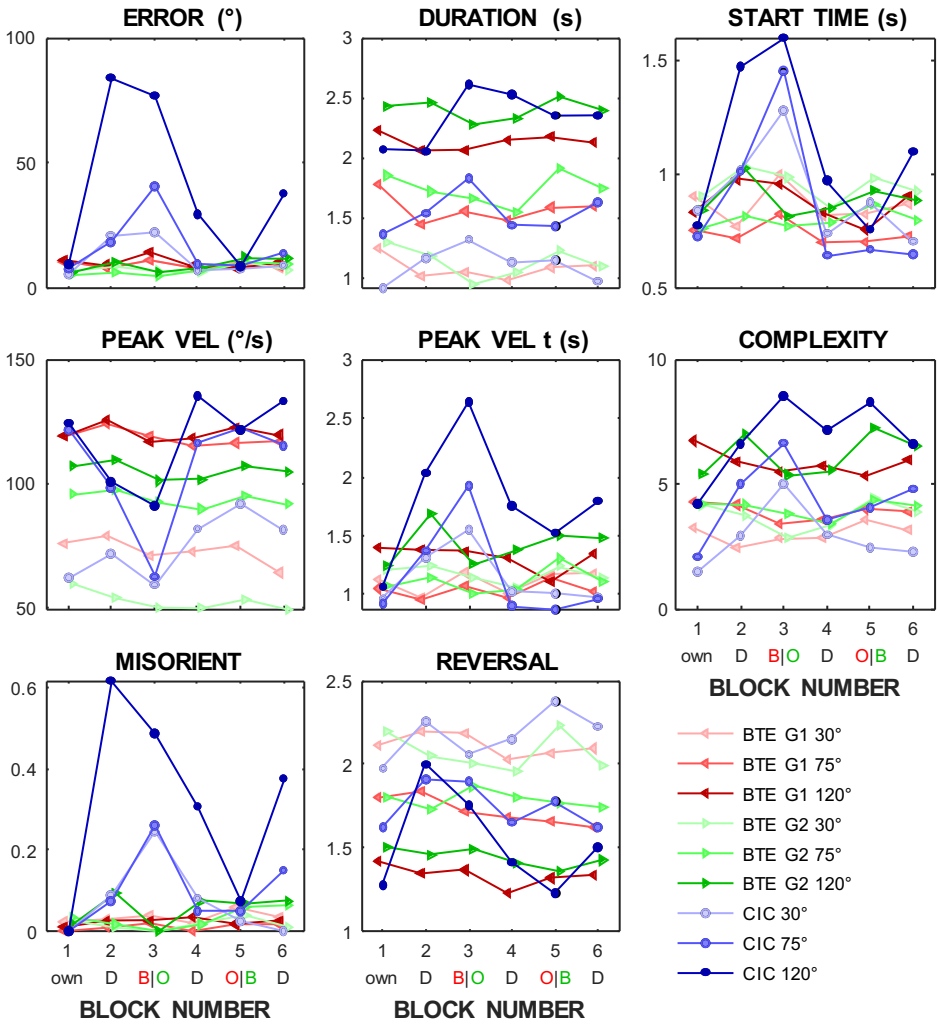


Fig. 3: Mean orientation measures as a function of experimental block (own device; D = directional; B = BiBF; O = pseudo-omnidirectional) for each group (colour/symbol) and target angle (shading; see legend).

To look for adaptation in the movement itself, the trajectories as opposed to measures derived from them were compared across blocks. For each repeat of each angle, the trajectory in the directional blocks (2, 4 and 6) were compared with the preceding block (1, 3 and 5). To derive a singular measure to analyse across conditions, the structural similarity index (SSI) was used, as it considers each sample in relation to others as opposed to independent error estimation (Wang *et al.*, 2004). Based on average start times and durations, the SSI was calculated for the 1-3 s segment of each

±120° trajectory (see Figure 4a). The mean results are shown in Figure 4b. The SSI for trajectories in the OMNI and subsequent DIR block was significantly greater than the DIR and either participants' own device or BiBF ($p < 0.05$); that is, participants' behaviours on average were more similar between DIR and OMNI modes than other modes. However, this adaptation evidence only applies to the SSI, not other similarity analyses (e.g., coherence).

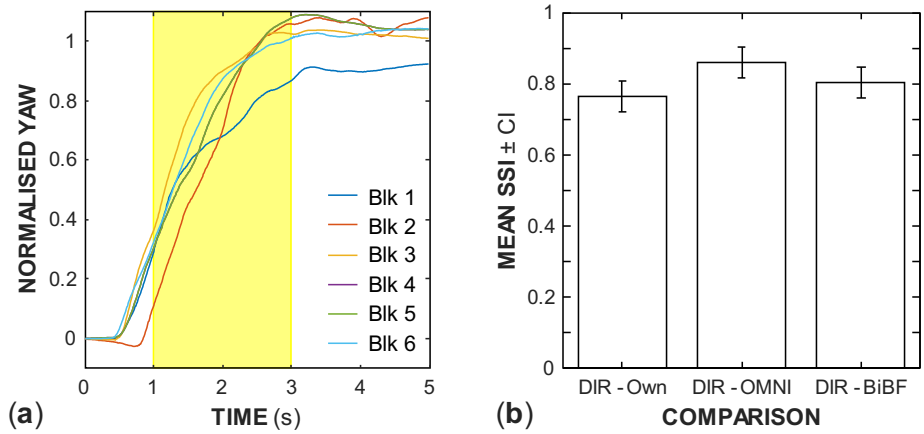


Fig. 4: Panel (a) shows sample trajectories – yaw (normalised to actual target angle) as a function of time – for a given trial in each block, highlighting the section analysed. Panel (b) shows mean SSI as a function of the similarity comparsion between the directional (DIR) blocks and the preceding block [own device, omnidirectional (OMNI) or bilateral beamformer (BiBF)]. Error bars indicate 95% within-subject confidence intervals.

DISCUSSION

Despite known changes to spatial cues, the bilateral beamformer here did not produce substantial changes in orientation behaviour on average. Nor did the fixed directional condition produce changes in behaviour from the omnidirectional condition, which contrasts with the previous orientation differences in Brimijoin *et al.* (2014). The tasks were slightly different in that here the task was timed, whereas the end of each trial was participant controlled in Brimijoin *et al.*, which may have elicited further searching or centring behaviour. The current timed task may have induced a particular global strategy to orient to the source that superseded any particular strategy for a given microphone mode. As the three microphone modes were all directional, it is possible that for an orienting task, they induce the same behaviour, though they produce different static localisation results (cf. Picou and Ricketts, 2019). Another difference is that the previous study used phantom sources between loudspeakers as opposed to single-loudspeaker sources in the current study; this difference could have caused more uncertainty in the precise location of the source, though only minimally. A direct comparison of the participant-controlled and timed methods would be necessary to determine whether it was the differences in method or similarities in microphone that limited the effects in the current study.

Adaptation was not an issue with microphone modes, but it was when the task involved a change in form factor. For the three experienced CIC users, the change to the new BTE device resulted in substantial issues in orienting over the first two blocks with the new devices, then returned to original performance for the remaining three blocks. As each block was 5'36" with a short break, this effect lasted approx. 12 minutes. The effects of this CIC-BTE adaptation (e.g., initially turning in the wrong direction on more than 50% of the trials) were most evident for the further off-axis sources. While there were differences in gain between the two devices, these gain differences were within the same range as the differences in the 12 experienced BTE users, who showed negligible changes when switching devices. Hence the cause of this major disruption in dynamic localisation ability was most likely due to the change in microphone positions. These results highlight the need for accommodating new patients who are changing to a different form factor and tempering their immediate spatial expectations.

ACKNOWLEDGEMENT

This study was funded in part by Sivantos GmbH. This work was also supported by the Medical Research Council [grant number MR/S003576/1]; and the Chief Scientist Office of the Scottish Government. Thanks to Patrick Howell for hearing-aid fitting and verification expertise and Drs. Maja Serman and Ronny Hannemann for guidance.

REFERENCES

Akeroyd, M.A., and Whitmer, W.M. (**2016**). "Spatial hearing and hearing aids," in Hearing Aids. Edited by G.R. Polpeka, B.C.J. Moore, R.R. Fay & A.N. Popper. doi: 10.1007/978-3-319-33036-5_7

Brimijoin, W.O., Whitmer W.M., McShefferty, D., and Akeroyd, M.A. (**2014**). "The effect of hearing aid microphone mode on performance in an auditory orienting task." Ear. Hear., **35(5)**, e204-e212. doi: 10.1097/AUD.0000000000000053

MacPherson, A., and Akeroyd, M.A. (**2013**). "The Glasgow Monitoring of Uninterrupted Speech Task (GMUST): A naturalistic measure of speech intelligibility in noise." Proc. Meet. Acous., **19**, 050068. doi: 10.1121/1.4805817

Maurer, L.K., Maurer, H., and Müller, H. (**2017**). Analysis of timing variability in human movements by aligning parameter curves in time. Behav. Res. Methods., **50(5)**, 1841-1852. doi: 10.3758/s13428-017-0952-0

Neher, T., Wagener, K.C., and Latzel, M. (**2017**). "Speech reception with different bilateral directional processing schemes: Influence of binaural hearing, audiometric asymmetry, and acoustic scenario." Hear. Res., **353**, 36-48. doi: 10.1016/j.heares.2017.07.014

Picou, E., and Ricketts, T. (**2019**). "An Evaluation of Hearing Aid Beamforming Microphone Arrays in a Noisy Laboratory Setting." J. Am. Acad. Audiol., **30(2)**, 131-144. doi: 10.3766/jaaa.17090

Wang, Z., Bovik, A.C., Sheikh, H.R., and Simoncelli, E.P. (**2004**). "Image quality assessment: From error visibility to structural similarity." IEEE Trans. Image Proc., **13(4)**, 600-612. doi: 10.1109/TIP.2003.81986

The vent effect in instant ear tips and its impact on the fitting of modern hearing aids

SUELI CAPORALI*, JENS CUBICK, JASMINA CATIC, ANNE DAMSGAARD AND ERIK SCHMIDT

Sound and Fitting, WS Audiology A/S, DK-3540 Lynge, Denmark

Today, approximately 70-80% of hearing aid fittings are made with instant ear tips. This may be due to ease of fit, improved physical comfort and the reduction in occlusion compared to custom earmolds. These tips can be completely open, vented, or closed. The acoustic properties of the ear tip depend on its type, size and the fit to the individual ear canal. Depending on the resulting real ear occluded gain (REOG) and vent effect (VE), the sound quality and aided benefit provided by the hearing aid may vary among individuals fitted with the same ear tip type. This study explored five Widex instant ear tips both in relation to REOG and VE using real ear measurements on 60 ears and in relation to subjective occlusion ratings. The results showed a large variation in REOG and VE both between ear tips and across subjects within the same ear tip type, and a high correlation between VE and perceived occlusion. These results imply that the acoustics of instant tips need to be assessed and considered as part of the hearing aid fitting process to ensure that fitting targets are matched.

INTRODUCTION

Open-fit and receiver-in-the-canal (RIC) hearing aids (HAs) have emerged in the last decade and become very popular due to their attractive and comfortable-to-wear design (Hallenbeck and Groth, 2008). The strong preference for RIC devices might also be a result of the HA performance, especially with the improvement of feedback-cancellation algorithms (Martin, 2008) and extended frequency bandwidth (Kuk and Baekgaard, 2008) in comparison to behind-the-ear HAs. In the United States RIC-style HAs constituted 82.6% of the private sector market and 77.9% of veteran affairs (VA) dispensing, for a total of 81.7% of all HAs dispensed in the first half of 2019 (Hearing Review , 2019). With the increased number of RIC devices, the use of instant ear tips has grown proportionally, whereas the number of custom moulds has decreased. Instant ear tips have been reported to be used in about 70 % of the fittings (Smith, *et al.*, 2008; Sullivan, 2018). The preference for instant ear tips might be mostly due to two reasons: time efficiency and increased comfort. With instant HA fitting during the first visit to the hearing clinic and no need to take earmould impressions, more time can be spent on counselling, fine-tuning, and verification (Caporali *et al.*, 2013). Instant ear tips also generally provide more comfort and reduce the occlusion effect (Pohlman and Kranz, 1926; Dillon, 2012; Stenfelt *et al.*, 2003). They are available from all manufacturers in a variety of shapes and sizes, ranging from so-called open ear tips to closed tips that aim to completely occlude the ear canal.

*Corresponding author: suca@widex.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Ear tips may be designed to include vents, or they might allow for leakage between the tip and the ear canal wall, which leads to an increased but uncontrolled effective vent size.

Every ear tip has an acoustic effect on the sound pressure at the eardrum, both for sounds amplified by the HA, for sounds produced outside the ear canal, and for the mixing of the two sound sources at the eardrum. For sound amplified by the HA, increasing the effective vent size decreases the amount of low-frequency energy at the eardrum, which we refer to here as the vent effect (VE, e.g., Dillon, 2012; Kuk and Keenan, 2006; Kuk and Nordahn, 2006). For sounds generated outside the ear canal, increasing the effective vent of the ear tip increases the level of the direct sound at the eardrum, an effect described by the real ear occluded gain (REOG). If the transmitted sound and the sound played by the HA are similar in level, the superposition of the direct and the slightly delayed amplified sound results in a comb filter effect (Dillon, 2012; Mueller and Ricketts, 2006). If the direct sound is higher in level than the amplified sound at certain frequencies, this can render HA processing algorithms like beamformer or noise reduction less effective, because the direct sound dominates the perception (Keidser *et al.*, 2007).

For traditional vented custom moulds, these effects are well-understood and well-documented (Kuk and Nordahn, 2006). There has been some research on the acoustics of open fittings, reviewed in Winkler *et al.* (2016). For other instant ear tips, the literature is very sparse. Jespersen and Møller (2013) found that the reliability of the acoustic properties of instant tips and custom moulds was comparable, and Smith *et al.*, (2008) found that instant tips offered a viable assess-and-fit option in clinical practice. This lack of knowledge is especially problematic given the popularity of instant tips, because the understanding of ear tip acoustics is essential for making the right fitting choices for a given hearing loss, especially for patients with a mild hearing loss at low frequencies, who require a compensation of the VE in order to match the fitting target. Similarly, knowledge of the REOG of a specific ear tip is important, because it allows for adjustments of the HA gain to minimise comb filter effects.

This study aimed to provide acoustic reference data for five types of silicone instant ear tips ranging from open to closed (no intentional venting) and to look at the inter-subject variability across ear tips. Three research questions have been addressed: 1) What is the REOG for the five different ear tips across subjects? 2) What is the VE for the different tips and how much does it vary across subjects? and 3) How much occlusion do the test subjects experience with these ear tips and how does that relate to measured VE?

METHODS

Test subjects

Thirty normal-hearing subjects (10 female) participated in the experiment. The average age was 45 years (min. 19, max 67). All participants were employees of WS Audiology. The subjects were screened before the experiment to determine the ear tip

size and wire length required to fit their ears. The selection criterion was to have a broad range of different sizes of ear canal and body heights and to include both genders. Before starting the experiment, they received written information about the experiment and gave informed consent to their participation and to the use of the resulting anonymised data for publication.

Procedure and apparatus

The main part of the experiment comprised a series of real-ear measurements using an Interacoustics Affinity 2.0 measurement system. The listeners were placed facing the system at a distance of 1 m in an acoustically dampened audiometric room. A pair of Widex Evoke 440 Passion receiver-in-the-canal (RIC) HAs with 'S' receivers was used for all measurements. They provided 10 dB linear gain in all 15 frequency channels. All adaptive processing was deactivated. The ear tips under investigation were Widex Open, Tulip, Round (2-vent), Round (1-vent), and Double-domes instant ear tips (cf. Fig. 1). The order of testing the different ear tips was counterbalanced across test subjects to avoid order effects.

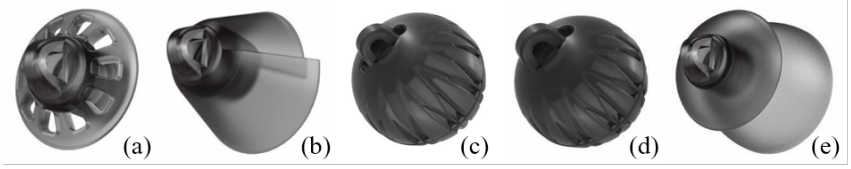


Fig. 1: Instant ear tips used in this study: Open (a), Tulip (b), Round (2-vent), c), Round (1-vent), d), Double domes (e)

After an initial otoscopic examination of the ear canal and the calibration of the probe microphones, the probe tubes were placed in the ear canals close to the eardrum and real-ear unaided responses (REUR) were measured using pink noise played back from the Affinity system at 0° azimuth at a distance of 1 m at a level of 65 dB SPL. Then the HAs were placed on the ears and the tips were inserted into the ear canal. Subsequently, real-ear occluded responses (REOR) were measured with the HAs switched off. The REOG was computed as the difference between the two measurements in dB (REOG = REOR – REUR).

The VE for each ear tip was obtained from a second pair of measurements. Brown noise was streamed to the HAs via a TV-DEX streaming device. Subsequently, the ear canal and the concha were filled completely with impression material with the receiver and ear tip still in place. This allowed for the measurement of a fully occluded response with no leakage. The VE was computed as the difference between the response measured in the 'normal' streaming measurement and the measurement with impression material.

Finally, the test subjects were asked to utter the names of the months of the year and rate the perceived occlusion based on the sound of their own voice on a scale ranging from 0 (no occlusion, "like normal listening") to 10 (complete occlusion, "as if they stuck their fingers in their ears").

RESULTS

Real ear occluded gain

The results for one test subject were excluded from the analysis due to issues with cerumen occluding the probe tube during some measurements. Fig. 2 shows the average REOG across the remaining 58 ears for the different ear tips (top left panel). The other panels show the same average REOG (thick line), +/- one standard deviation (shaded area), and the individual measurements (light grey lines) for each of the ear tips separately. The average results (top left) show that the Open ear tips are mostly transparent for sound generated outside the ear canal, apart from a slight attenuation of about 2 dB at mid- and high frequencies above about 2 kHz. The remaining tips form two groups with different gain responses. The Tulip, Round 1 and 2 tips show very similar attenuation patterns with a transparent response up to about 1 kHz and a maximum attenuation of about 9 (Tulip) to 12 dB (Round, 1-vent) at a frequency of about 2.6-2.8 kHz. Double domes are on average only transparent up to about 600 Hz, and they show the highest attenuation of 16 dB at 3 kHz. The data also show high variability between subjects across ear tips, which is largest for double domes.

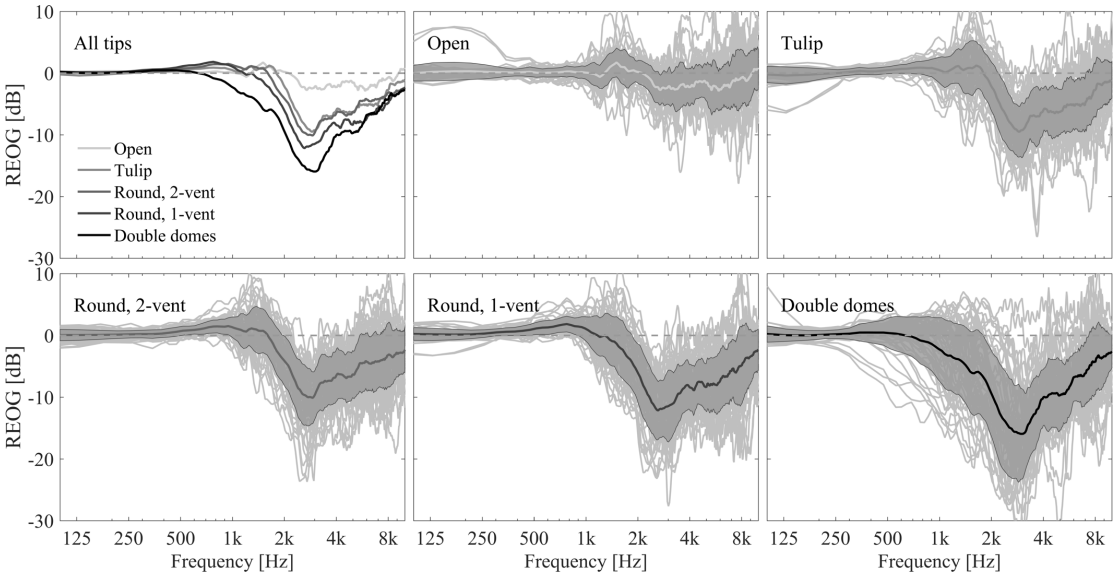


Fig. 2: Average REOG across 58 ears for the five ear tips (top left). The other panels show the average REOG per tip (thick line) and +/- 1 standard deviation (shaded area). The light grey lines represent individual measurements.

Vent effect

Fig. 3 shows the average VE across 58 ears for each of the tips (top left) and the average value (thick line) +/- one standard deviation (shaded area) and the individual measurements (light grey lines) for each tip separately. On average, the largest VE was found for open tips. Here, the vent loss starts just below 2 kHz and reaches a

maximum of about 40 dB around 125 Hz. Noteworthy is also the peak in the VE curve between about 2 and 6 kHz. This peak is caused by the difference in the magnitude of the ear canal resonance between the two underlying measurements. While the Open tip hardly affects the natural ear canal resonance at all, its magnitude is clearly reduced when the impression material is inserted. Tulip, Round (2-vent) and Round (1-vent) show very similar responses with a vent loss starting between 1 and 1.5 kHz and a maximum attenuation of about 30 dB at the lowest frequencies. Double domes show the least pronounced vent loss with a cut-off frequency of about 1150 Hz and an average VE of 24 dB around 125 Hz. There is also a high variability in the VE measurements across ear tips. Open shows the lowest and Double domes show the highest variability, which also was seen for the REOG. At the lowest frequencies, the VE with Double domes might be as low as 6 dB (nearly fully closed) or as high as 38 dB (nearly completely open).

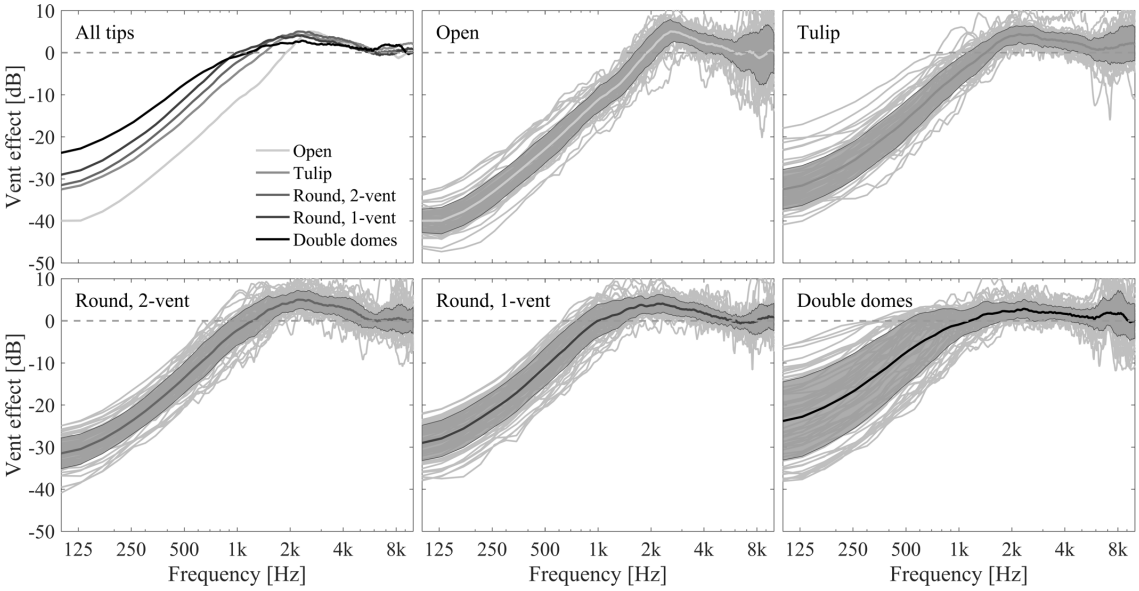


Fig. 3: VE, averaged across 58 ears for each ear tip (top left). The other panels show the average VE per tip (thick line) and +/- 1 standard deviation (shaded area). The light grey lines represent individual measurements.

Occlusion ratings

Fig. 4 shows the average occlusion ratings for the different ear tips (open circles) +/- one standard deviation and for the condition with impression material (filled circles). The smaller grey symbols represent individual responses. Circles indicate that the occlusion was perceived equally strong on both ears. Triangles point in the direction of the ear with higher occlusion. On average, open tips were perceived as least occluding with an average rating of 0.82, followed by Tulip (2.31), Round 2 (2.87), Round 1 (3.23), and Double domes (3.74). The individual ratings showed a trend of increased variability with increasing average occlusion rating. Whereas all ratings for Open tips were rated between 0 and 4, the ratings for Double spread from 0 to 8.

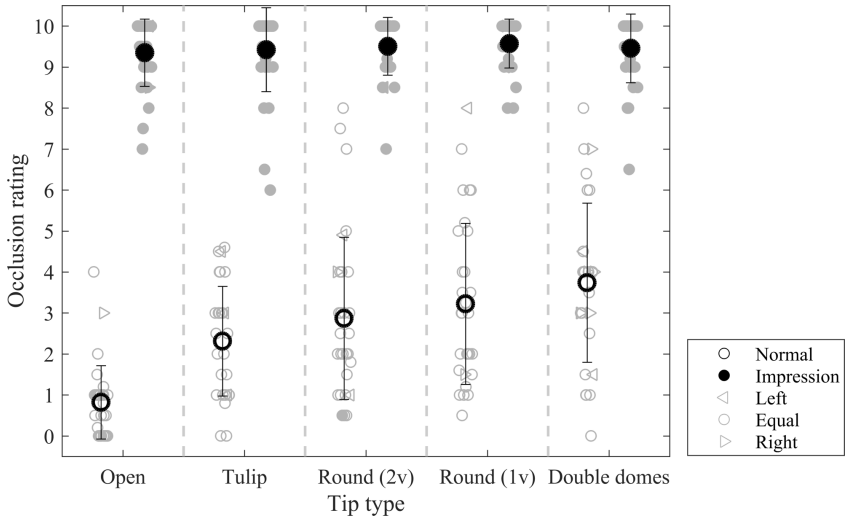


Fig. 4: Individual occlusion ratings (grey) and mean +/- one standard deviation (black). Circles indicate equal amounts of occlusion on both ears, triangles point in the direction of stronger occlusion.

All average ratings for the conditions with impression material were above 9.3, indicating that the impression material was an efficient means for completely occluding the ear canal. A correlation analysis of the occlusion ratings and the average VE between 200 and 400 Hz, which is the most critical frequency range for occlusion (Dillon, 2012), showed a highly significant negative correlation (Spearman's rho: -0.61, $p < 0.0001$).

DISCUSSION

The measured REOG indicates that all tested instant tips were, on average, acoustically transparent up to about 1 kHz except Double domes, which were transparent only up to about 600 Hz. Similar findings are described in Mueller *et al.* (2017). However, the data show substantial between-subject variability. A transparent tip may have a positive influence on the perceived sound quality for users with residual hearing at low frequencies, who will likely be able to fully utilize low-frequency acoustic cues. However, care must be taken for users who need amplification at low frequencies, because the interaction of the direct sound transmitted through the ear tip and the slightly delayed amplified sound from the HA can cause a comb filter effect. This effect is most pronounced when both sounds are similar in level (Dillon, 2012). If the REOG is known and considered in combination with the individual gain requirement, such comb filter effects can be reduced considerably.

All tested instant tips showed acoustic leakage (energy loss at low frequencies), even if no venting is intended. The average VEs can be grouped into three distinct acoustic identities for the five ear tips: open (Open), semi-open (Tulip, Round 1, Round 2), and semi-closed (Double). However, inter-subject variability was large, probably due to differences in coupling between tip and individual ear canal. These findings suggest

that instant ear tips are not the best solution for patients who need significant low-frequency gain, because a lot of the signal generated by the HAs will 'escape' through the vent and not reach the eardrum. If a lot of gain is required at low frequencies, a custom earmold is still a better choice.

Especially the results for Double domes show a high inter-subject variability with VEs ranging from completely open for some subjects to completely closed for others. This suggests that compensating for the average vent loss might compromise sound quality for individual users, who would either experience loss of bass if their individual VE is larger than the average value or boominess if their VE is smaller. Hence, an estimation of the VE on an individual level is needed to apply a suitable amount of VE compensation. The large variability found in the study also highlights the importance of choosing the correct tip size for the individual ear canal to optimise the acoustic coupling between receiver and ear canal and thereby minimising the VE.

The test subjects' rating of occlusion correlates well with the measured VE, which is consistent with findings in Kiessling *et al.* (2005), showing the importance of choosing the right tip type for the individual client, based on their hearing loss profile. In the end, the hearing care professional will always need to find a compromise between comfort and acoustical demands. They have to select a tip that is open enough to avoid occlusion, while it is closed enough to allow for sufficient gain to match the fitting target.

CONCLUSIONS

On average, instant ear tips seal the ear canal less effectively than custom earmoulds, resulting both in more transparency for sounds from outside the ear canal and in a more pronounced VE. This is desirable for sloping hearing losses, because it reduces the occlusion effect. However, such ear tips might not be the best option for flatter losses that require gain also at low frequencies. Furthermore, the large inter-subject differences found in this study indicate the need for an individual in-situ estimate of the VE and REOG, since using inappropriate average values for VE compensation in the HA fitting can negatively impact the sound quality and lead to either a "boomy" or a "tinny" sound from the HA. Measuring the VE as part of the HA fitting procedure will give the best preconditions to maximise the sound quality for each end user individually, since this will allow the fitting software to provide the correct amount of gain for vent compensation and for avoiding comb-filter effects. Even though only Widex ear tips were used in the current study, we expect to find similar results for the instant tips from other manufacturers, since the coupling between each ear tip and the individual ear depends on the tip shape and size and the individual ear canal configuration. Ultimately, the results show that real-ear verification is crucial in clinical practice to verify that the prescribed gain has been matched.

ACKNOWLEDGEMENTS

The authors would like to thank Steen Rose, who helped with the data acquisition, Dina Lelic for her assistance with the statistical analysis, and the test subjects for their patience and willingness to participate.

REFERENCES

Caporali, S. A., Schmidt, E., Eriksson, Å., Sköld, B., Popecki, B., Larsson, J., and Auriemmo, J. (**2013**). "Evaluating the physical fit of receiver-in-the-ear hearing aids in infants", J. Am. Acad. Audiol., **24**(3), 174–191.

Dillon, H. (**2012**). *Hearing aids*, Thieme, New York.

Hallenbeck, S. A., and Groth, J. (**2008**). "Thin-tube and receiver-in-canal devices: There is positive feedback on both!", Hear. J., **61**(1), 28–30.

Hearing Review (**2019**). "Hearing Aid Sales Increase by 3.8% in First Half of 2019", Retrieved September 12, 2019, from http://www.hearingreview.com/2019/07/hearing-aid-sales-increase-3-8-first-half-2019/

Jespersen, C. T., and Møller, K. N. (**2013**). "Reliability of real ear insertion gain in behind-the-ear hearing aids with different coupling systems to the ear canal", Int. J. Audiol., **52**(3), 169–176.

Keidser, G., Carter, L., Chalupper, J., and Dillon, H. (**2007**). "Effect of low-frequency gain and venting effects on the benefit derived from directionality and noise reduction in hearing aids", Int. J. Audiol., **46**(10), 554–568.

Kiessling, J., Brenner, B., Jespersen, C. T., Groth, J., and Jensen, O. D. (**2005**). "Occlusion effect of earmolds with different venting systems". J. Am. Acad. Audiol., **16**(4), 237–249.

Kuk, F., and Baekgaard, L. (**2008**). "Hearing Aid Selection and BTEs : Choosing Among Various " Open-ear " and " Receiver-in-canal " Options", Hear. Rev., **15**(3), 22–36.

Kuk, F., and Keenan, D. (**2006**). "How Do Vents Affect Hearing Aid Performance ?", Hear. Rev., **13**(2).

Kuk, F., and Nordahn, M. (**2006**). "Where an Accurate Fitting Begins : Assessment of In-Situ Acoustics (AISA)", Hear. Rev., **13**(7), 34–42.

Martin, R. L. (**2008**). "BBs: The hearing aids of the near future", Hear. J., **61**, 46–48.

Mueller, B. H. G., and Ricketts, T. A. (**2006**). "Open-canal fittings : Ten take-home tips", Hear. J., **59**(11), 24–39.

Mueller, H. G., Ricketts, T. A., and Bentler, R. (**2017**). *Speech mapping and probe microphone measurements*, Plural Publishing.

Pohlman, A. G., and Kranz, F. W. (**1926**). "The influence of partial and complete occlusion of the external auditory canals on air and bone transmitted sound", Ann. Otol., Rhinol. Laryngol., **35**(1), 113–121.

Smith, P., Mack, A., and Davis, A. (**2008**). "A Multicenter Trial of an Assess-and-Fit Hearing Aid Service Using Open Canal Fittings and Comply Ear Tips", Trends Amplif., **12**(2), 121–136.

Stenfelt, S., Wild, T., Hato, N., and Goode, R. L. (**2003**). "Factors contributing to bone conduction: The outer ear", J. Acoust. Soc. Am., **113**(2), 902–913.

Sullivan, R. F. (**2018**). "A Simple and Expedient Method to Facilitate Receiver-in-Canal (RIC) Non-custon Tip Insertion", Hear. Rev., **25**(3), 12–13.

Winkler, A., Latzel, M., and Holube, I. (**2016**). "Open Versus Closed Hearing-Aid Fittings: A Literature Review of Both Fitting Approaches", Trends Hear., **20**, 1–13.

Individual hearing aid benefit: Ecological momentary assessment of hearing abilities

PETRA VON GABLENZ[1,2*], ULRIK KOWALK[1,2], JÖRG BITZER[1,2], MARKUS MEIS[3,2], AND INGA HOLUBE[1,2]

[1] *Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, D-26121 Oldenburg, Germany*

[2] *Cluster of Excellence Hearing4all, Oldenburg, Germany*

[3] *Hörzentrum Oldenburg GmbH, D-26129 Oldenburg, Germany*

Questionnaires are often used to address the subjective perspective on hearing abilities in the course of hearing aid (HA) fitting. Weaknesses of this approach are, e.g., memory bias and possible mismatch of the pre-defined and individually experienced listening situations. Ecological momentary assessment (EMA) including in-situ surveys in real-life, could tackle these issues. We conducted an EMA study to examine how HA uptake changes the perception of everyday hearing abilities. In collaboration with local hearing aid acousticians, 16 first-time and follow-up HA wearers were recruited. They used the smartphone-based EMA system olMEGA for 3-4 full days before HA fitting and after HA acclimatization. This system allows for specifying situations and sound sources as well as for assessing hearing related dimensions like speech understanding and listening effort. Nine hundred thirty-three surveys out of a total of 1705 surveys related to speech listening events. Results showed a considerable individual variability regarding the type of reported events, the distribution and position of assessments. Overall, speech understanding improved by 1.1 scores and listening effort decreased by 1.3 scores on 7-point scales in post-intervention EMA compared to pre-intervention EMA.

INTRODUCTION

Questionnaires are widely used in hearing rehabilitation and research to capture the subjective perspective on hearing abilities. Apart from their advantages, the standardized inventories certainly do have weaknesses as well. Filled in retrospectively, the assessment might be biased by memory effects. Moreover, the pre-defined listening situations described in the questionnaires might not meet the real-life challenges experienced by the individual – neither in frequency nor in importance. Therefore, ecological momentary assessment (EMA), an in-situ survey including prompt and repeated assessments in real-life is increasingly used in audiological research (Holube *et al.*, under review). Galvez *et al.* (2012) were among the first who demonstrated the feasibility of EMA in elderly hearing aid users. They already claimed further studies to determine if EMA "can be used in the clinical setting with patients, both before and after receiving hearing aids." Against this

*Corresponding author: petra.vongablenz@jade-hs.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

background, we used an EMA approach in a field study with clients seeking hearing health care in order to trace the change of self-reported hearing abilities, particularly listening effort and speech understanding, associated with hearing aid uptake.

METHODS

Study design

The interventional field study was carried out from 2018 to 2019 in Oldenburg, Germany. Adults who were medically advised for hearing aid uptake conducted EMA surveys before being fitted with hearing aids ("pre" condition) and after hearing aid acclimatization ("post" condition). Along with the EMA surveys, the study protocol included various other measures not reported here such as comprehensive audiometric tests, questionnaires, interviews, and external assessment of communication behavior. The research design and procedures passed examination by the ethics commission of the Carl von Ossietzky University in Oldenburg. Written informed consent was obtained from all participants. Study participation was remunerated on an hourly basis for visits at the institute and blanket per day for EMA periods.

Participants

In collaboration with local hearing aid acousticians, 24 adults with a medical prescription for hearing aids were recruited. Seven participants cancelled hearing aid (HA) uptake during the fitting process and one participant left the study. In total, 16 adults (8 males, 8 females) aged 48 to 76 yrs (median 67 yrs) completed the study protocol, among them 13 first-time HA users and three follow-up users. HA choice and fitting were left to the participant and the HA acoustician, respectively, since they reflect decisions made in real-life professional care. The participants' hearing losses were mild to moderate and, except in participant no. 9, rather symmetric (Figure 1). HA acclimatization took 3.3 months on average (min. 0.7, max. 5.4).

EMA equipment and sampling

The participants used olMEGA, a smartphone-based EMA device. olMEGA provides an adaptive questionnaire app and allows for privacy-aware storage of acoustical feature data (Kowalk *et al.*, 2018). Every participant was instructed in the handling of the device for about 30 minutes and received an illustrated manual as well as the experimenter's phone number in case they needed support. Every participant used the EMA device for 3 to 4 days both before HA fitting and after HA acclimatization. Subject-initiated entries were possible at any time and a reminder was scheduled every 25 to 35 minutes.

The adaptive questionnaire app provided staggered option menus to specify situations, activities, speech familiarity, sound sources, and target signals. Assessments were requested for various hearing dimensions and related items, such as sound localization, importance of good hearing, listening effort, loudness, pleasantness of sounds, disability, and speech understanding (in that order) using 7-point ordinal

scales. One survey took approximately 1 to 1.5 min. Delimited by the first and the last survey of each day, the daily usage of the EMA device was 11 h on average.

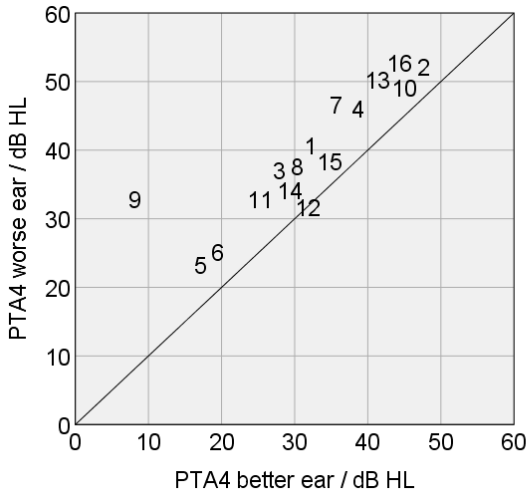


Fig. 1: Study participants' hearing loss at PTA (0.5, 1, 2, and 4 kHz) in the better and the worse ear. The numbers refer to the participants' randomized ID (alphanumeric strings) in ascending order for first-time HA users (no. 1 to 13) and follow-up HA users (no. 14 to 16).

Statistical analysis

To reduce the variety of listening and non-listening events, the participants' responses to the target sound sources were aggregated to five categories: Natural speech in quiet, natural speech in the presence of other sound sources, electro-acoustically presented speech (radio, TV, mobile or landline phone, loudspeaker), non-speech listening targets, and non-listening events. For further analysis, numerical values ranging from 1 to 7 were assigned to assessments given on the ordinal scales and the corresponding variables were treated as metric. Assessments given for speech listening events as dependent variables were regressed on condition as independent variable (pre versus post HA intervention) using a mixed model approach, though the distributional assumptions for linear regression were not fully met. Participants were included as fixed factor and intercepts as well as slopes were defined as random factors. Moreover, correlation coefficients were calculated. Pearson correlational analyses included the individual differences at the mean score in pre- and post-intervention EMA, days of HA acclimatization, and PTA in the better and worse ear. Spearmen's correlation coefficients were calculated to clarify the relationship of assessments for different dimensions of hearing abilities.

RESULTS

In total, 1705 EMA surveys were collected in the pre and post condition (mean 106, min. 48, max. 155 per participant). Of these 1705 surveys, 1109 related to listening

events of any type. In total, 933 surveys related to speech listening events with assessment of speech understanding and listening effort.

Comparability of listening events

Before comparing pre- and post-intervention assessments, we examined whether the proportion of target sound types matched in pre and post EMA on the individual level. Figure 2 shows the high inter-subject variability regarding the events assessed. In many cases, the proportion of surveys in non-listening events is high, most probably due to reminder-initiated responses, whereas listening to speech in background noise or listening to electro-acoustically presented speech mostly account for a small fraction only. The match of target sound types in pre and post EMA was moderate ($r_{Spearman}$= 0.64).

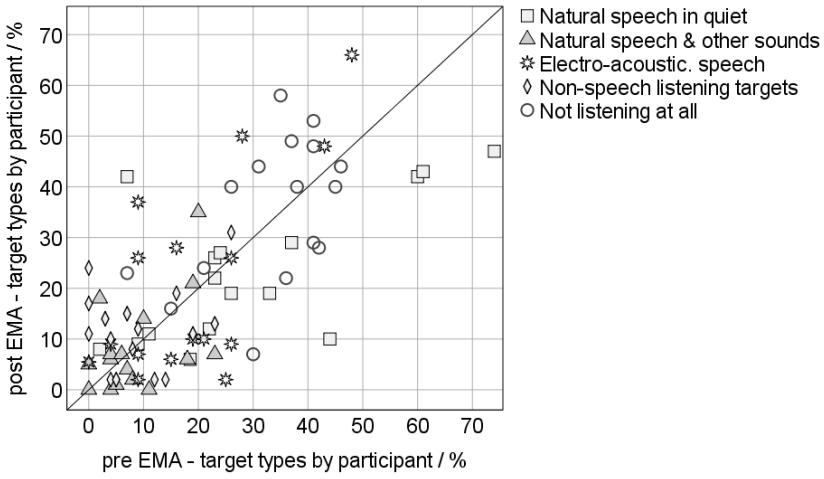


Fig. 2: Percent target types by participant in pre and post EMA.

Listening effort in pre and post EMA surveys

Individual assessments of listening effort in pre and post condition are shown in Figure 3 for three first-time users separately for three types of speech listening events. Note that these examples are randomly chosen out of the 13 first-time HA users.

As can be seen already from these few examples, the type of events, the position and distribution of assessments have different patterns. Participants 1 and 3 almost exclusively reported quiet environments when listening to natural speech, whereas participant 2 predominantly reported other sounds along with natural speech. Some participants used the entire scale, others only a quite narrow scale section for assessments. In this respect, EMA of HA first-time and follow-up users did not indicate any systematic difference. Correlational analyses showed that mean and standard deviation of listening effort assessments do not significantly relate to the degree of better or worse ear hearing loss neither in pre nor in post EMA. Absolute $r_{Pearson}$ estimates ranged from 0.04 to 0.28. In the post-intervention EMA, assessments

of speech understanding and listening effort showed ceiling effects and variance was mostly lower than in pre-intervention EMA.

In general, assessments of listening effort and speech understanding are highly correlated ($r_{Spearman}$=0.78) and both are similarly high correlated to self-reported disability and the ability of sound localization with correlation coefficients ranging between 0.73 and 0.81.

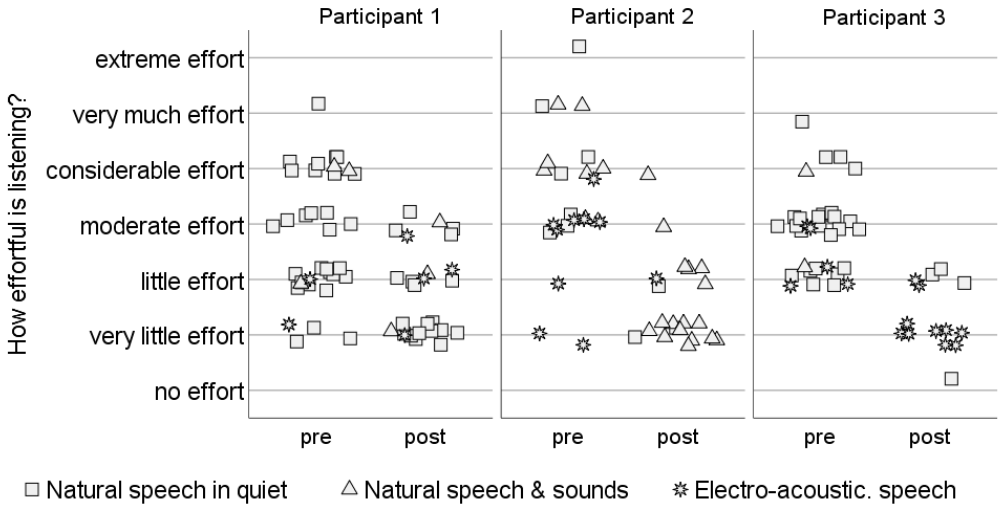


Fig. 3: Individual assessments of listening effort from pre- and post-intervention EMA. Data jittered for display.

Individual Benefit

Most but not all participants assessed real-life listening being less effortful with (new) HA than before the intervention. Figure 4 shows the mean assessment of listening effort in pre and post EMA separately for all study participants and regardless of the type of speech listening event. As expected, individuals differ with regard to HA benefit in terms of absolute change in mean assessment, with follow-up users having overall less benefit than first-time users. HA benefit, however, is uncorrelated to both degree of hearing loss and duration of HA acclimatization. Absolute $r_{Pearson}$ estimates range from 0.16 to 0.35 and failed statistical significance.

A linear mixed regression model was established to estimate the impact of the HA intervention on various hearing dimensions and related items. Table 1 reports intercept and beta coefficients with bootstrapped 95%-confidence intervals for listening effort, speech understanding, and the pleasantness of sounds for both all speech listening events combined and separately for types of speech targets. HA intervention showed a significant effect on these hearing dimensions (p = 0.001, 2-tailed). Beta estimates indicate that listening effort decreased by 1.3 scores and speech understanding improved by 1.1 scores in the post-intervention EMA compared to the pre-intervention EMA. Estimates of speech understanding are almost the same for all

speech target types, whereas the effect on listening effort was largest for natural speech in the presence of other sounds and smallest for electro-acoustically presented speech. Significant effects were stated also for localization of sounds and self-reported disability. Somewhat surprisingly, the pleasantness of sounds was also better assessed in post EMA than in pre EMA. Self-rated loudness and the importance of good hearing did not materially change (beta estimates ≤ 0.09).

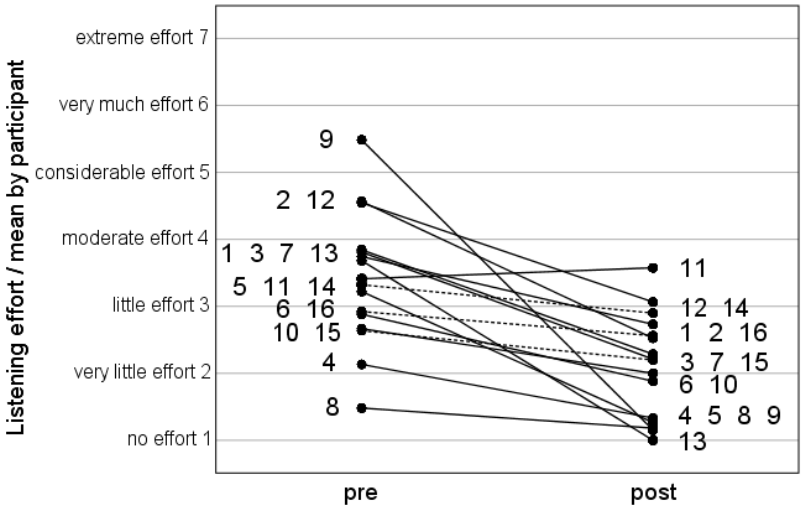


Fig. 4: Assessment of listening effort averaged by participant for all speech listening events in pre- and post-intervention EMA. Results are shown for first-time HA users (solid lines) and follow-up HA users (dashed lines).

Covariance parameters estimated in the mixed models reported in Table 1 confirmed that slopes as well as intercepts differed significantly between the participants. Moreover, the covariance of slopes and intercepts was negative in every one of the models indicating that higher pre-interventional assessments result in smaller intervention benefit.

DISCUSSION

This study used EMA to trace the change of hearing abilities and related dimensions in real-world environments associated with hearing aid uptake. Particular emphasis was put on the assessment of listening effort and speech understanding before hearing aid uptake and after hearing aid acclimatization. At large, self-report for both hearing dimensions was found to be shifted towards improvement after HA acclimatization, presumably indicating HA benefit. This was shown on the individual level, exemplarily also for different types of speech listening events, and with effect estimates based on a mixed model approach. It was further shown that pre-interventional assessments and HA benefit were connected and that both were not significantly related to the degree of hearing loss and the duration of the acclimatization period. Note that three participants were follow-up HA users and 10 out of 13 HA first-time users had mild hearing loss, partly retaining pure-tone hearing

abilities not considered as "impaired" according to the WHO criterion for hearing impairment (better ear PTA > 25 dB HL). For this reason, ceiling effects were to be assumed not only in post-, but also in pre-interventional assessments while larger intervention effects would be expected in participants with a more pronounced and previously unaided hearing loss. In the present study sample, the mixed model estimate established an overall improvement by 1 to 1.6 categories on the 7-point scales for listening effort and speech understanding. Given the heterogeneity of the study sample, these results must be interpreted with caution since the sample size does not support including further covariates in the model to control, e.g., for the degree and type of hearing loss, socio-demographics, first- and follow-up HA use. However, estimates based only the data of twelve HA first-users with symmetric hearing loss (excluding the data of participant no. 9 due to pronounced asymmetric hearing), were very similar to the estimates derived in the total sample (95%-CI widely overlap). Timmer *et al.* (2018) also researched the effect of HA rehabilitation using EMA in 10 elderly subjects with mild hearing impairment in unaided and aided conditions. In contrast to the present study, the subjects were fitted with HA as part of the study and had prior experience with the EMA method. Using different item wording and scales, Timmer *et al.* also found significant improvements in the aided condition compared to the unaided baseline condition, though somewhat stronger for speech understanding (beta = 0.9) than listening effort (beta = 0.7).

Dimension	**Intercept**	**Beta**	**[95% CI]**
How effortful is listening?			
All speech listening events	3.4	-1.3	[-1.4, -1.2]
Natural speech in quiet	3.5	-1.4	[-1.6, -1.1]
Natural speech & sounds	3.8	-1.6	[-1.9, -1.1]
Electro-acoustic. speech	3.1	-1.1	[-1.3, -0.9]
How good or bad do you understand?			
All speech listening events	4.6	1.1	[1.0, 1.2]
Natural speech in quiet	4.6	1.1	[0.9, 1.2]
Natural speech & sounds	4.3	1.2	[0.7, 1.7]
Electro-acoustic. speech	4.7	1.0	[0.8, 1.2]
How pleasant are the sounds?	4.5	0.5	[0.2, 0.7]

Table 1: Effect of HA intervention estimated in a mixed model regression analysis. Assessments of speech listening events regressed on condition (pre versus post intervention). Scale orientation: nothing at all (1), no effort (1), very unpleasant (1).

More importantly, the results of this study emphasize the between-subjects variability with regard to the types of situations that were assessed as well as with regard to the assessment patterns. Providing snapshots of real-life hearing, EMA has pinpointed the diversity of listening experiences. Since individualization is crucial in hearing health

care, EMA has been found to be a suitable method for specifying individual challenges and issues. Thus, the main strength of this field study was to demonstrate that EMA can be incorporated in the process of HA fitting with elderly adults completely naïve towards hearing studies and the respective tests and questionnaires. Participants were normal clients of HA acousticians, initially not prepared to take part in any study of this kind. They received feedback of their EMA results and shared them with their HA acoustician. By this means, EMA has the potential to encourage, guide and substantiate the dialogue in the clinical practice, especially when clients are reserved or unable to find the appropriate terms to describe their experiences. However, there are limitations to this study. The main limitation is that carry-over effects are not controlled in the AB design used in this study. Thus, the robustness of effects is not confirmed. An ABA-design including a withdrawal of HA as reported by Timmer *et al.* (2018) was not viable due to ethical reasons. In this context, the positive HA intervention effect on the pleasantness of sounds should be examined more closely. The order of items in the EMA survey might impact the assessments. It is unclear, for example, whether the assessments of listening effort and speech understanding are unchanged if either one or the other is assessed first. Additionally, it is still an open question whether the participants were able to keep the concepts of the hearing dimensions, particularly listening effort and speech understanding, separated and consciously present. These issues cannot be settled based on the data from this study, but certainly merit further attention.

ACKNOWLEDGEMENT

Thanks to Alejandro Garavito Arango, Maximilian Hehl, Florian Schmitt, Kristin Sprenger, Annäus Wiltfang, and the local hearing aid acousticians. Funded by the Hearing Industry Research Consortium (IRC) and the Research Fund of Jade University of Applied Sciences.

REFERENCES

Galvez, G., Turbin, M.B., Thielman, E.J., Istvan, J.A., Andrews, J.A., and Henry, J.A. (**2012**). "Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users," Ear Hearing, **33**, 497-507. doi:10.1097/AUD.0b013e3182498c41.

Holube, I., von Gablenz, P., and Bitzer, J. (under review) "Ecological Momentary Assessment (EMA) in audiology – current state, challenges, and future directions." Submitted to Ear Hearing (special issue).

Kowalk, U., Kissner, S., von Gablenz, P., and Holube, I. (**2018**). "An improved privacy-aware system for objective and subjective ecological momentary assessment," Proc. ISAAR, **6**, 25-30B. Retrieved from https://proceedings.isaar.eu/index.php/isaarproc/article/view/2017-04

Timmer, B., Hickson, L., and Launer, S. (**2018**). "Do hearing aids address real-world hearing difficulties for adults with mild hearing impairment? Results from a pilot study using Ecological Momentary Assessment," Trends Hear., **22**. doi: 10.1177/2331216518783608.

Hearing-aid settings in connection to supra-threshold auditory processing deficits

RAUL H. SANCHEZ-LOPEZ[1,*], TORSTEN DAU[1] AND MORTEN LØVE JEPSEN[2]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark*

[2] *WS Audiology, Lynge, Denmark*

Plomp (1986) described the consequences of hearing impairment in speech communication as the sum of two components: attenuation and distortion. Recent studies have shown that the sensitivity to spectro-temporal modulations (STM) might be linked to speech intelligibility in noise, suggesting that supra-threshold, or "internal", distortions would affect both speech and STM perception similarly. Furthermore, reduced sensitivity to STM may also affect a listener's preference for a hearing aid (HA) compensation strategy. Here, speech intelligibility and STM sensitivity were measured in 20 hearing-impaired (HI) listeners. One group of the listeners (Group A) showed an inability to detect STM, whereas the other listeners (Group B) exhibited similar thresholds as the control group with young normal-hearing (NH) listeners. The two HI groups participated in a perceptual evaluation experiment using multi-stimulus comparisons (MUSHRA). The audio files were processed by a HA simulator fitted to the individual hearing loss and the performance was rated in terms of four attributes: clarity, comfort, preference and listening effort. A correlation analysis showed that clarity and preference were correlated in Group A whereas comfort and listening effort were correlated in Group B. The classification of HI listeners in auditory profiles might be valuable for efficient HA fitting.

INTRODUCTION

Plomp (1986) proposed a model based on observations of speech-in-noise intelligibility tests. The model describes the consequences of hearing loss (HL) in speech communication as the sum of two components: attenuation and distortion. According to the model, listeners with only an attenuation component exhibit elevated speech reception thresholds (SRT) in quiet but their performance in speech-in-noise tests is comparable to the one of a normal-hearing listener. In contrast, listeners with a distortion component show elevated SRTs both in quiet and in noise. The SRT model of Plomp constitutes a scientifically founded way for quantifying the effects of hearing impairment beyond the audiogram (Soli and Wong, 2008). The model relies on restrictive assumptions related to the test procedure (Plomp, 1986), which makes the comparison of results across different studies difficult, especially when

*Corresponding author: rsalo@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

using other speech materials, spatial configurations or noise types. Therefore, other auditory tests, able to indicate whether a distortion component is present, would be of interest for a further hearing loss characterization. Several studies have aimed to characterize the distortion component by studying the relationship between different aspects of auditory processing, such as frequency selectivity and temporal processing, and speech intelligibility (Houtgast and Festen , 2008). However, none of the studied auditory processing deficits could fully account for the degraded speech perception results.

Hearing aid (HA) fitting is commonly based on applying a frequency-dependent non-linear amplification based on the pure-tone audiogram. Current hearing-aid technology allows advanced signal processing that can improve the signal-to-noise ratio (SNR) of the acoustic input signals presented to the ears. The advanced features are usually activated on-demand and not adjusted to the individual needs. The personalization of the HA settings, based on outcome measures that reflect supra-threshold auditory processing deficits, may therefore improve the individual listener's satisfaction. Here, two types of noise management strategies were evaluated by means of subjective assessments. It is hypothesized that listeners with a high degree of supra-threshold deficits would indicate a preference for more aggressive noise reduction.

The mammalian primary auditory cortex encodes dynamic signals by performing a spectro-temporal decomposition of the neural response (Kowalski *et al.*, 1996). The analysis of spectro-temporally modulated signals by the auditory system might be crucial for the discrimination of complex sounds, such as speech. Mehraei *et al.* (2014) investigated the spectro-temporal modulation (STM) sensitivity in normal-hearing (NH) and hearing-impaired (HI) listeners and its relation to speech intelligibility in noise. Results using one-octave wide carriers showed a significant difference between NH and HI listeners for some specific combinations of spectral and temporal modulations. These conditions, combined with the individual audibility predicted by the speech intelligibility index (SII; ANSI S3.5-1997), were able to account for 89% of the variance in the data. The main hypothesis of the present study was that a reduced STM sensitivity is associated with supra-threshold auditory processing deficits, which cause distortions in the internal representation of the acoustic stimuli, such that listeners with a reduced STM sensitivity would exhibit a distortion component that causes elevated speech reception thresholds in noise.

The goal of the present study was to explore the viability of a classification of the HI listeners based on their supra-threshold deficits by means of a simple STM test. The listeners were divided into two groups based on their STM sensitivity: Group A, with a reduced STM sensitivity; and Group B, with a fairly high STM sensitivity. This classification was used to explore the differences between both groups in terms of speech intelligibility and preference for hearing-aid processing. It is hypothesized that Group A would show a poorer speech-in-noise discrimination and a preference for more aggressive settings in a HA noise management algorithm, whereas Group B would show near-normal speech-in-noise intelligibility.

METHOD

Twenty-six subjects participated in this study, divided into a control group (NH) and a HI group with different degrees of sensorineural hearing loss. The NH group consisted of five listeners (1 female) aged between 22 and 58 (median 24) years. Their audiograms did not show any threshold above 20 dB hearing level (HL), and no air-bone gaps were observed. The age of the HI group (20 listeners) ranged between 26 and 86 years (median 69, 12 females). The listeners were divided into two groups. The criterion for placing subjects in one of the two groups was chosen to be at -3 dB STM sensitivity in the listeners' better ear. This criterion was based on the results from Bernstein *et al.* (2016), where the average STM sensitivity was found to approximately -3 dB. The audiometry was performed following the standard ISO 8253-1:2010 and using a two-channel audiometer Interacoustics AA222 and Sennheiser HDA200 headphones. At least one ear of each subject was explored by the whole test battery.

Experimental set-up

All tests were carried out with the same equipment. For the behavioural tests, the stimuli were generated in MATLAB at a sampling frequency of 44.1 kHz and converted in analogue signals by a sound card (RME Fireface). The software Senselab Online 3.1.6 was used for the subjective assessments. The signals were amplified (RME QuadMic) in the analogue domain and presented to the listener through Etymotic research ER-2 insertion earphones with foam ear tips. The tests were performed in a double-walled sound-attenuating booth.

Behavioural tests

Speech reception thresholds were measured using the Danish hearing-in-noise test (HINT; Nielsen and Dau, 2011). The SRT was measured in quiet (SRT_Q) and in speech-shaped stationary noise (SRT_N). The level of the noise was set at $SRT_Q + 30$ dB. For convenience, Lists 1 and 3 were always used for SRT_Q, and Lists 7 and 8 were used for the SRT_N measurements. The SRT_Q was defined as the speech level, in dB sound pressure level (SPL), at which 50% of the sentences are correctly recognized. The SRT_N was defined in terms of SNR. The adaptive procedure for the speech-in-noise test, as described in Nielsen and Dau (2011), was used for estimating SRT.

One-octave wide moving-ripple stimuli were generated in a similar way as in Mehraei *et al.* (2014). Two STM conditions were found to be the most significant predictors of speech-in-noise performance in Mehraei *et al.* (2014). However, only the condition with the ripple centered at $f_c = 1000$ Hz, a spectral density of $\Omega = 2c/o$, and an amplitude modulation frequency of $f_m = 4$ Hz was considered. The stimuli were generated in the frequency domain as the sum of 32 equal-amplitude carrier tones per octave band, logarithmically spaced. All carriers were presented in random phase. Sinusoidal amplitude modulation was applied to the carriers by additional side-bands with instantaneous phases increasing according to the frequency space. Unmodulated

carriers, presented at 15 dB below the level of the modulated band, served to control for off-frequency listening. The stimuli in the STM detection task were presented at the same level as the speech signal at SRT_N. The subjects' task was to detect which interval contained the STM stimulus in a 3-interval 3-AFC paradigm. In the initial trial, the target signal was fully modulated whereas the other two intervals were unmodulated. A three-down one-up tracking procedure approximated the 79.4% point of the psychometric function. The modulation depth in dB [20log(m)] was decreased in steps of 6 dB until the first reversal. The step-size was then decreased to 4 dB for the next two reversals. The threshold was estimated as the mean of six additional reversals with the final step size of 2 dB. Two repetitions were obtained for each ear. All signals were 500 ms long, including 5 ms raised-cosine ramps, separated by 500 ms silence.

Subjective assessment

The subjective assessment was based on a multi-stimulus comparison paradigm (ITU-R.BS.1534-1, 2001) combined with a HA simulator. The hearing-aid simulator consisted of an 18-channel wide-dynamic-range compressor (WDRC) followed by a noise-management algorithm based on a speech intelligibility index (SII) optimizer (Kuk and Paludan, 2006), referred to here as the "speech enhancer" (SE). Individual gain, following the NAL-NL2 prescription, was applied before the noise management algorithm. Two settings of the SE were evaluated: $SE1^{0}_{-12}$ (aggressive noise reduction) where the gain in each frequency band was reduced in a range from 0 to -12 dB to optimize the speech intelligibility of the input signal according to SII; and $SE2^{+6}_{-6}$ (less aggressive noise reduction) where the gain could be increased or reduced in a range from -6 dB to 6 dB to optimize SII. Although the two settings aimed for improved speech intelligibility, the additional gain provided by SE2 may affect listening comfort (Kuk and Paludan, 2006). The noise management algorithm analyzed the long-term signal and applied the gain reduction over the whole sound sample.

The multi-comparison test was implemented in SenseLab Online 3.1.6 (SenseLab, 2015). Five HA settings were presented in each run, including a reference (+6 dB SNR), a hidden anchor (-6 dB SNR) and a control (noise-management OFF). The evaluation was performed in three challenging noisy environments (café, party and traffic) and two SNR conditions, a difficult condition (+1 dB SNR) and a more favourable (+4 dB SNR). The target signal was running speech taken from an audiobook. Four attributes were considered: clarity (contrast between the speech signal and the noise), comfort (comfortability of the whole sound scene), listening effort (difficulties to understand the speech target signal) and preference.

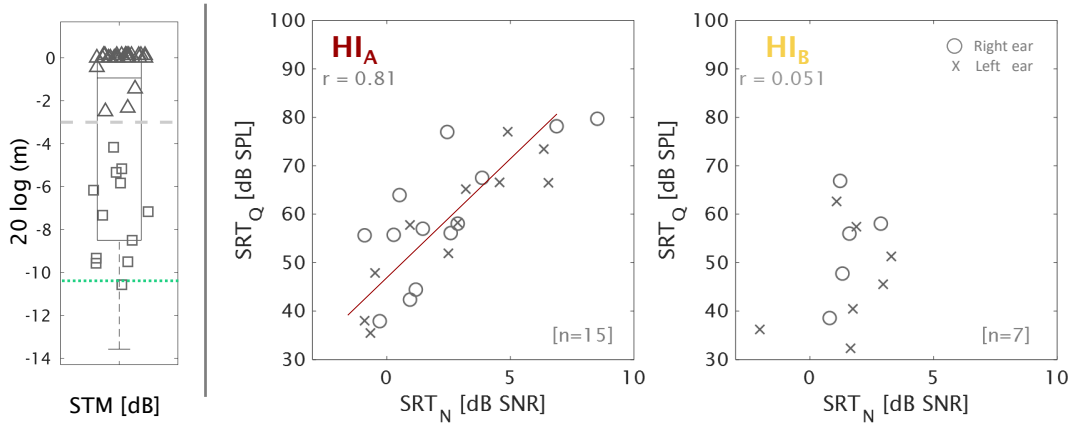


Fig. 1: Results of the behavioural tests. Left: Spectro-temporal modulation detection. The grey dashed line corresponds to -3 dB and the green dotted line corresponds to the averaged results of the NH group. Right: Speech intelligibility tests in quiet and noise for the two HI groups. The origin of coordinates corresponds to the averaged thresholds of the NH listeners.

RESULTS

Figure 1 (left) shows the STM sensitivity results of Group A, indicated as (△), and of Group B, indicated as (□). In Group A, eight listeners were not able to perform the test, reporting that there was no difference between the unmodulated and the STM stimuli. The middle and right panels of Fig. 1 show the results of the speech intelligibility tests. SRT_Q is shown as a function of SRT_N. Group A (middle panel) showed a correlation between SRT_Q and SRT_N whereas for Group B (right panel) SRT_N was not correlated to SRT_Q. Group B's SRT_N values were slightly higher than the SRT values expected for NH listeners (-2 dB SNR).

Figure 2 shows the overall averaged results of the subjective ratings and for the noise types and SNR conditions. Results are shown separately for Group A (top) and Group B (bottom). The ratings of both groups were within a small range between 35 and 65 %. Group A disliked SE2 ($< 40\%$), which provided less comfort ($< 35\%$) and higher listening effort ($> 60\%$), but this group was indifferent in terms of clarity. Group B prefered SE2 ($> 45\%$), which provided more clarity ($> 50\%$) and less listening effort ($< 50\%$) but also less comfort ($< 45\%$). Even though the results were highly correlated across the four attributes, Group A's preference was more significantly correlated[†] with comfort ($r = 0.49$) than with clarity ($r = 0.39$). In contrast, for Group B, preference and listening effort were more significantly correlated with clarity ($r = 0.40$) than with comfort ($r = 0.37$).

[†]Spearman's correlation

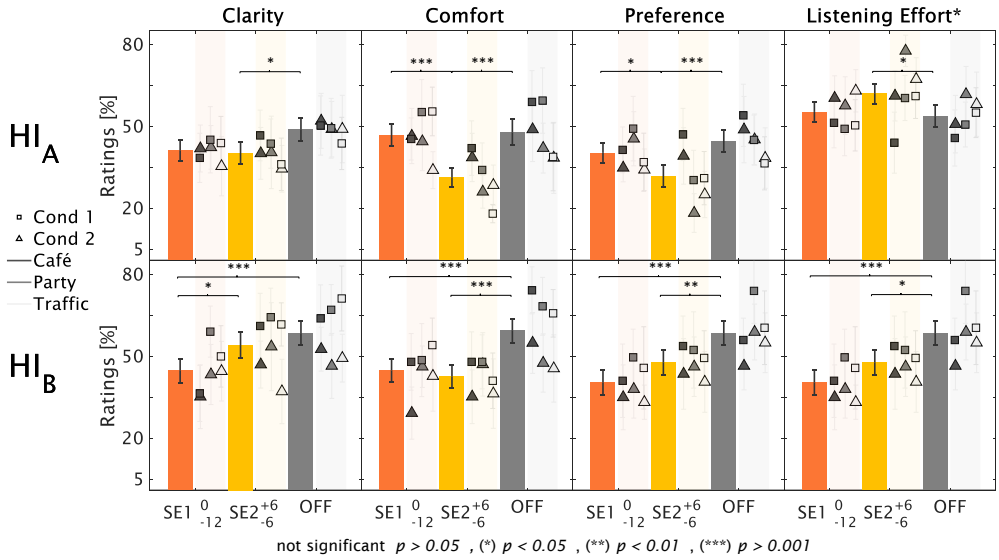


Fig. 2: HA evaluation. Mean and confidence intervals of the ratings of the 4 attributes for the three HA settings. Results of overall Group data (bars) as well as divided by noisy environments and SNR conditions together with 95% confidence intervals. Only seven listeners from each group participated in the subjective assessments

The overall data of the subjective assessments were analyzed using a mixed-linear model that includes the main fixed effects and first-order interactions. The results of the analysis of variance (ANOVA) revealed that (i) the noise type only affected listening comfort ($F(2,800) = 0.7;\ p < 0.001$); (ii) the SNR affected listening effort ($F(1,787) = 11.2;\ p < 0.001$); and (iii) the interaction between HA settings and HI group was significant for all attributes. The higher influence of this interaction was found for preference ($F(4,800) = 6.6;\ p < 0.001$) and the lowest for listening effort ($F(4,787) = 4.5;\ p < 0.01$). The results of the statistical analysis showed only a significant difference between the groups in the ratings of less aggressive setting (SE2) which was preferred by Group B.

DISCUSSION

The participants were divided into two groups based on their performance in the STM sensitivity test. The overall results spanned between -10 and 0 dB and eight listeners were not able to perform the test. This was in line with Bernstein *et al.* (2016), where a substantial number of subjects was not able to perform a similar test and had to be tested with a different test paradigm. Zaar *et al.* (2020) tested 30 HI listeners with similar stimuli as in Bernstein *et al.* (2016) but with an individual hearing-loss compensation that ensures that the presentation level was at 15 dB sensation level (SL) or above. Additionally, the stimuli used in Zaar *et al.* (2020) were slightly longer (1

s). In their study, the STM detection thresholds were between -15 and -5 dB and all listeners were able to perform the test. Furthermore, their results of aided speech-in-noise intelligibility were significantly correlated with the aided STM performance. The frequency-selective amplification and the duration of the stimuli might have made the cues provided by the STM stimulus more salient. In the present study, the goal was to separate the listeners into two groups, associated with different degrees of supra-threshold deficits. Therefore, STM stimuli, such that some listeners are not able to perform the test, might be more sensitive for that purpose.

Plomp argued that the distortion component affects both speech intelligibility in quiet and noise, whereas the attenuation component of the hearing loss only affects speech in quiet and does not yield elevated SRT in noise. In the present study, the group of listeners with lower STM sensitivity (Group A) showed elevated SRT_Q and SRT_N. In this group, the results obtained for speech intelligibility in quiet and in noise were highly correlated to each other. In contrast, in Group B, the speech intelligibility results (in quiet and in noise) were not correlated, suggesting that SRT_Q, in this case, might have only been affected by the attenuation component. Despite the fact that SRT_N values of Group B were slightly elevated compared to the SRT_N values of NH listeners, Group A is more likely to reflect supra-threshold distortions than Group B.

Both groups preferred SE OFF over the two evaluated algorithms. The reason for this could be that the signals were not adjusted in terms of loudness after noise management. Besides, the frequency responses of the reference (+6dB SNR condition) and the SE OFF setting were identical. This might have influenced the judgments of the participants by rating SE OFF higher because of its similarity with the reference. In the group with reduced STM sensitivity (Group A), there was not a significant preference for aggressive noise reduction compared to the SE OFF setting. However, the listeners of Group A preferred the aggressive noise reduction over the speech enhancer with additional gain. Zaar *et al.* (2020) evaluated the benefit of different parameters of a noise-reduction algorithm by means of speech intelligibility and listening effort, as well as the subjective preference in a field study with hearing aids. STM sensitivity was correlated with a preference for noise-reduction settings (i.e., listeners with poorer STM sensitivity preferred more aggressive noise reduction). Thus, in both studies, the adjustment of noise-management algorithms based on the STM sensitivity showed potential for an individualized HA fitting.

CONCLUSION

STM sensitivity can be connected to speech-in-noise intelligibility and hearing-aid fitting strategies. Listeners with a low STM sensitivity (i.e. higher thresholds) seemed to prefer more aggressive noise reduction and higher listening comfort, whereas listeners with high STM sensitivity preferred additional gain that may have improved speech clarity. The individualization of HA parameters based on listeners' supra-threshold hearing abilities might increase HA users' satisfaction.

ACKNOWLEDGEMENTS

This work was carried out at the Center of Applied Hearing Research (CAHR) at the Technical University of Denmark (DTU) as part of Raul's MSc thesis, in collaboration with Widex A/S. We want to thank the participants in the study and the colleagues working at both sites at that time, especially P. Holtegaard and A. Wiinberg.

REFERENCES

ANSI S3.5-1997 (**1997**), "Methods for the calculation of the speech intelligibility index," New York: American National Standards Institute, **19**, 90-119.

Bernstein, J. G., Danielsson, H., Hällgren, M., Stenfelt, S., Rönnberg, J., and Lunner, T. (**2016**). "Spectrotemporal modulation sensitivity as a predictor of speech-reception performance in noise with hearing aids," Trends Hear., **20**, 2331216516670387. doi: 10.1177/2331216516670387

Hougaard S, Ruf S. (**2011**) "EuroTrak I: A consumer survey about hearing aids in Germany, France and the UK," Hearing Review, **18**(2), 12-28.

Houtgast, T., and Festen, J. M. (**2008**). "On the auditory and cognitive functions that may explain an individual's elevation of the speech reception threshold in noise," Int. J. Audiol., **47**(6), 287-295. doi: 10.1080/14992020802127109

Kowalski, N., Depireux, D. A., and Shamma, S. A. (**1996**). "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra," J. Neurophysiol., **76**(5), 3503-3523. doi: 10.1152/jn.1996.76.5.3503

Kuk, F. K., and Paludan-Müller, C. (**2006**). "Noise-management algorithm may improve speech intelligibility in noise," Hear. J., **59**(4), 62. doi: 10.1097/01.HJ.0000286697.74328.32

Mehraei, G., Gallun, F. J., Leek, M. R., and Bernstein, J. G. (**2014**). "Spectrotemporal modulation sensitivity for hearing-impaired listeners: Dependence on carrier center frequency and the relationship to speech intelligibility," J. Acoust. Soc. Am., **136**(1), 301-316. doi: 10.1121/1.4881918

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiolog., **50**(3), 202-208. doi: 10.3109/14992027.2010.524254

ITU-R, Recommendation. "Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra)," ITU, Geneva (**2001**).

Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J. Speech Lang. Hear. R., **29**(2), 146-154. doi: 10.1044/jshr.2902.146

SenseLabOnline (v3.1.6), Listening test software, Hørsholm, Denmark: Delta, **2015**.

Soli, S. D., and Wong, L. L. (**2008**). "Assessment of speech intelligibility in noise with the Hearing in Noise Test," Int. J. Audiol., **47**(6), 356-361. doi: 10.1080/14992020801895136

Zaar, J., Simonsen, L. B., Bherens, T., Dau, T., and Laugesen, S. (**2020**) "Investigating the relationship between spectro-temporal modulation detection, aided speech perception, and directional noise reduction preference in hearing-impaired listeners," Proc. ISAAR, **7**, 181-188.

Hearing aid feature profiles: the success of rehabilitation

SIMON LANSBERGEN[1,*] AND WOUTER DRESCHLER[1]

[1]*Amsterdam UMC, Univ of Amsterdam, Clinical and Experimental Audiology, Meibergdreef 9, Amsterdam, Netherlands*

We recently developed a method to objectively classify hearing aids, using technical data (e.g., compression, noise reduction, etc.) from over 3900 different devices. This yielded hearing aid subgroups called 'modalities', that were characterized as distinct feature profiles, independent of manufacturer or type. Our present study aims to combine these objectively defined modalities with audiologically relevant rehabilitation needs, using data including audiological diagnostic tests and two questionnaires for subjective ratings. We investigated which hearing aid modalities contribute to successful rehabilitation results, and to which extent these modalities can be associated with specific rehabilitation needs. Our results indicate that more adjustable hearing feature channels or levels do not necessarily lead to better rehabilitation results.

INTRODUCTION

The choice of a hearing aid that covers the rehabilitation needs of a hearing-impaired person is an important starting point for a successful rehabilitation. This study combines technical hearing aid data based on publicly available information (e.g., hearing aid datasheets) with user data that includes the individual rehabilitation needs of the hearing impaired. The user data consists of various measures, such as audiogram, demographical data, and subjective data. The hearing aid data used in this study was made independent of brand or manufacturer using previously defined subgroups of hearing aids, based on a clustering of hearing aid features (Lansbergen and Dreschler, 2020). The resulting subgroups of hearing aids, referred to as 'modalities', were characterized by particular profiles, representing the complex interplay between the selected hearing aid features.

This study is a first step in using hearing aid modalities as a selection tool to achieve successful hearing aid rehabilitation at group level. We consider the combined effects of hearing aid features by investigating the effects of modalities rather than isolated features on user perceived benefit.

METHOD

Two datasets were used in this study. One dataset with user data on hearing aid rehabilitation, and a second dataset on technical data from hearing aids. User data were collected between 2015 and 2017 during the regular hearing aid rehabilitation process in the Netherlands, including both new and experienced users. Hearing aid selection was always done by an audiologist or dispenser, where the professional

*Corresponding author: s.e.lansbergen@amsterdamumc.nl

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

ideally presented a choice of two or more hearing aids to the patient. Subjects assessed their hearing aid rehabilitation process before and after a trial period, that included selection and fitting of a hearing aid. They were asked to evaluate perceived benefit based on personal rehabilitation goals and the degree of auditory disability. For this subjective evaluation two questionnaires were used: the Client Oriented Scale of Improvement (COSI) by Dillon *et al.* (1997) and a slightly adapted version of the Amsterdam Inventory for Auditory Disability and Handicap (AIADH), called AVAB[1]. COSI evaluates personal rehabilitation goals by measuring the degree of change (DC) and the final ability (FA) due to the hearing aid fit. AVAB evaluates predefined listening conditions before and after the hearing aid fit on six dimensions of auditory disability[2]. AVAB post-evaluation results could be thought of as a measure of FA, similar to the FA of COSI. Likewise, the difference between AVAB item scores prior to the hearing aid fit and post hearing aid fit, could be thought of as a measure of DC. To estimate AVAB DC measure, differences between pre- and post-AVAB were normalized based on the ratio of the actual pre- and postscore difference and the maximum possible difference (maximum benefit is 100%)[3].

Each of the 32 AVAB questionnaire items were related to one of six dimensions of auditory disability: detection of sounds (Det), sound discrimination (Dis), auditory localization (Loc), speech in quiet (SiQ), speech in noise (SiN), and noise tolerance (Tol). This resulted in mean AVAB scores per dimension of auditory disability, and overall mean AVAB scores. Personal COSI goals were matched to one or more dimensions of auditory disability (Dreschler and de Ronde-Brons, 2016), consequently, mean COSI scores for the dimensions of auditory disability could also be obtained (Lansbergen *et al.*, 2018).

Hearing aid modalities

The hearing aid data contain technical information of hearing aids, such as the number of compression channels, and were available on the Dutch market in March 2018. Data was provided by the manufactures through their hearing aid datasheet and was checked by a group of audiologists. After a selection procedure, data from 2106 behind-the-ear (BTE) hearing aids were included, containing all major, but also some lesser known, hearing aid manufacturers. The selection relied on the following criteria: (i) no missing data and (ii) no ambiguity (i.e., conflicting technical details).

The dataset contained about 50 of the most important characteristics of a hearing aid. After applying a data processing procedure, a set of 10 key hearing aid features were identified as relevant for audiological rehabilitation and were used as input for

[1]The original version of AIADH was developed by Kramer *et al.* (1995), details on the used AVAB questionnaire could be found in Dreschler and de Ronde-Brons (2016).

[2]Kramer *et al.* (1995) defined 'auditory disability' as the difficulties experienced in everyday hearing.

[3]This research used the same data as described by Lansbergen *et al.* (2018). A more detailed description of the specific data gathering methods of the user data can be found in their paper.

further analysis (Lansbergen and Dreschler, 2020). The latter was done using Latent Class Tree Analysis (van den Bergh *et al.*, 2017), which is an extension of the better known Latent Class Analysis clustering method. Using this method, we extracted six mutually exclusive hearing aid modalities from the hearing aid data, see Figure 1.

Statistical analyses

Computation of mean COSI scores for the dimensions of auditory disability resulted in missing data for one or more dimensions. We therefore used a linear mixed-effect model on the COSI and AVAB scores because such models allow unequal variances and can accommodate unbalanced data. The type of hearing aid modality was used as between-subject factor and the type of auditory disability dimension as within-subject factor. Complementary post-hoc analysis was done using the Games-Howell pairwise multiple comparison procedure, because this procedure can accommodate unbalanced group sizes. Cohen's *d* was computed to examine the effect sizes between hearing aid modalities that showed significant differences on the post-hoc analyses.

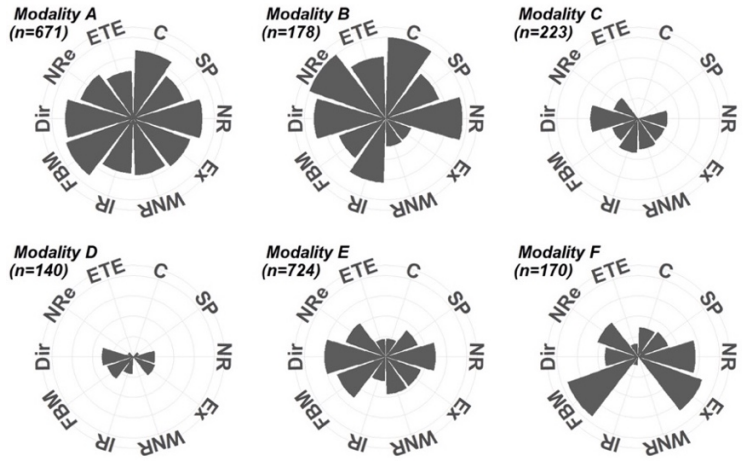


Figure 1: Six mutually exclusive hearing modalities, for behind-the-ear type hearing aids. The modalities were defined by ten hearing aid features: (C) compression channels; (SP) sound processing channels; (NR) noise reduction levels; (Ex) expansion levels; (WNR) wind noise reduction levels; (IR) impulse reduction levels; (FBM) feedback manager; (Dir) directionality type; (NRe) Noise reduction environments; (ETE) ear to ear communication type. The values on the radar chart represents mean features measures (i.e. data rescaled between 0 and 1). The n-values indicate the number of unique hearing aids associated with the modality.

RESULTS

User data from 1149 subjects were included. Because the hearing aids that were used for rehabilitation were known for each subject, user data could be merged with matching hearing aid data, including the corresponding modality. Only a small number of subjects were fitted with in-the-ear type hearing aids (29 subjects) and

were excluded from the data. Remaining subjects were all fitted with a behind-the-ear type hearing aid. The mean age of the included subjects was 67.7 years (SD ± 13.2 years, range: 20-98 years). The weighted binaural hearing loss for 0.5, 1, 2 and 4 kHz were calculated using the average values of the better ear and the worst ear in a ratio of 5:1, considering that overall hearing disability is mainly determined by the better hearing ear in subjects with asymmetric hearing loss. The mean binaural hearing loss of all subjects was 45.3 dBHL (±14.6 dBHL). Mean DC and FA scores for COSI and AVAB are shown in Table 1, along with the distribution of matched modalities. The vast majority of subjects were fitted with hearing aids from modality A (n=376), E (n=628) or F (n=96).

Modality	***A***	***B***	***C***	***D***	***E***	***F***
n	376	29	13	7	628	96
%	32.7%	2.5%	1.1%	0.6%	54.7%	8.4%
Mean COSI DC	4.03	4.03	4.27	4.24	4.25	4.24
Mean COSI FA	4.17	4.27	4.42	4.48	4.38	4.35
Mean AVAB DC	43.8%	52.8%	43.7%	49.7%	46.6%	38.3%
Mean AVAB FA	3.07	3.18	3.20	3.57	3.24	3.25

Table 1: Questionnaire data matched to hearing aid modalities (A-E), n indicates number of devices in each modality. COSI DC, COSI FA and AVAB FA are expressed in mean scores: COSI scores ranges between 1-5, AVAB FA scores ranges between 1-4. AVAB DC is expressed in a ratio of pre- and post-rehabilitation scores.

The results of linear mixed-effect model showed that differences between modalities were significant for the measures: COSI DC ($F(5,1125) = 5.38$, $p < 0.001$); COSI FA ($F(5,1122) = 5.22$, $p < 0.001$); and AVAB FA ($F(5,1143) = 9.73$, $p < 0.001$). Post-hoc analysis revealed that mean scores for COSI DC and FA were significantly higher for subjects that used modality E hearing aids relative to modality A ($p < 0.05$, Table 2). Furthermore, differences between AVAB FA and DC scores were also significant for modality A and F. Effect sizes of the differences between modalities varied between -0.28 and -0.42 (Table 2).

	COSI	AVAB
Degree of Change	E>A; -0.33	
Final Ability	E>A; -0.33 F>A; -0.28	E>A; -0.42 F>A; -0.40

Table 2: Post-hoc analysis results (Games Howell, α=0,05). Only significant differences between hearing aid modalities were shown, followed by the effect size (Cohen's *d*).

Dimensions of auditory disability

A linear mixed-effect model showed that interaction between the type of modality and the dimensions of auditory disability was significant for AVAB DC ($F(25,5715)$

= 1.71, $p = 0.01$) and AVAB FA ($F(25,5715) = 2.52$, $p < 0.001$). COSI DC and FA per dimension of auditory disability were not significantly dependent on the type of modality. The results from the post-hoc analysis are displayed in Table 3. Effect sizes for the significant post-hoc results were calculated to interpret the impact of the results. There was an overall trend that COSI and AVAB scores for modalities E and F were better than those of modality A. Though, one exception was found in the dimension SiQ, where AVAB DC scores were better for modality A as compared to modality F. Effect sizes of significant differences between modalities varied between -0.12 and -0.47.

AVAB	*Det*	*Dis*	*Loc*	*SiN*	*SiQ*	*Tol*
Degree of Change				**E>A**; -0.12	**A>F**; 0.43 **E>F**; 0.47	
Final Ability	**E>A**; -0.35 **F>A**; -0.44	**E>A**; 0.32	**E>A**; -0.45 **F>A**; -0.35	**E>A**; -0.43 **F>A**; -0.46	**E>A**; -0.20	

Table 3: Post-hoc (Games Howell, $\alpha = 0,05$) analysis of mean AVAB DC and FA scores between the (most relevant) hearing aid modalities for each of the six dimensions of auditory disability, followed by the effect size (Cohen's *d*).

DISCUSSION

The focus of this study was to combine user experience of hearing aid rehabilitation with objective, technical hearing aid data, expressed in terms of hearing aid modalities. In this paper we present preliminary results, which address the relation between hearing aid modalities and user benefit. Modalities were defined using a data driven approach, which resulted in groups of hearing aids, independent of brand or type. Our results indicate that on a group level, significant differences exist between hearing aid modalities and individual rehabilitation goals evaluated with COSI. Similarly, this was also true for modalities and the dimensions of auditory disability evaluated with AVAB. Our study found that better COSI and AVAB scores could not be explained by an overall increase in feature potential. This is important information for the selection of hearing aids, as it suggests that more advanced hearing aids might not always solve auditory disability or be supportive for individual rehabilitation goals. The results imply that modalities can be considered a suitable and objective tool to support evidenced based hearing aid selection.

Feature potential expresses the mean number of channels, levels or type of a particular feature within a modality, rescaled between 0 and 1 (Figure 1). Accordingly, hearing aids related to modality B should be considered hearing aids with the highest feature potential. Yet, only in a few instances a hearing aid from this modality was used for rehabilitation. Nevertheless, modality A can also be considered as a modality with a high overall mean feature potential and was used in a large number of subjects. Interestingly, we found that over 95% of all hearing aids used for rehabilitation were related to either modality A, E, or F. Also, most subjects were fitted with a hearing aid with intermediate feature potential (modalities E and

F). The low number of used hearing aids that were related to modalities C and D, was perhaps less surprising. These two modalities represent hearing aids with a very limited mean feature potential. The selection of a specific hearing aid might be biased by the preference of the professional towards a certain brand or type. In the case of the dispenser, selection might also be driven by commercial interest. Further analysis of the existing data will provide a better understanding of the hearing aid selection process.

It is striking that the subjects in this study reported higher scores for hearing aids with an intermediate feature potential (modalities E and F), as compared to hearing aids that were related to a modality with a high overall feature potential (modality A). This clearly implies that the availability of a wider range of adjustable hearing aid feature channels or levels, does not necessarily result in more beneficial rehabilitation. This seems to be especially true for the number of compression channels, as differences for this feature were most pronounced between modality A and modalities E and F. Previously, Cox *et al.* (2014) reported a very small effect (d = -0.06) between hearing aids with 'premium' and 'basic' technology for older listeners (mean age 70.4 years) with mild to moderate sensorineural hearing loss, using several questionnaires. They considered that result to be non-significant. The terms 'premium' and 'basic' were also related to the number of available and/or adjustable hearing aid features and can thus be considered as a first-order approach of the modalities used in this study. In line with their findings, we also didn't find evidence that hearing aids with a higher feature potential (modalities A and B) were reported more beneficial as compared to hearing aids with a limited feature potential (modalities C and D).

Using objective measures (HASPI and HASQI) to evaluate hearing aid benefit, Kates *et al.* (2018) found some significant differences between manufactures, yet, they also found no significant differences between 'basic' and 'premium' hearing aids. They conclude that 'the similarity in performance between basic and premium devices suggest that increased processing complexity does not necessarily lead to improved performance'. Again, we conclude similar results. However, we compared a large number of commercially available hearing aids that were evaluated by a large group of users. This enabled us to extend comparisons beyond the dichotomy between 'premium' and 'basic' hearing aids. Furthermore, higher DC and FA scores were found for hearing aid types with a more intermediate feature potential as compared to hearing aids types with a high feature potential. This might be explained by the fact that this research included a vastly larger selection of different hearing aids. As a result, we were able to use a refined hearing aid classification method. Hence, differences between modalities are more detailed than just the dichotomy between 'basic' and 'premium'. The AVAB dimensions of auditory disabilities relate to distinct real-life hearing difficulties. The benefit of hearing aid features might therefore also differ between these dimensions. We found that AVAB DC results for the dimensions SiN and SiQ were dependent on the type of modality, but only for modalities A, E, and F. Interestingly, the only instance that subjects reported a significant higher benefit from a 'premium' type hearing aid, was in

relation to the DC for the SiQ dimension. Differences in the level of AVAB scores for the 'speech' dimensions were significant, with mostly a medium effect size. This translates into important differences in perceived benefit. A possible explanation for this might be that problems related to these dimensions were prioritized during the hearing aid fit, regardless the selection of the hearing aid.

The AVAB FA results showed more significant differences between AVAB results per dimension and type of modality, except for the Tol dimension. The effects were most pronounced for differences between modalities A and E, and A and F. It could be argued that hearing aids with less complex directionality and less processing channels are more capable to handle problems related to auditory localization. In general, for most dimensions subjects reported the highest FA with a hearing aid that has an intermediate mean feature potential This indicates that hearing aids with the highest available and adjustable feature channels/levels are not always required for a good hearing aid fitting result.

Limitations and future

The existing data has a large potential for more elaborate analysis. For instance, it could be fruitful to examine differences between new and experienced users or between females and males, but also the relation to hearing loss and age. We hypothesize that there will be a clear effect of hearing aid experience on the perceived benefit in relation to modalities or relevant features. There might be a relation between hearing loss and/or age with respect to perceived benefit within different modalities. Furthermore, the data was not limited to the current six modalities and the data used to model the modalities could also be used for further analysis. In this respect, a regularization process can be considered to investigate which hearing aid features might be sensitive for the measures DC or FA. Such analyses would complement the present results and, as we expect, will lead to stronger predictions and more detailed conclusions. In a follow-up project the group results have to be translated to a more individual approach, feasible to support the clinical process of hearing aid selection and fitting.

The limitations of the COSI and AVAB questionnaires, as well as the added value of combing these two questionnaires, were previously discussed in Lansbergen *et al.* (2018). Especially, the ceiling effect observed in the COSI scores was found to be a limiting factor with respect to the overall sensitivity of this measure. This lack of sensitivity also translates to a poor analytic power when using the COSI with respect to the dimensions of auditory disability. On the other hand, main effects between mean COSI scores and the type of modality were clearly significant and in agreement with the results obtained with AVAB.

CONCLUSION

This study illustrated that more adjustable hearing aid feature channels/levels do not necessarily result in a larger perceived benefit. This implies that the ability to cope with real-life hearing problems is not solved by merely using more advanced hearing

aids. Differences between perceived benefit of a hearing aid and available feature potential, encapsulated in modalities, were found to be specific for self-formulated rehabilitation goals and hearing problems associated with different dimensions of auditory disability. Although it is too early to fully understand what the underlying reasons for this outcome could be, our results are in line with results that were reported previously. Our results may be of interest for readers that work in the field of rehabilitation, and in particular for hearing aid dispensers and audiologists as it might help in hearing aid selection using an evidenced based method.

ACKNOWLEDGEMENTS

Data collection was organized by the PACT Foundation, a network of co-operating Audiological Centres in the Netherlands. The authors like to thank Bert van Zanten (Academic Medical Centre Utrecht), André Goedegebure (Erasmus Medical Center in Rotterdam), and Wim Soede (Leiden University Medical Centre) for their efforts.

REFERENCES

Cox, R. M., Johnson, J. A., and Xu, J. (**2014**). "Impact of advanced hearing aid technology on speech understanding for older listeners with mild to moderate, adult-onset, sensorineural hearing loss," Gerontology, **60,** 557-568. doi: 10.1159/000362547

Dillon, H., James, A., and Ginis, J. (**1997**). "Client Oriented Scale of Improvement (COSI) and its relationship to several other measures of benefit and satisfaction provided by hearing aids," J. Am. Acad. Audiol. **8,** 27-43.

Dreschler, W. A., and De Ronde-Brons, I. (**2016**). "A profiling system for the assessment of individual needs for rehabilitation with hearing aids," Trends Hear. **20,** doi: 10.1177/2331216516673639

Kates, J. M., Arehart, K. H., Anderson, M. C., Muralimanohar, R. K. and Harvey JR, L. O. (**2018**). "Using objective metrics to measure hearing aid performance," Ear Hearing, **39,** 1165-1175. doi: 10.1097/AUD.0000000000000574

Kramer, S. E., Kapteyn, T. S., Festen, J. M. and Tobi, H. (**1995**). "Factors in subjective hearing disability," Audiology, **34,** 311-320. doi:10.3109/00206099609071948

Lansbergen, S., De Ronde-Brons, I., Boymans, M., Soede, W. and Dreschler, W. A. (**2018**). "Evaluation of Auditory Functioning and Rehabilitation Using Patient-Reported Outcome Measures," Trends Hear. **22,** doi: 10.1177/2331216518789022.

Lansbergen, S., and Dreschler, W. A. (**2020**). "Classification of Hearing Aids Into Feature Profiles Using Hierarchical Latent Class Analysis Applied to a Large Dataset of Hearing Aids," Ear Hearing, Manuscript in press. doi: 10.1097/AUD.0000000000000877

Van den Bergh, M., Schmittmann, V. D. and Vermunt, J. K. (**2017**). "Building latent class trees, with an application to a study of social capital," Methodology - Eur., **13,** 13. doi: 10.1027/1614-2241/a000128

Using BEAR data to obtain reduced versions of the SSQ-12 and IOI-HA-7 questionnaires

TOBIAS PIECHOWIAK[1,*] AND DAVID ZAPALA[2]

[1] *GN Store Nord A/S, DK-2750 Ballerup, Denmark*

[2] *Mayo Clinic, Jacksonville, FL 32224, USA*

The Speech, Spatial and Qualities of Hearing scale (SSQ-12) and the International Outcome Inventory for Hearing Aids (IOI-HA-7) are questionnaires containing 12 and 7 items, respectively. They are designed to subjectively assess hearing ability and are complementary to behavioral measures. Both questionnaires have been applied across a range of clinical and clinical research-related contexts, for example for assessing outcomes of e.g., cochlear implants and hearing aids. However, due to time constraints neither of the questionnaires seem to be an inherent part of standard clinical quality control. The Better Hearing Rehabilitation (BEAR) database contains SSQ-12 and IOI-HA-7 scores of around 1600 subjects. Applying an Exploratory Factor Analysis (EFA) on the data from the 2nd visit allowed us to reduce the SSQ-12 to 5 questions and the IOI-HA to 3 remaining questions. The SSQ-5 explains 79 % of the variance in the SSQ-12 data while the IOI-HA-3 accounts for 70 % of the variance in the original IOI-HA-7. These new versions have the potential to be used more efficiently by shortening time and focusing on the items that are most effective to reflect individual benefit. Furthermore, the analysis seems to confirm the validity of such a reduction from similar findings in the literature that were done on different datasets.

INTRODUCTION

Our understanding of the many factors that may influence a person's auditory capacity has grown substantially over the past decades. We now recognize the importance of identifying and accounting for conditions such as depression, cognitive impairment, willingness to pursue audiological treatment, and overall health. In an effort to account for these factors, there is a growing pressure to deploy multiple questionnaires in the clinic waiting room. Validated questionnaires can be an efficient way to extract useful information in a busy clinic. However, answering long sets of questions can be taxing for the patient to complete and may adversely affect the patient experience. So there is a need to ask questions efficiently. The Speech, Spatial and Qualities of Hearing scale (Gatehouse and Noble, 2004) is a popular questionnaire tool that was designed for use as a complement to behavioral or experimental measures of hearing ability. The scale was intended to sample every day experiences related to speech understanding, auditory spatial perception and related abilities, and sound quality in a way that scores

*Corresponding author: tpiechowiak@gnresound.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

could generalize across individuals and life situations. The original scale has 49 items and has been applied across a range of clinical and clinical research-related contexts (see Akeroyd *et al.*, 2014). For example, it has been used to assess the effects of age on the hearing ability in subjects with "normal" hearing (Banh *et al.*, 2012) and the effects of unilateral hearing loss (Olsen *et al.*, 2012). It has been used to compare the effects of one versus two cochlear implants (e.g., Laske *et al.*, 2009), one cochlear implant versus implant and contralateral linear frequency transposing hearing aid fitting (Hua *et al.*, 2012), and the effects of musical training on cochlear implant performance (Fuller *et al.*, 2012). There are shorter versions of the SSQ already available. Noble *et al.* (2013) proposed a 12-item version of the SSQ (the SSQ-12). Deemester *et al.* (2012) proposed a 5-item SSQ-5. Akeroyd *et al.* (2014) found four factors on the SSQ-49, three clear factors "speech understanding", "spatial perception", and "clarity" with a possible fourth factor named "effort".

The International Outcome Inventory for Hearing Aids (Cox *et al.* (2000); referred to as IOI-HA-7) is a 7-item questionnaire outcome measure which is considered sufficiently general to apply to many different types of investigations carried out across the world and for many different applications. It was developed at a workshop on *Measuring Outcomes in Audiological Rehabilitation Using Hearing Aids* in Eriksholm in Denmark. The IOI-HA is not intended to replace existing outcome measures but to serve as a useful add-on to already existing measures in a research context. Similar to the SSQ it has been used as a standalone tool for quality assessment of hearing-aids and cochlear implants (Noble, 2002; Erixon and Rask-Andersen, 2015).

The Better-Hearing-Rehabilitation Project (BEAR) is a five year project whose purpose is to promote research in clinical audiology, particularly the development of new clinical methods for diagnosis and hearing-aid fitting. The overall goal is to come up with a new extended fitting procedure. One part of this project is to collect a relatively large database of general as well as auditory information about the subjects and their current fitting. This database contains information of around 2000 subjects including the scores from SSQ-12 and IOI-HA-7 for around 1600 subjects. Subjects had to fill out the SSQ at two occasions: inital fitting and a follow-up visit (2nd visit) two-to three month after the inital visit took place which allows for a direct rating of improvement. The IOI-HA was only rated at the 2nd visit. Based on this data in this work we investigated if and to what extend the SSQ-12 and IOI-HA-7 can be reduced further in order to speed up their use and if the structure of those reduced versions correspond to what has been found in earlier studies (Deemester *et al.* (2012) for the SSQ).

METHOD

Multivariate data are often viewed as multiple indirect measurements arising from an underlying source, which typically cannot be directly measured. Exploratory Factor Analysis is a classical technique developed in the statistical literature that aims to identify these latent sources. In this sense it is a dimensionality reduction technique.

The classical factor analysis model was developed by researchers in psychometrics, like Hastie *et al.* (2008).

With $q < p$, a factor analysis model has the form

$$\begin{aligned} X_1 &= a_{11}S_1 + ... + a_{1q}S_q + \varepsilon_1 \\ X_2 &= a_{21}S_1 + ... + a_{2q}S_q + \varepsilon_2 \\ &\vdots \\ X_p &= a_{p1}S_1 + ... + a_{qp}S_q + \varepsilon_p \end{aligned}$$

or in matrix notation

$$\mathbf{X} = \mathbf{AS} + \varepsilon \qquad \text{(Eq. 1)}$$

The parameters basically all reside in the covariance matrix

$$\sum = \mathbf{AA^T} + \mathbf{D}_\varepsilon \qquad \text{(Eq. 2)}$$

with $D_\varepsilon = diag[Var(\varepsilon_1), ..., Var(\varepsilon_p)]$. A is a *pxq* matrix of *factor loadings* and the ε_j are uncorrelated zero-mean disturbances. The factor loadings are used to name and interpret the factors.The **singular value decomposition** (SVD) is a way to decompose quadratic and non-quadratic matrices into their singular-or eigenvalue representation from which the principal factors can be derived. The number of factors is determined by comparing the eigenvalues generated from a data matrix to the eigenvalues generated from a Monte-Carlo simulation created from a random data matrix of the same size and retains those with eigenvalues larger than zero. This is known as parallel analysis and the criteria as *Horn's criterium* for factor analysis (see Dinno , 2014, for an overview).

The analysis here was done with the R programming language (see R Core Team, 2017).

RESULTS AND DISCUSSION

The Exploratory Factor Analysis yields 5 factors for the SSQ-12 and 3 factors for the IOI-HA-7. They explain 79 and 70% of the cumulative variance, respectively. These are also the number of factors with eigenvalues larger than zero (see eigenvalues for SSQ analyis in Figure 1).

The following two figures (2 and 3) show the factor loadings for the individual SSQ and IOI-HA items, respectively. In the figures the factor loadings are stacked in order to see the entire dependency of each single item on the factors.

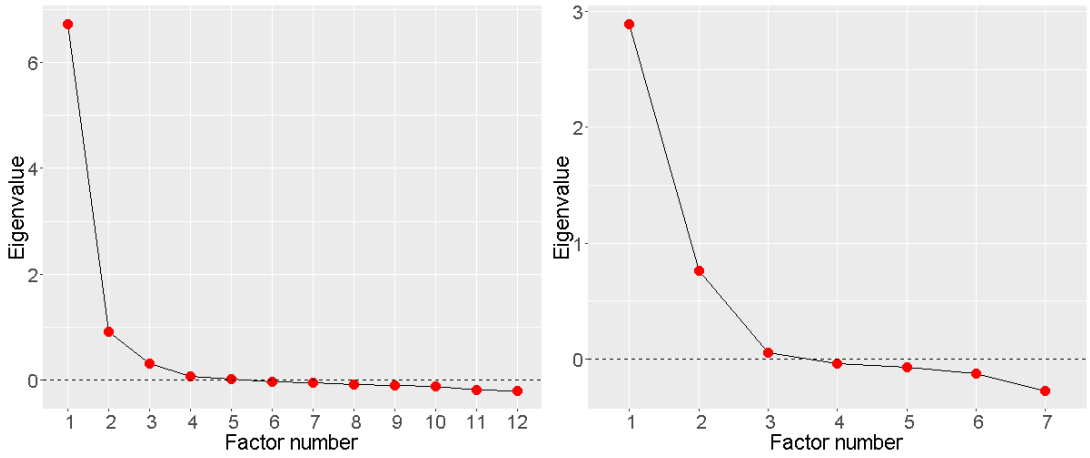


Fig. 1: Scree plot showing the eigenvalues of factors for the SSQ (left panel) and IOI data (right panel). The plot suggests retaining five factors for SSQ and three for IOI according to a parallel analysis which retains eigenvalues larger than zero.

When analysing the factor loadings on each SSQ-12 item one can see that *Factor 1* mainly loads on items relating to speech, while *Factor 2* loads on item relating to space. *Factor 3* loads mainly on sound item #2 which deals with the ability to separate different sound streams. *Factor 4* is clearly associated with listening effort since it correlates mainly with sound item #14. *Factor 5* on the other hand load equally on sound items #7 and #9. They deal with the clarity of everyday sounds and music but not necessarily involving multiple streams of sounds. They are somehow comparable to the findings from Akeroyd *et al.* (2014) who found four factors: *Speech Understanding, Spatial Perception, Clarity and Effort*. In general, good sound clarity means that the sound is as close to reality, or ”like-being-there”, as possible. In this sense, factors *Source Separation* and *Frequency Resolution* contribute to sound clarity. Thus, the findings from this study and those find by Akeroyd *et al.* (2014) do not contradict each other but point to the same factors contributing to the assessment of hearing aid qualities.

One could summarize the factor analysis by tagging each factor with a new label:

Factor labels for SSQ

- Factor 1: *Speech Understanding*
- Factor 2: *Spatial Perception*
- Factor 3: *Source Separation*
- Factor 4: *Listening Effort*
- Factor 5: *Frequency Resolution*

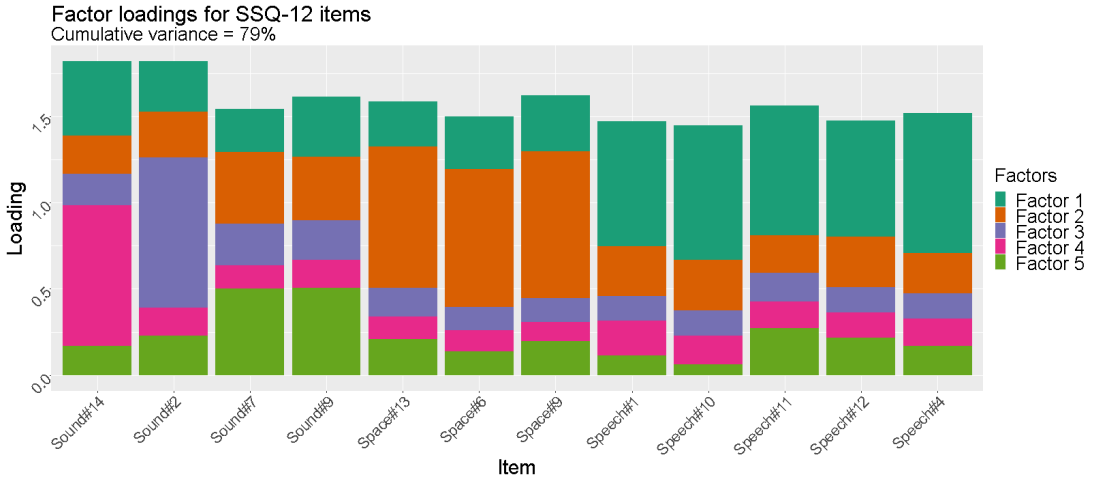


Fig. 2: Results of the factor analyis for the SSQ-12. **Horn's criterium** suggests five remaining factors as seen in Figure 1. These five factors show high correlation with speech understanding, spatial perception, source separation, listening effort, and frequency resolution. See text for details.

Factor labels for IOI-HA

- Factor 1: *Benefits with Hearing Aids*
- Factor 2: *Residual Difficulties with Hearing Aids*
- Factor 3: *Usage*

How do we come up with the items for a potential five item SSQ-5 and three item IOI-HA-3?

In this study, we simply choose those items from the original questionnaires which show the highest correlation (which are the loadings) with the respective factors. Based on this approach we propose the following SSQ-5 and IOI-HA-3:

SSQ-5

- #1: *You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow the conversation?*
- #2: *Can you tell from the sound which direction a bus or truck is moving, e.g., from your left to your right or right to left?*
- #3: *When you hear more than one sound at a time, do you have the impression that it seems like a single jumbled sound?*

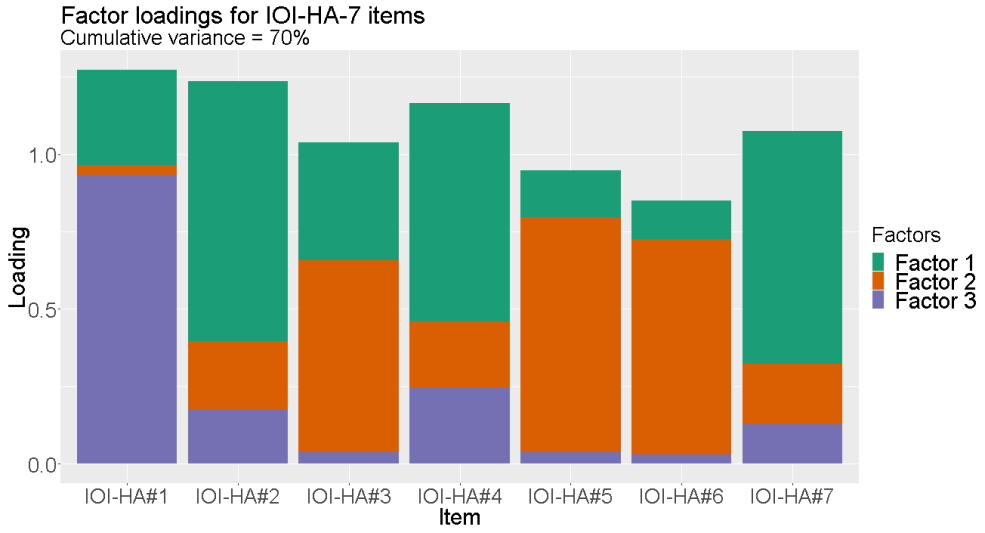


Fig. 3: Results of the factor analyis for the IOI-HA-7. **Horn's criterium** suggests three remaining factors. These three factors show high correlation with speech understanding, spatial perception, source separation, listening effort, and frequency resolution. See text for details.

- #4: *Do you have to concentrate very much when listening to someone or something?*
- #5: *Do everyday sounds that you can hear easily seem clear to you (not blurred)?*

IOI-HA-3

- #1: *Think about the situation where you most wanted to hear better, before you got your present hearing aid. Over the past two weeks, how much has the hearing aid helped in that situation?*
- #2: *Over the past two weeks, with your present aids, how much have your hearing difficulties affected things you can do?*
- #3: *Think about how much you used your present hearing aids over the past two weeks. On an average day, how many hours did you use it?*

Finally, in order to compare the original SSQ-12 with the newly obtained SSQ-5, individual SSQ-12 scores as well as IOI-HA-7 with the obtained IOI-HA-3 original data is plotted against transformed scores with a power function linking the two for the SSQ and the IOI-HA. (left-and right panel Figure 4).

Regarding the SSQ both a power function line and a 1:1 line are also shown in red and black color, respectively. There is close agreement between the two versions of the

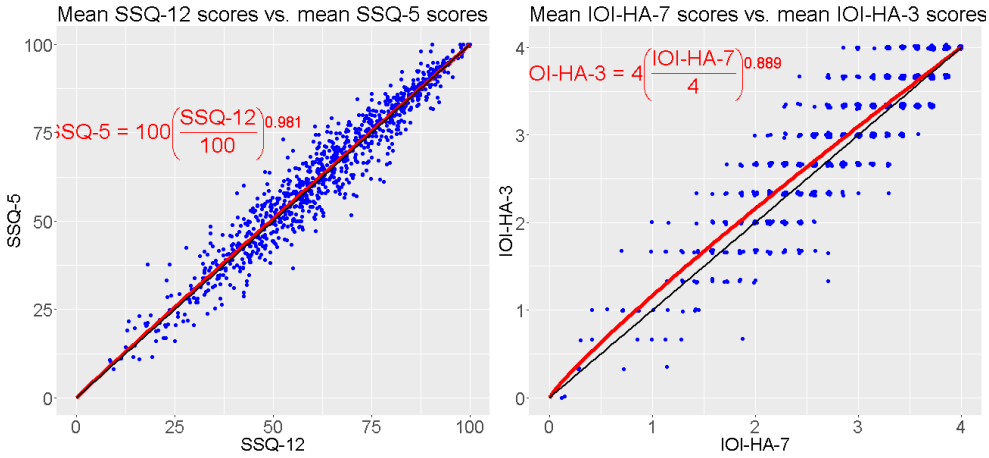


Fig. 4: Left: Individual SSQ-5 scores plotted against SSQ-12 scores. The red line indicates the best fit to a power function given by the equation in the top left corner while the black line shows an 1:1 relationship. They indicate an almost linear relationship between the scales. Right: Individual IOI-HA-3 scores plotted against IOI-HA-7 scores. The red line indicates the best fit to a power function given by the equation in the top left corner while the black line shows a 1:1 relationship

scale the small exponent indicating an almost linear relationship and no bias visible which is indicated by the proximity of the two lines.

For the IOI-HA, the lines indicate two things: (a) a modestly higher IOI-HA-3 average score compared to the IOI-HA-7 due the points lying above the 1:1 line and (b) a slightly steeper slope for the IOI-HA-7 relative to IOI-HA-3 which indicates a higher sensitivity between low-and high scores for the longer IOI-HA-7.

In general, it can be concluded from the current study that it is possible to further reduce the existing SSQ-12 and IOI-HA-7 questionnaires without critical loss of explanatory power (79% and 70% cumulative variance explained). This is in line with findings from other studies (see e.g., Akeroyd *et al.*, 2014; Deemester *et al.*, 2012; Moulin *et al.*, 2019) whose suggestion in reduction of size is similar to our findings. Shorter and more concise questionnaire likely lower the threshold of use of both from a clinician's standpoint and also from the subjects' point of view in the sense that the remaining questions are more precise and contrast more from each other.

REFERENCES

Akeroyd, M. A., Guy, F. H., Harrison, D. L., and Suller, S. L. (**2014**). "A factor analysis of the SSQ (Speech, Spatial, and Qualities of Hearing Scale).," Int. J. Audiol., **53**(2), 101-114

Banh, J., Singh, G., and Pichora-Fuller, M. K. (**2012**). "Age affects reponses on the Speech, Spatial and Qualities of Hearing scale (SSQ) by adults with minimal

audiometric loss.," J. Am. Acad. Audiol., **23**, 81-91.

Cox, R., Hyde, M., Gatehouse, S., Noble, W., Dillon, H., Bentler, R., Stephens, D., Arlinger, S., Beck, L., Wilkerson, D., Kramer, S., Kricos, P., Gagne, J., Bess, F., and Hallberg, L. (**2000**). "Optimal outcome measures, research priorities, and international cooperation," Ear Hear., **21**, 106S-115S.

Deemester, K., Topsakal,V., Hendrickx, J-J, Fransen, E., Van Laer, L., and Van Camp, G. (**2012**). "Hearing disability measured by the Speech, Spatial, and Qualities of Hearing scale in clinically normal-hearing and hearing-impaired middle-aged persons, and disability screening by means of a reduced SSQ (the SSQ5)," Ear Hear., **33**, 615-626.

Dinno, A. (**2014**). "Gently Clarifying the Application of Horn's Parallel Analysis to Principal Component Analysis Versus Factor Analysis," unpublished, `https://alexisdinno.com/Software/files/PA_for_PCA_vs_FA.pdf`

Erixon, E., and Rask-Andersen, H. (**2015**). "Hearing and Patient Satisfaction Among 19 Patients Who Received Implants Intended for Hybrid Hearing: A Two-Year Follow-Up," Ear Hear., **36**, 271-8.

Fuller, C., Free, R., Maat, B., and Başkent, D. (**2012**). "Musical background not associated with self-perceived hearing performance or speech perception in postlingual cochlear-implant users.," J. Acoust. Soc. Am., **132**, 1009-1016.

Gatehouse, S., and Noble, W. (**2004**). "The Speech, Spatial and Qualities of Hearing scale (SSQ)," Int. J. Audiol., **43**, 85-99.

Hastie, T., Tibshirani, R., and Friedman, J. (**2008**). "Elements of Statistical Learning," Springer Series in Statistics.

Hua H., Johansson B., Jönsson R., and Magnusson, L. (**2012**). "Cochlear implant combined with a linear frequency transposing hearing aid.," J. Am. Acad. Audiol., **23**, 722-732.

Laske R. D., Veraguth D., Dillier N., Binkert A., Holzmann D., and Huber, A. M. (**2009**). "Subjective and objective results after bilateral cochlear implantation in adults.," Otol. Neurotol., **30**, 313-318.

Moulin, A., Vergne, J., Gallego, S. and Micheyl, C. (**2019**). "A new speech, spatial, and qualities of hearing scale short-form: factor, cluster, and comparative analyses," Ear Hear., **40**(4), 938-950.

Noble, W. (**2002**). "Extending the IOI to significant others and to non-hearing-aid-based interventions," Int. J. Audiol., **41**, 27-9.

Noble, W., Jensen N. S., Naylor G., Bhullar N., and Akeroyd, M. A. (**2013**). "A short form of the Speech, Spatial and Qualities of Hearing scale suitable for clinical use: The SSQ12," Int. J. Audiol., **52**, 409-412.

Olsen S. O., Hernvig L. H., and Nielsen, L. H. (**2012**). "Self-reported hearing performance among subjects with unilateral sensorineural hearing loss.," Audiol. Med., **10**, 83-92.

The R Core Team. (**2017**). "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna Austria.

Perceptual evaluation of six hearing-aid processing strategies from the perspective of auditory profiling: Insights from the BEAR project

MENGFAN WU[1,*], RAUL SANCHEZ-LOPEZ[2], MOUHAMAD EL-HAJ-ALI[1], SILJE GRINI NIELSEN[2], MICHAL FERECZKOWSKI[1,2], TORSTEN DAU[2], SÉBASTIEN SANTURETTE[2,3], AND TOBIAS NEHER[1]

[1] *Institute of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark*

[2] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark*

[3] *Oticon A/S, Smørum, Denmark*

The current study forms part of the Better hEAring Rehabilitation (BEAR) project, which aims at developing new clinical tools for characterizing individual hearing loss and for assessing hearing-aid (HA) benefit. Its purpose was to investigate potential interactions between four auditory profiles and three measures of HA outcome obtained for six HA processing strategies. Measurements were carried out in a realistic noise environment at signal-to-noise ratios that were set based on individual aided speech reception thresholds (SRT_{50}). Speech recognition scores and ratings of overall quality and noise annoyance were collected in two spatial conditions. The stimuli were generated with the help of a HA simulator and presented via headphones to 60 older, habitual HA users who had previously been profiled based on a data-driven approach (Sanchez-Lopez *et al.*, 2019). The four auditory profiles differed significantly in terms of mean aided SRT_{50} and interacted significantly with the HA processing strategies for speech recognition in one spatial condition. Moreover, the correlation-pattern between the speech recognition scores and subjective ratings differed among the auditory profiles.

INTRODUCTION

Hearing-aid (HA) benefit in noisy environments is known to vary substantially among users, and several researchers have investigated ways to improve individual HA outcome (e.g., Lopez-Poveda *et al.*, 2017). Additionally, modern HA technology offers various features to improve speech intelligibility, e.g., directional microphones (Keidser *et al.*, 2013), noise reduction (Brons *et al.*, 2014), and dynamic range compression (Picou *et al.*, 2015). Despite these efforts, clinical HA fittings are still mainly based on the audiogram, even though pure-tone hearing thresholds are unable to capture all the supra-threshold deficits induced by a hearing loss (Johannesen *et al.*, 2016; Plomp, 1978). Moreover, the advanced features are not utilized in a systematic way.

*Corresponding author: awu@health.sdu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

The Better hEAring Rehabilitation (BEAR) project aims at developing new clinical tools for individual hearing loss characterization and HA benefit assessment. For that purpose, an auditory test battery and a data-driven approach for classifying listeners into four distinct auditory profiles were proposed in an earlier study (Sanchez-Lopez *et al.*, 2019). In that study, 75 participants from four auditory profiles differed in terms of their performance on various auditory measurements, as shown in Table 1. In the present study, 60 of the subjects tested by Sanchez-Lopez *et al.* (2019) participated and evaluated six processing strategies for HA treatment in three perceptual tasks.

The main purpose of the current study was to evaluate the perceptual HA outcomes of these six HA processing strategies in relation to the four auditory profiles. Furthermore, correlations between aided speech-in-noise intelligibility and the subjective ratings of overall quality and noise annoyance were analysed. Since a better speech recognition score with a given HA setting does not necessarily correspond to high preference for that HA setting (Cox *et al.*, 2016), we hypothesized that the four auditory profiles may help explain this inconsistency.

Auditory Profile	Audibility LF	Audibility HF	Binaural processing	Loudness	Speech perception	Spectro-temporal
A (n=14)	☺	😐	☺	☺	☺	☺
B (n=13)	☺	☹	😐	😐	☹	☹
C (n=20)	☹	☹	☹	☹	☹	☹
D (n=8)	☹	😐	☺	☹	☺	😐

Table 1: Overall relative performance on the main measures from the BEAR auditory test battery. LF = low frequencies, HF = high frequencies. ☺: better performance, ☹: poorer performance, and 😐: average performance.

METHODS

The perceptual evaluation was carried out in a simulated speech-in-noise environment and consisted of a speech recognition task and a subjective rating task. To achieve high face validity, testing conditions were chosen to reflect the difficulties that older HA users often encounter in complex noisy scenarios (Neher *et al.*, 2011; Prosser *et al.*, 1991).

Participants

Sixty subjects aged 60-80 years (mean = 70.8 years) were recruited for the study. Twenty-nine of them were tested at Odense University Hospital, Odense, while the other ones were tested at Bispebjerg Hospital, Copenhagen. All participants had bilateral symmetrical sensorineural hearing loss and were experienced HA users. The range of hearing loss configurations was chosen to lie in-between the N1 and N4 standard audiograms (Bisgaard *et al.*, 2010).

Prior to this study, all participants completed a comprehensive auditory test battery developed by Sanchez-Lopez *et al*. (2020). Based on these measurements, the participants were classified into one of four auditory profiles using a data-driven approach (Sanchez-Lopez *et al.*, 2019). Five of the participants tested here could not be reliably allocated to any of these profiles and were thus not included in the data analysis described here. The distribution of the remaining 55 participants was as shown in the first column of Table 1.

Test setup

The measurements were performed either in an anechoic chamber or a soundproof booth. Audio playback was via an RME Fireface UC soundcard, an SPL Phonitor Mini amplifier and a pair of Sennheiser HDA200 headphones. All stimuli were generated with the help of a hearing-aid simulator (HASIM) implemented in Matlab (Sanchez-Lopez *et al.*, 2018).

Stimuli

The target speech stimuli were DANTALE-II sentences spoken by a female native Danish speaker (Wagener *et al.*, 2003). The target speech was presented from either 0º (front) or 90º (the side of the 'better' ear according to previously conducted unaided speech-in-noise measurements). The background noise was a spatially diffuse cafeteria noise recorded in a university canteen with a pair of HA satellites. In addition, the International Speech Test Signal (Holube *et al.*, 2010) was used as a directional distractor from either 90º (target speech from 0º) or 0º (target speech from 90º). The directional distractor was presented at a signal-to-noise ratio (SNR) of +2 dB relative to the diffuse cafeteria noise.

Hearing-aid simulator (HASIM)

The HASIM included directional processing (omnidirectional, fixed cardioid or fixed binaural beamformer setting), noise reduction (maximal attenuation of 0, 5 or 15 dB) and amplitude compression (attack times of 5 or 250 ms and release times of 10 or 1250 ms for 'fast' and 'slow', respectively). For each listener, gains were set according to the NAL-NL2 fitting rule (Keidser *et al.*, 2011). Four HA processing strategies (Table 2) were selected to maximize differences in the sound processing. HA1 corresponded to very basic processing and served as a reference. HA6 resembled typical 'commercial' HA processing. For further details about the HASIM, see Sanchez-Lopez *et al*. (2018).

	Directional processing	Noise reduction	Amplitude compression
HA1	Omnidirectional	Off	Slow
HA2	Omnidirectional	Strong	Fast
HA3	Binaural beamformer	Off	Slow
HA4	Binaural beamformer	Strong	Slow
HA5	Binaural beamformer	Strong	Fast
HA6	Cardioid	Mild	Slow

Table 2: Description of the six tested HA processing strategies

Procedure

Each participant completed two visits. At the first visit, aided speech reception thresholds (SRT_{50}) were measured in an adaptive procedure (1-down 1-up procedure with a step size of 4 dB for the first five trials and 2 dB afterwards) to establish a baseline performance level for each participant. For the aided SRT_{50} measurements, the baselines of the stimuli were amplified according to individual gains (NAL-NL2 prescription for an input level of 65 dB SPL) and the target was amplified linearly during measurements. Aided SRT_{50} was only tested in the 0° condition. The six HA processing strategies were then evaluated for both spatial conditions using a speech recognition task at a fixed SNR that corresponded to the individual aided SRT_{50}. The speech recognition measurements were repeated at the second visit.

The subjective assessment included ratings of overall quality and noise annoyance for the six HA in two spatial conditions. A multi-stimulus comparison method with a hidden anchor ('MUSHA') was implemented in the SenseLabOnline 4.0.2 software (SenseLab, 2017). The anchor stimulus used for the subjective ratings was a speech-in-noise stimulus that had been heavily distorted using random binary mask processing to approximate undesired spectral distortion of the tested noise reduction scheme. On a given trial, participants were presented with a graphical user interface containing seven playback buttons and sliders (6 HA settings + 1 anchor stimulus). Each stimulus was rated four times per spatial condition. The test SNR used for the subjective ratings corresponded to SRT_{50} + 4 dB.

RESULTS

Effect of auditory profile on SRT_{50}

On average, profile A had the lowest SRT_{50} (mean = –0.5 dB SNR, SD = 1.2 dB SNR) while profile C had the highest (mean = 5.1 dB SNR, SD = 3.6 dB SNR). According to a series of independent *t*-tests, profile B (mean = 2.7 dB SNR, SD = 2.3 dB SNR) and profile C differed significantly from profile A and profile D (mean = 0.6 dB SNR, SD = 1.2 dB SNR), respectively (all $p < 0.01$).

Effects of auditory profile on HA outcomes

For both speech recognition (Figure 1) and the subjective ratings, listeners from the four auditory profiles showed similar patterns of benefit from the six HA processing strategies. More specifically, all auditory profiles gained larger benefits from the same or similar HA processing strategies for each outcome measure.

To assess the effect of auditory profile on the different HA outcomes, linear mixed effects models were implemented. The dependent variable was the individual standardized score. For speech recognition, due to the data being split based on spatial condition, the model included four components (HA, auditory profile (AP), HA*test SNR, HA*AP). The random effect was the individual intercept. For the subjective

ratings, the model included nine parts (HA, spatial condition (spa), AP, HA*spa, HA*AP, AP*spa, HA*test SNR, spa*test SNR, HA*spa*AP).

For all three outcomes, a significant effect of HA was found (all $p < 0.001$). For the subjective ratings, the effects of spa and HA*spa were also significant (all $p < 0.001$). Furthermore, for speech recognition assessed in the 90º spatial condition there was a significant interaction between AP and HA ($F_{9, 201} = 4.3, p < 0.001$), which was driven by low-benefit HA strategies (HA2 and HA3, see Fig. 1). Overall, there were no significant main effects of auditory profile or significant interaction with auditory profile (all $p > 0.05$).

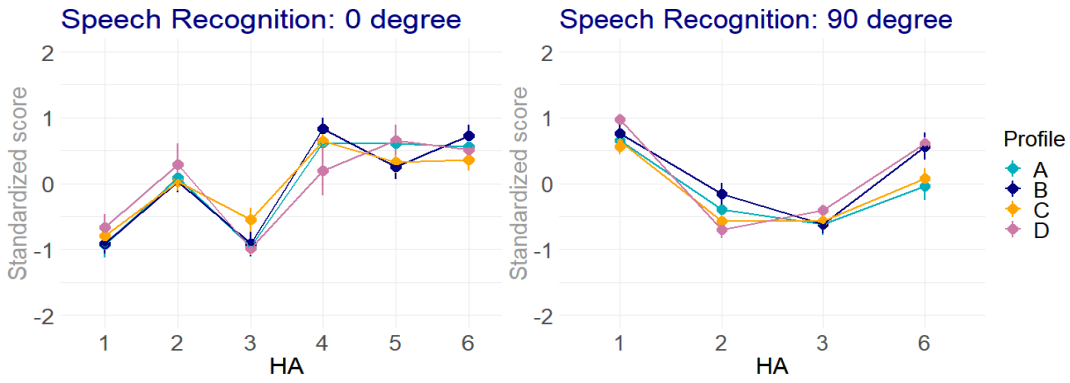


Fig. 1: Mean standardized speech recognition scores and standard errors for each test condition and auditory profile. Scores were averaged across test and retest. HA4 and HA5 were excluded in the 90º condition because of strong flooring effects.

Correlation analysis

Spearman's correlation coefficients were calculated to investigate potential relations between the three outcome measures across the four auditory profiles (Table 3). In general, more correlations were found for the 90º spatial condition than for the 0º spatial condition. In particular, the overall quality ratings were positively correlated with the speech scores for all auditory profiles in the 90º (but not the 0º) condition. Some differences among the four profiles were observed. Participants from profiles B showed relatively large, positive correlations between sentence recognition scores and both types of subjective ratings, while for profile A, which had a near-normal SRT_{50}, the different outcomes were not significantly correlated in most cases.

Profile		OVERALL & SPEECH		NOISE & SPEECH	
		0°	90°	0°	90°
A	*r*	0.07	**0.40**	0.02	0.17
	p	0.52	<0.01	0.88	0.22
B	*r*	**0.29**	**0.60**	**0.34**	**0.29**
	p	0.02	<0.001	<0.01	0.04
C	*r*	-0.01	**0.61**	**0.36**	**0.25**
	p	0.96	<0.001	<0.001	0.04
D	*r*	0.08	**0.71**	0.04	**0.57**
	p	0.60	<0.001	0.81	<0.01

Table 3. Results of correlation analyses performed on the speech scores and subjective ratings for each auditory profile. OVERALL = overall quality, SPEECH = speech recognition, NOISE = noise annoyance.

DISCUSSION

In the current study, speech recognition measurements and subjective ratings were applied to investigate potential links between four auditory profiles and response to six different HA processing strategies in a simulated speech-in-noise environment. Differences in aided SRT_{50} between four auditory profiles indicate different needs in terms of SNR improvement in HA processing. However, the four profiles barely differed in terms of their responses to the six tested HA processing strategies. One possible explanation could be that the participants were equated in terms of baseline performance level, which was based on their aided SRT_{50}. In other words, both the HASIM and the participants were exposed to different input signals.

Another potential explanation for the lack of differences among the four profiles could be that the acoustic scene contained only one type of noise. It is possible that the use of a multi-talker scenario or more fluctuating noises would elicit more pronounced differences among the profiles in terms of their ability to utilize spatial and temporal cues in such scenarios.

Moreover, in the present study, a limited set of HA settings were considered, with gains being prescribed according to the NAL-NL2 rule in all conditions. Previous research suggested that individuals with sloping audiograms obtain larger benefits from different HA amplification than individuals with flat audiograms (Keidser and Grant, 2001). Thus, it is possible that individuals from four auditory profiles obtain high HA benefit from different amplification rationales. Whether there is a three-way interaction between HA setting, amplification rationale and auditory profile in terms of perceptual HA outcome requires further study in the future.

The correlation analyses revealed that the four auditory profiles differed in terms of the extent to which speech recognition is related to overall quality and noise annoyance. For profile B, there were consistent positive correlations between the two types of measurements. This result might indicate that for profile B listeners HA preference is governed by the clarity or naturalness of the target speech. However, for profiles A and D, this was only the case in the 90° condition. Considering that these

two groups were tested at lower SNRs, it is reasonable to think that the HA processing strategies rendered the speech more unclear or distorted in this condition.

It is well established that HA benefit in complex speech-in-noise environments depends on both auditory and non-auditory factors (Gatehouse *et al.*, 2006). Our study suggests that preference for HA processing can be broken down into different types of psychoacoustic function. Whether those auditory factors are indeed linked to a general preference for speech naturalness requires further research. More generally, the question of whether the auditory profiles tested here influence HA outcome still needs further investigation. Ideally, this work should use real HAs, various background noises and aided outcome measures, and should also provide the participants with the possibility to acclimatize to the tested HA settings.

ACKNOWLEDGEMENTS

This work was supported by Innovation Fund Denmark Grand Solutions 5164-00011B (BEAR project), Oticon, GN Resound, Widex and other partners (Aalborg University, University of Southern Denmark, the Technical University of Denmark, Force, Aalborg, Odense and Copenhagen University Hospitals). The funding and collaboration of all partners is sincerely acknowledged. The authors sincerely thank Rikke Skovhøj Sørensen (Technical University of Denmark) and Christer P. Volk (SenseLabOnline, FORCE Technology) for their support. We also want to thank the participants and student helpers in this study.

REFERENCES

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (**2010**). "Standard audiograms for the IEC 60118-15 measurement procedure," Trends Amplif., **14**(2), 113-120. doi:10.1177/1084713810379609

Brons, I., Houben, R., and Dreschler, W. A. (**2014**). "Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners," Trends Hear., **18**, 2331216514553924.

Cox, R. M., Johnson, J. A., and Xu, J. (**2016**). "Impact of hearing aid technology on outcomes in daily life I: the patients' perspective," Ear Hearing, **37**(4), e224.

Gatehouse, S., Naylor, G., and Elberling, C. (**2006**). "Linear and nonlinear hearing aid fittings–1. Patterns of benefit," Int. J. Audiol., **45**(3), 130-152. doi:10.1080/14992020500429518

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). "Development and analysis of an international speech test signal (ISTS)," Int. J. Audiol., **49**(12), 891-903. doi:10.3109/14992027.2010.506889

Johannesen, P. T., Pérez-González, P., Kalluri, S., Blanco, J. L., and Lopez-Poveda, E. A. (**2016**). "The influence of cochlear mechanical dysfunction, temporal processing deficits, and age on the intelligibility of audible speech in noise for hearing-impaired listeners," Trends Hear, **20**, 2331216516641055.

Keidser, G., Dillon, H., Convery, E., and Mejia, J. (**2013**). "Factors influencing individual variation in perceptual directional microphone benefit," J. Am. Acad. Audiol., **24**(10), 955-968.

Keidser, G., Dillon, H., Flax, M., Ching, T., and Brewer, S. (**2011**). "The NAL-NL2 prescription procedure," Audiol. Res., **1**(1).

Keidser, G., and Grant, F. (**2001**). "Comparing loudness normalization (IHAFF) with speech intelligibility maximization (NAL-NL1) when implemented in a two-channel device," Ear Hearing, **22**(6), 501-515.

Lopez-Poveda, E. A., Johannesen, P. T., Perez-González, P., Blanco, J. L., Kalluri, S., and Edwards, B. (**2017**). "Predictors of hearing-aid outcomes," Trends Hear, **21**, 2331216517730526.

Neher, T., Laugesen, S., Søgaard Jensen, N., and Kragelund, L. (**2011**). "Can basic auditory and cognitive measures predict hearing-impaired listeners' localization and spatial speech recognition abilities?," J. Acoust. Soc. Am., **130**(3), 1542-1558. doi:10.1121/1.3608122

Picou, E. M., Marcrum, S. C., and Ricketts, T. A. (**2015**). "Evaluation of the effects of nonlinear frequency compression on speech recognition and sound quality for adults with mild to moderate hearing loss," Int. J. Audiol., **54**(3), 162-169. doi:10.3109/14992027.2014.961662

Plomp, R. (**1978**). "Auditory handicap of hearing impairment and the limited benefit of hearing aids," J. Acoust. Soc. Am., **63**(2), 533-549. doi: 10.1121/1.381753

Prosser, S., Turrini, M., and Arslan, E. (**1991**). "Effects of different noises on speech discrimination by the elderly," Acta Otolaryngol., **111**(sup476), 136-142.

Sanchez-Lopez, R., Fereczkowski, M., Bianchi, F., Piechowiak, T., Hau, O., Pedersen, M. S., Behrens, T., Neher, T., Dau, T. and Santurette, S. (**2018**). "Technical evaluation of hearing-aid fitting parameters for different auditory profiles," Euronoise 2018.

Sanchez-Lopez, R., Fereczkowski, M., Neher, T., Santurette, S., and Dau, T. (**2019**). "Robust auditory profiling: Improved data-driven method and profile definitions for better hearing rehabilitation," Poster presented at the 7th International Symposium on Auditory and Audiological Research, August 21 - 23 2019, Nyborg, Denmark. doi: 10.13140/RG.2.2.19762.35526

Sanchez-Lopez R, Nielsen. S. G,, El-Haj-Ali M, Bianchi F, Fereckzowski M, Cañete O, Wu M, Neher T, Dau, T and Santurette S. (**2020**). "Auditory tests for characterizing hearing deficits: The BEAR test battery," medRxiv. doi:10.1101/2020.02.17.20021949

SenseLab. (**2017**). SenseLabOnline (4.0.2 ed.). Hørsholm, Denmark: FORCE Technology.

Wagener, K., Josvassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol., **42**(1), 10-17. doi: 10.3109/14992020309056080

Assessing daily-life benefit from hearing aid noise management: SSQ12 vs. ecological momentary assessment

LINE STORM ANDERSEN[1], KLAUDIA EDINGER ANDERSSON[1], MENGFAN WU[1], NIELS PONTOPPIDAN[2], LARS BRAMSLØW[2], AND TOBIAS NEHER[1,*]

[1] *Institute of Clinical Research, University of Southern Denmark, Odense, Denmark*

[2] *Eriksholm Research Centre, Snekkersten, Denmark*

In audiological research, assessing daily-life benefit from hearing aid (HA) noise management (NM) is a challenge. While ecological momentary assessment (EMA) using smartphone-connected HAs has recently emerged as a promising tool for real-life data acquisition, there is a lack of research linking this method to established ones such as the SSQ12 questionnaire. In the current study, 12 hearing-impaired participants were asked to assess two HA fittings using a well-known questionnaire and a smartphone-based EMA method combining soundscape logging with momentary self-reports. The two HA fittings differed in terms of their NM settings (no NM vs. cardioid microphones and noise reduction activated). The participants were aged 23-75 years and had different occupations and lifestyles. The testing period for each fitting was 2 weeks. Overall, the EMA and SSQ12 scores were higher for the setting with NM activated, but this difference was only statistically significant in case of the SSQ12. The soundscape data showed that only few participants experienced noisy surroundings frequently. Future work on EMA-based HA assessment should therefore address the interplay between the tested HA features and the auditory ecology of the participants.

INTRODUCTION

For the assessment of daily-life experiences, questionnaires are widely used in both clinical and research settings and are generally of much value. Nevertheless, a well-known shortcoming of questionnaire-based assessments is the so-called memory bias (Schwarz, 2011). Memory bias describes how the human memory system compromises memory recollection, resulting in potentially imprecise data. More recently, ecological momentary assessment (EMA) has emerged as a promising alternative for subjective data acquisition. In contrast to the retrospective assessments performed with questionnaires, EMA is based on momentary assessments and thus avoids memory bias. In the field of audiology, EMA using smartphone-connected hearing aids (HAs) has gained much interest lately. In addition to the participants' self-reports provided via a smartphone app, the acoustic environments or 'soundscapes' can be logged by the HAs. This has made it possible to explore the acoustic environments that HA users typically encounter (Jensen and Nielsen, 2005), the efficacy of advanced HA features (Wu *et al*., 2018) or the way someone's sense

*Corresponding author: tneher@health.sdu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

of hearing loss is influenced by the regular assessment of daily-life listening experiences (Henry *et al*., 2012; Galvez *et al*., 2012).

The purpose of the current study was to compare EMA to an established method for daily-life HA assessment – the 12-item version of the Speech, Spatial and Qualities of Hearing scale (SSQ12) questionnaire (Noble *et al*., 2013). More specifically, we investigated (i) if EMA and the SSQ12 can document real-life benefit from a directional microphone setting combined with noise reduction, and (ii) if the results from the two types of assessments are in agreement with each other.

METHODS AND MATERIALS

Participants

Participants were recruited via social media groups for HA users, the hearing clinic at Odense University Hospital, private otologists and communication centres. The inclusion criteria were a minimum age of 18 years, mild-to-moderate, symmetric, bilateral, sensorineural hearing loss, a general aptitude to handle hearing aids and smartphones, and prior experience with smartphone use. A total of 13 participants were enrolled in the study, 12 of whom completed it. Table 1 provides an overview of their characteristics, while Figure 1 shows their audiograms. Hearing thresholds varied considerably, with pure-tone average hearing losses ranging from 20-54 dB HL (mean: 42 dB HL).

Participant	Age (yrs)	HA experience (yrs)	Occupation
1	47	41.0	Student
2	73	21.0	Retired
3	24	14.0	Employed
4	72	15.0	Retired
5	42	0.04	Maternity leave
6	75	6.0	Retired
7	23	18.0	Student
8	66	8.0	Retired
9	40	0.04	Employed
10	71	5.0	Retired
11	72	5.0	Retired
12	44	5.0	Employed

Table 1: Overview of the participants' characteristics.

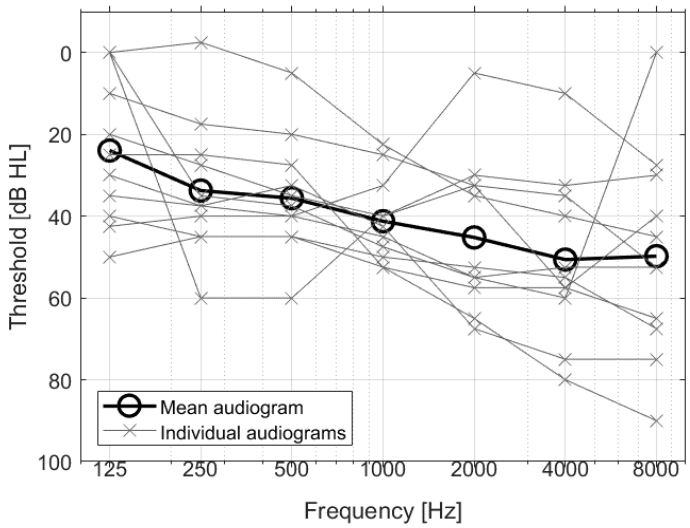


Fig. 1: Mean and individual audiograms averaged across left and right ears.

HA fittings

The HAs used were research prototypes (Oticon EVOTION mini-RITE). The devices classified the acoustic environment (or soundscape) continuously into one of four possible categories: Quiet, Speech, Noise or Speech in Noise. Two HA settings were tested: (i) Pinna-omni without noise reduction (NM_{OFF}), and (ii) fixed cardioid microphones with default noise reduction activated (NM_{ON}). The two settings were chosen with the aim of creating a clear acoustic contrast. The participants tested each HA setting for two weeks (see Study design).

SSQ12 assessments

The SSQ12 questionnaire consists of 12 items from the original SSQ questionnaire (Gatehouse and Noble, 2004). More specifically, there are five speech-related items, four items related to spatial hearing and three related to other qualities of hearing. For each item, a rating scale from 0-10 is used, with a higher score indicating a better outcome. In the current study, the participants were asked to complete a paper version of the SSQ12 twice, that is, once after each 2-week HA trial period.

EMAs

The EMAs were carried out using a custom smartphone app designed by Oticon A/S, which was installed on an iPhone SE device. The app prompted the participants to assess their listening experiences eight times a day. The prompts occurred randomly between 8 am and 8 pm. Since the study lasted for four weeks (see Study design), the participants were expected to complete 224 assessments to obtain a compliance score of 100%. The participants were able to carry out additional, voluntary assessments (resulting in a compliance score >100%). An assessment started with three questions (see Figure 2). The first two questions related to the overall listening experience (pleasant/good vs. unpleasant/bad) with the current HA setting, while the third

question enquired if the current assessment was related to speech understanding. If the participant reported this to be the case, four additional questions followed. These questions related to the ability to follow a conversation, the perceived difficulty in following a conversation and the experience of effort. For each of these questions a rating scale from 0-10 was used, with a higher score indicating a better outcome. The questions and rating scales were used as implemented in the app, that is, they were not adapted to correspond to the SSQ12 items.

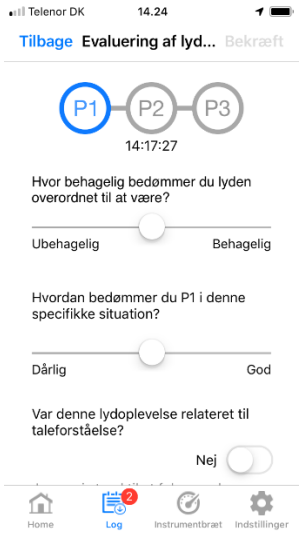


Fig. 2: Screen dump of the smartphone app showing the first three questions relating to the user's overall listening experience with the current HA program ('P1') and the relation to speech understanding.

Study design

The current study followed a single-blinded, balanced crossover design with a duration of 2 × 2 weeks. Figure 3 illustrates the general layout. The order in which the two HA settings were tested was counterbalanced across the participants.

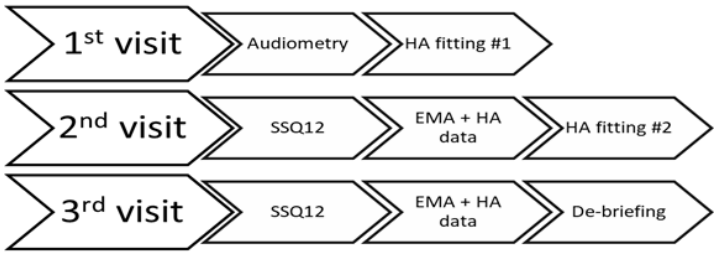


Fig. 3: Outline of study design.

Statistical analyses

The collected data were analysed using Microsoft Excel and IBM SPSS Statistics v25. Statistical significance was assessed using 2-tailed paired *t*-tests. The EMA scores

were pre-processed by taking the median across all ratings made by a given participant in a given soundscape (Quiet, Noise, Speech, Speech in Noise). Due to the design of the app, assessments reported as being unrelated to speech understanding contained data from the first three questions only whereas speech-related assessments contained data from all seven questions (see above). For the sake of brevity, the analyses below focus on the data from the speech-related assessments. Furthermore, they focus on overall SSQ12 or EMA scores rather than scores for individual items.

RESULTS

SSQ12 assessments

Table 2 summarises the SSQ12 data. Seven participants gave the NM_{ON} setting higher (better) SSQ12 ratings; the other participants rated the NM_{OFF} setting higher. Overall, the difference in mean scores for the two HA settings was statistically significant ($t_{11} = -2.9$, $p = 0.01$). At the individual-item level, the NM_{ON} setting scored higher on all but two SSQ12 items. However, the scores for the two HA settings did not differ from each other for any single SSQ12 item (all $p > 0.05$).

Participant	Mean score NM_{OFF}	Mean score NM_{ON}	Difference
1	5.5	6.0	0.5
2	5.0	4.5	-0.5
3	6.0	6.6	0.6
4	4.0	6.4	2.4
5	5.5	6.7	1.2
6	6.2	6.2	0.0
7	6.0	6.7	0.7
8	5.5	6.8	1.3
9	5.0	6.8	1.8
10	8.0	8.8	0.8
11	7.0	8.4	1.4
12	6.2	7.7	1.5
Average	**5.8**	**6.8**	**1.0**

Table 2: Mean SSQ12 scores for, and differences between, the NM_{OFF} and NM_{ON} settings for each participant and across them.

EMAs

In total, the participants provided 3140 EMAs, 1749 of which they reported as being related to speech understanding. Furthermore, 1398 assessments were made with the NM_{ON} setting engaged and 1742 assessments with the NM_{OFF} setting engaged. Given the requirement to complete at least eight assessments per day, the participants had to carry out a minimum of 224 assessments over the 2 × 2 week test period. From Table 3 it is apparent that compliance varied greatly across the 12 participants.

Participant	No. of EMAs	Compliance (%)
1	99	44
2	470	210
3	75	33
4	317	142
5	261	117
6	373	167
7	88	39
8	141	63
9	217	97
10	402	179
11	346	154
12	351	157
Total	**3140**	**117**

Table 3: Number of EMAs and compliance score for each participant and the group as a whole.

Overall, the EMA scores were 0.24 scale units higher for the NM_{ON} setting than for the NM_{OFF} setting. At the individual level, participants 1, 2, 7, 9 and 12 rated the NM_{ON} setting higher, whereas participants 4, 6 and 8 gave the NM_{OFF} setting slightly higher ratings. The difference in mean EMA scores between the two HA settings was not statistically significant ($t_{11} = 1.6$, $p > 0.1$). The same was true for all individual EMA items (all $p > 0.1$).

SSQ12 assessments vs. EMAs

On average, the speech-related EMA scores were higher than the SSQ12 scores (means: 8.4 vs. 5.4 scale units) and showed also more spread across participants (standard deviations: 1.7 vs. 1.0 scale units). The observed difference in mean scores for the two methods was statistically significant ($t_{11} = -2.6$, $p = 0.03$). In spite of these differences, the two datasets showed the same overall preference for a given HA setting for all but one participant.

Soundscape logging

Table 4 shows, for each participant, the number of EMAs made per soundscape category. As can be seen, only 10% of the data were collected under noisy conditions. When comparing the EMA scores for the NM_{ON} and NM_{OFF} settings as a function of soundscape, it was found that the NM_{ON} setting received higher scores in all but the quiet category, for which the NM_{OFF} setting dominated (data not shown). Moreover, EMAs classified by the HAs as containing speech were ~50% of the time *not* reported as being speech-related by the participants.

Participant	Quiet	Noise	Speech	Sp. in No.	Total
1	58	3	34	4	99
2	239	15	177	39	470
3	15	16	32	12	75
4	225	2	90	0	317
5	78	33	99	51	261
6	303	8	58	4	373
7	51	0	37	0	88
8	82	2	50	7	141
9	87	9	111	10	217
10	183	18	178	23	402
11	150	20	173	3	346
12	236	4	83	28	351
Total	**1707**	**130**	**1122**	**181**	**3140**
Percentage	**54%**	**4%**	**36%**	**6%**	**100%**

Table 4: Number of EMAs per participant for the Quiet, Noise, Speech and Speech in Noise (Sp. in No.) soundscapes as well as in total.

DISCUSSION

The current study explored (i) if EMA and the SSQ12 can document daily-life benefit from NM, and (ii) if the results from the two methods are in agreement with each other. Both assessment methods indicated a difference in mean scores between the two tested HA settings in favour of the NM_{ON} setting. However, this difference was only statistically significant in case of the SSQ12. Moreover, the EMA scores were generally higher than those obtained using the SSQ12, and they also showed more spread across participants. One possible explanation for these differences in outcome could be that the EMA items were used as implemented in the app (see above). Consequently, they were not identical to the SSQ12 items, which might be more suited for evaluating HA fittings (e.g. in terms of the formulations used to describe daily-life listening situations).

Another factor that could have influenced the outcomes of the current study was the heterogeneity of the participants. The participants varied considerably across several parameters such as audiometric configuration, age, occupation and HA experience. Humes *et al*. (2018) found that with increasing HA experience HA users spend more time in noisy environments. Comparing Tables 1 and 3, it is also apparent that the

older users tested here were more inclined to complete EMAs, which could potentially have biased the collected data in the direction of their auditory ecologies.

The shortage of assessments carried out in noisy conditions agrees with the findings of Jensen and Nielsen (2005) who reported that HA users spend most of their time at home or in conversation with less than three people. The general lack of noise in the participants' daily surroundings likely restricted the efficacy of the NM_{ON} setting, since directional microphones and noise reduction are meant to attenuate noise. In future work, it could be beneficial to target EMAs of this type of HA technology more specifically, e.g., by recruiting participants with specific auditory ecologies.

The discrepancies observed when comparing HA- and user-logged soundscapes were probably related to the fact that current HA soundscape classification algorithms are oblivious to the user's intent. Whether or not a speech signal is of relevance to the user depends on acoustic and non-acoustic factors, which can change over time. To achieve better correspondence between HA- and user-logged soundscapes, future EMAs would have to be able to capture user-related intentional aspects, too.

REFERENCES

Galvez, G., Turbin, M. B., Thielman, E. J., Istvan, J. A., Andrews, J. A., and Henry, J. A. (**2012**). "Feasibility of ecological momentary assessment of hearing difficulties encountered by hearing aid users," Ear. Hearing, **33**, 497-507, doi: 10.1097%2FAUD.0b013e3182498c41.

Gatehouse, S., and Noble, W. (**2004**). "The Speech, Spatial and Qualities of Hearing Scale (SSQ)," Int. J. Audiol., **43**, 85-99, doi: 10.1080/14992020400050014.

Henry, J. A., Galvez, G., Turbin, M. B., Thielman, E. J., McMillan, G. P., and Istvan, J. A. (**2012**). "Pilot study to evaluate ecological momentary assessment of tinnitus," Ear. Hearing, **33**, 179-290, doi: 10.1097%2FAUD.0b013e31822f6740.

Humes, L. E., Rogers, S. E., Main, A. K., and Kinney, D. L. (**2018**). "The acoustic environments in which older adults wear their hearing aids: Insights from datalogging sound environment classification," Am. J. Audiol., **27**, 594-603, doi: 10.1044/2018_AJA-18-0061.

Jensen, N., and Nielsen, C. (**2005**). "Auditory ecology in a group of experienced hearing-aid users: Can knowledge about hearing aid users' auditory ecology improve their rehabilitation?," Research Report, Eriksholm Research Centre.

Noble, W., Jensen, N. S., Naylor, G., Bhullar, N., and Akeroyd, M. A. (**2013**). "A short form of the Speech, Spatial and Qualities of Hearing scale suitable for clinical use: The SSQ12," Int. J. Audiol., **52**, doi: 10.3109/14992027.2013.781278.

Schwarz, N. (**2011**). "Why researchers should think "real-time": A cognitive rationale," in: M. R. Mehl & T. S. Conner (eds.), *Handbook of Research Methods for Studying Daily Life*, New York: Guilford.

Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., DeVries, S., and Oleson, J. (**2019**). "Efficacy and effectiveness of advanced hearing aid directional and noise reduction technologies for older adults with mild to moderate hearing loss," Ear. Hearing, **40**, 805-822, doi: 10.1097/AUD.0000000000000672.

Robust auditory profiling: Improved data-driven method and profile definitions for better hearing rehabilitation

RAUL SANCHEZ-LOPEZ[1,*], MICHAL FERECZKOWSKI[1,2], TOBIAS NEHER[2], SÉBASTIEN SANTURETTE[1,3], AND TORSTEN DAU[1]

[1] *Hearing Systems Section, Dept. of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark*

[2] *Institute of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark*

[3] *Oticon A/S, Smørum, Denmark*

Currently, the clinical characterization of hearing deficits for hearing-aid fitting is based on the pure-tone audiogram only. This relies on the assumption that the audiogram can predict performance in complex, supra-threshold tasks. Sanchez-Lopez *et al.* (2018) hypothesized that the hearing deficits of a given listener, both at threshold and supra-threshold levels, result from two independent types of auditory distortions. The authors performed a data-driven analysis of two large datasets with results from several tests, which led to the identification of four auditory profiles. However, the definition of the two types of distortion was challenged by differences between the two datasets in terms of the tests and listeners considered. In the Better hEAring Rehabilitation (BEAR) project, a new dataset was generated with the aim of overcoming these limitations. A heterogeneous group of listeners was tested using measures of speech intelligibility, loudness perception, binaural processing abilities and spectro-temporal resolution. As a consequence, the auditory profiles of Sanchez-Lopez *et al.* (2018) were refined. The updated auditory profiles, together with the investigation of optimal hearing-aid compensation strategies, are expected to form a solid basis for improved hearing-aid fitting.

INTRODUCTION

Hearing deficits are typically characterized by hearing loss severity as defined by the World Health Organization. The severity and shape are assessed based on the individual's sensitivity to pure tones, i.e., the audiogram. However, while the audiogram describes hearing thresholds, it does not reflect supra-threshold deficits. In general, such deficits are not addressed systematically in clinical practice. Nowadays, profiling has gained broad attention as a tool for typifying groups of observations (e.g., users, recordings or patients) that follow similar patterns. Data-driven profiling allows for uncovering complex and hidden structures in the data and has been used

*Corresponding author: rsalo@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

in psychology and audiology, for example in connection to hearing-aid features (Lansbergen *et al.*, 2020). Data-driven auditory profiling might thus help to identify groups of listeners that are defined by specific hearing deficits, as a basis for rehabilitative precision audiology.

Recently, Sanchez-Lopez *et al.* (2018) proposed a method for auditory profiling that was tested and verified by analyzing two datasets from previous studies. Their method is tailored to the hypothesis that a listener's hearing deficits can be characterized along two independent auditory distortion types, type-I and type-II. A normal-hearing listener is placed at the origin whereas other listeners, with a hearing loss that differs in the degree of the two types of distortion, are placed at different points in the two-dimensional space. While profile C represents a high degree of both distortions, profile B and D reflect hearing deficits dominated by one of the two distortions. Profile A, the group with a low degree of distortions, represents near-normal hearing. Each type of distortion would then covary with specific deficits observed in behavioral tasks, such as temporal processing or modulation sensitivity, that define a given auditory profile.

In Sanchez-Lopez *et al.* (2018), it was hypothesized that distortion type-I covaries with the audiometric thresholds, whereas distortion type-II is not related to audibility. However, the results of the analysis of two different datasets did not confirm this hypothesis. Distortion type-I was found to be connected to high-frequency hearing loss and reduced speech intelligibility in the analysis of both datasets. In contrast, distortion type-II was found to be linked to reduced binaural processing abilities in the case of one dataset (Thorup *et al.*, 2016) and to a low-frequency hearing loss in the case of the other dataset (Johannesen *et al.*, 2016). These mixed results were attributed to differences between the two datasets in terms of listeners and behavioral tests. It was concluded that there was a need for a new dataset that included listeners with more heterogeneous audiograms to better define the distortions and, thus, the auditory profiles. Furthermore, the tests should investigate several aspects of auditory processing while being clinically feasible. In the Better hEAring Rehabilitation (BEAR) project, a new dataset was therefore generated with the aim of overcoming these limitations. The considered tests were chosen based on a literature review.

The resulting dataset includes a large and heterogeneous group of seventy-five listeners that were tested in a clinical environment in several behavioral tasks divided into five domains: audibility, loudness perception, binaural processing abilities, speech perception and spectro-temporal processing. In the present study, the analysis of the new dataset did not aim to disentangle the effects of audibility and supra-theshold deficits but to identify four clinically relevant patient subpopulations by refining the definition of the auditory profiles and the two types of distortions.

METHOD

Description of the dataset

Seventy-five listeners participated in the study. Seventy participants presented various degrees and shapes of symmetrical sensorineural hearing loss and five had near-normal audiometric thresholds. The participants were recruited from clinical databases at Odense University Hospital (OUH) and Bispebjerg Hospital (BBH). All listeners completed the BEAR test battery described in Sanchez-Lopez *et al.* (2019). The selected tests have shown potential for auditory profiling and their outcomes[†] may be informative for hearing-aid fitting. Table 1 summarizes the chosen tests.

Name of the test	Category	Variables[†]
A. Pure-tone audiometry	Audibility	AUD_xF
B. Fixed level frequency threshold		FLFT
C. Word recognition scores	Speech Perception	SRT_Q, maxDS
D. Hearing in noise test		SRT_N, SS^{4dB}
E. Maximum frequency for IPD detection	Binaural processing abilities	IPD_{fmax}
F. Binaural pitch		BP20
G. Extended binaural audiometry in noise		BMR
H. Adaptive categorical loudness scaling	Loudness perception	HTL_{xF}, MCL_{xF} $DynR_{xF}$, $Slope_{xF}$
I. Fast spectro-temporal modulation sensitivity	Spectro-temporal processing	$sSTM_8$, $fSTM_8$ TiN_{xF}
J. Extended audiometry in noise		SMR_{xF}, TMR_{xF}.

Table 1: Overview of the tests used for the collection of the BEAR dataset. The tests are divided by categories, and the outcome variables are presented in the right column. When a variable is frequency specific, the sufix $_{xF}$ follows the variable. It can be LF for lower frequencies or HF for higher frequencies.

The results of the tests were processed in a similar fashion as in Sanchez-Lopez *et al.* (2018). First, for each of the tests, the outcome measures of interest were extracted from the raw results. When the tests had frequency specific values, the results were grouped into low ($\leq$ 1 kHz) and high ($>$ 1 kHz) frequency averages. In case of monaural examination, the mean of the two ears was used. The variables with more than 12 missing data points were excluded as well as listeners with more than two missing outcome measures. The data were then normalized between the 25th and 75th percentiles. In total, 26 variables were selected from the outcome measures, as shown in Table 1.

Unsupervised learning

The data-driven analysis used here is based on the unsupervised learning stage of the method described in Sanchez-Lopez *et al.* (2018). This is divided into three

[†]The outcome variables are further explained in Sanchez-Lopez *et al.* (2019)

steps: I) Dimensionality reduction: based on principal component analysis (PCA), a subset of variables highly correlated with the principal components PC1 and PC2 was kept for the following steps; II) Archetypal analysis: the data were decomposed into two matrices, the test matrix, which contained the extreme patterns of the data (archetypes), and the subject matrix, which contained the weights of each archetype. A given subject can then be represented as a convex combination of the archetypes. The analysis was limited to four archetypes; III) Profile identification: the subject matrix was used to estimate the distance between observations and the four archetypes. Each listener was then assigned to an auditory profile based on the nearest archetype.

In the present study, the method for profile identification was refined according to the following modifications: 1) Iterative bagging: Since a data-driven analysis can be highly influenced by the data themselves, here, the data were decimated randomly in terms of listeners and tests. Then, one thousand iterations of the steps I, II and III were performed with only 83% of the data (69 listeners and 24 variables); 2) Dimensionality reduction: Since the use of several correlated variables in PCA can bias the results, the dimensionality reduction was refined by removing highly correlated variables. If two variables resulting from step I were highly correlated (Pearson's $r > 0.85$), one of them was dropped and this step repeated. The number of selected variables (6, 8 or 10) was also randomly chosen in each iteration; 3) Profile identification: Instead of using the criterion of the nearest archetype, listeners with a weight larger than 0.5 for one of the four archetypes were identified as belonging to that auditory profile. Otherwise, they were left "unidentified".

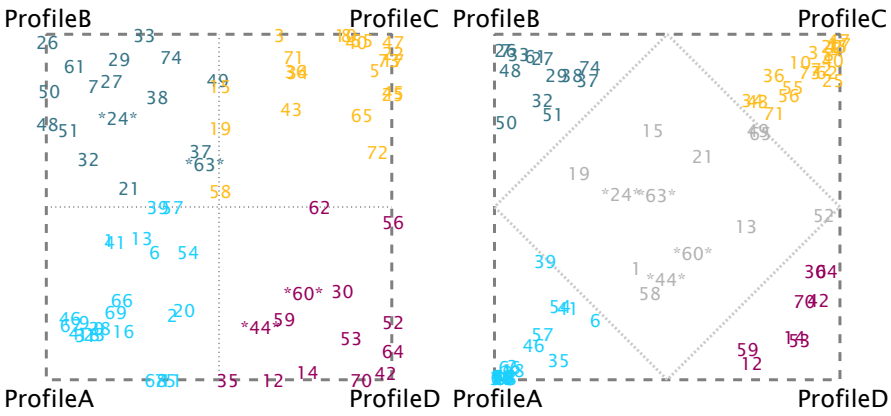


Fig. 1: Square representation result of the data-driven method. Left panel: results from the Sanchez-Lopez *et al.* (2018) method where the listeners are represented based on the subject matrix of the archetypal analysis. Right panel: results from the refined method where the listeners' representation is based on the probabilities. The listeners indicated by grey squares show a low probability of belonging to any of the four profiles ($P < 0.5$) and listeners labelled as $*Number*$ show a high probability ($P > 0.5$) of being "unidentified".

The final output of the refined method is the probability of being identified as belonging to an auditory profile. This was estimated by the fraction of times a listener was assigned to a profile after the 1000 iterations. The probability of being "unidentified" was also calculated.

RESULTS

The left panel in Figure 1 shows the results of the data-driven auditory profiling as used in Sanchez-Lopez *et al.* (2018). Listeners are located in the square representation based on the results of the archetypal analysis. The representation does not show clear clusters and the listeners identified as belonging to an auditory profile may have been wrongly identified. The right panel shows the results of the refined method where the listeners are located in the two-dimensional space based on the probability of belonging to any of the four auditory profiles. Listeners located close to a corner exhibited a high probability of belonging to that profile and four clear clusters can now be observed. The "unidentified" listeners are placed in between the four quadrants.

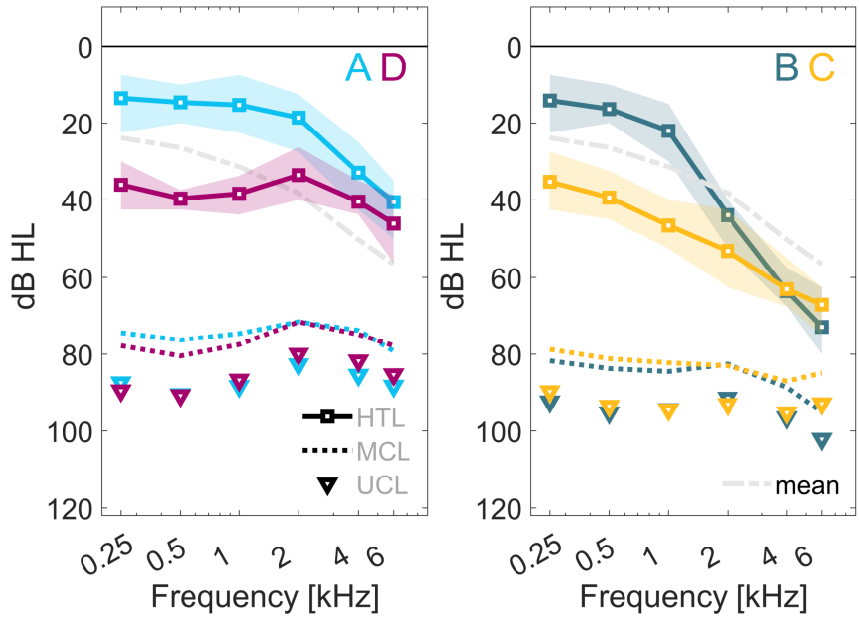


Fig. 2: Results of the ACALOS measurements presented in terms of hearing thresholds (HTL), most comfortable level (MCL) and uncomfortable levels (UCL). Results are divided by profiles (solid lines) and averaged across the whole dataset (dashed line).

Figure 2 shows average hearing threshold levels (HTL) as well as the corresponding interquartile ranges, most comfortable levels (MCL) and uncomfortable levels (UCL) ($\bigtriangledown$). Profiles A and D (left panel) exhibit a mild-to-moderate high-frequency hearing loss with hearing levels below the average high-frequency hearing loss (dashed line). The difference between profiles A and D is at low frequencies where profile D exhibits

thresholds that are more elevated than the average values ($\gtrsim$ 30 dB HL). Profiles B and C (right panel) also exhibit a high-frequency hearing loss but with values above the average ($\gtrsim$ 45 dB HL). The difference between profiles B and C is mainly in terms of the low-frequency hearing thresholds where profile C exhibits a low-frequency hearing loss that is above the average values.

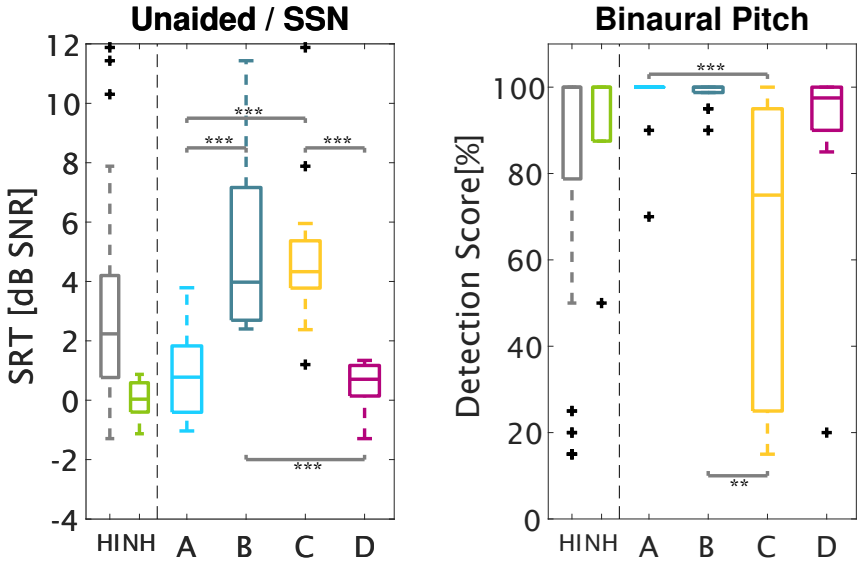


Fig. 3: Boxplots of the results from suprathreshold measures. Left: Unaided speech intelligibility in speech-shaped noise (SSN). Right: Binaural pitch detection. In each panel, the average data of the whole group of the hearing-impaired (HI) listeners and the normal-hearing (NH) listeners are shown on the left, whereas the results for the sub-populations according to the profiles are shown on the right.

Figure 3 (left panel) shows unaided speech reception thresholds (SRT) measured for sentences presented in speech-shaped noise. The median SRT for the hearing-impaired (HI) listeners data was at about 2.5 dB signal-to-noise ratio (SNR). Profiles A and D showed a better performance (median: $\approx$ 1 dB SNR) than profiles B and C (median: $\approx$ 4-5 dB SNR).

The right panel of Figure 3 shows the results from the binaural pitch detection test. The detection scores for listeners with profiles A, B and D were significantly higher (better performance) than for the listeners with profile C. Some profile C listeners were not able to detect the dichotic pitch percept, whereas they could detect the diotic pitch.

DISCUSSION

The refined method showed four clusters of listeners with significant differences in terms of several supra-threshold tasks. In the previous study of Sanchez-Lopez *et al.* (2018) the data-driven analysis of the Thorup *et al.* (2016) data set showed that distortion type-II was associated with binaural processing abilities in listeners with near-normal or mild-moderate high-frequency hearing loss. The binaural pitch test was found to be the most relevant prediction in that study and the listeners in profiles D_T* and C_T showed scores lower than 95 %. In the present study, only listeners with profile C showed such a behaviour. In contrast, the analysis of the Johannesen *et al.* (2016) data set in Sanchez-Lopez *et al.* (2018) showed that distortion type-II was associated with outer hair cell loss at low frequencies and their participants had moderate-to-severe hearing loss. Profiles D_J† and C_J showed higher audiometric thresholds and a loss of cochlear non-linearity at low frequencies. In the present study, the results are in better agreement with the analysis performed on the Johannesen *et al.* (2016) data set than on the Thorup *et al.* (2016) data set. This suggests that the use of data from a representative sample of different degrees of hearing loss and a normal hearing reference was crucial for a robust profile-based hearing-loss characterization.

The audiometric thresholds corresponding to the four robust auditory profiles were significantly different in terms of the degree and shape of the hearing loss. Interestingly, the four audiometric profiles look similar to the audiometric phenotypes of Dubno *et al.* (2013). According to this view, the flat "attenuation" observed in profile D may be ascribed to a metabolic hearing loss (endocochlear potential loss) and the sloping hearing loss observed in profile B may be associated with a sensory loss. Metabolic hearing loss yields flat elevated thresholds but does not affect speech-intelligibility in noise (Pauler *et al.*, 1986), which is consistent with the results of the present study. Furthermore, according to Plomp's model, profiles B and C exhibited a distortion component because of their elevated speech reception thresholds in noise (Plomp, 1978). (Wu *et al.*, 2020) carried out an accompanying study assessing aided speech-in-noise performance with the same test participants. Their results also showed elevated SRT_N for the listeners with profiles B and C suggesting that amplification did not fully compensate for their hearing deficits.

CONCLUSION

Based on a refined data-driven method and a new dataset, a solid definition of the auditory profiles was obtained. The different profiles showed significant differences in terms of low- and high-frequency hearing loss, speech-in-noise intelligibility, binaural processing abilities and spectro-temporal modulation sensitivity. Overall, the results of the present study suggest that the shape and degree of sensitivity loss can be

*The subindex T refers to the profiles from the analysis of Thorup *et al.* (2016)

†The subindex J refers to the profiles from the analysis of Johannesen *et al.* (2016)

a consequence of specific impairment mechanisms with associated supra-threshold deficits. Moreover, stratifying the listeners in clinically relevant subgroups has potential for further investigating the independence of sensitivity vs supra-threshold deficits as well as physiological correlates of the perceptual auditory distortions.

ACKNOWLEDGEMENTS

We want to thank S. G. Nielsen, M. El-Haj-Ali, M. Wu and O. Cañete for their collaboration in the data collection. This work was supported by Innovation Fund Denmark Grand Solutions 5164-00011B (BEAR project), Oticon, GN Resound, Widex and other partners. The funding and collaboration of all partners are sincerely acknowledged.

REFERENCES

Dubno, J. R., Eckert, M. A., Lee, F. S., Matthews, L. J., and Schmiedt, R. A. (**2013**). "Classifying human audiometric phenotypes of age-related hearing loss from animal models," J. Assoc. Res. Oto., **14**(5), 687-701. doi: 10.1007/s10162-013-0396-x.

Johannesen, P. T., Pérez-González, P., Kalluri, S., Blanco, J. L., and Lopez-Poveda, E. A. (**2016**). "The influence of cochlear mechanical dysfunction, temporal processing deficits, and age on the intelligibility of audible speech in noise for hearing-impaired listeners," Trends Hear., **20**(0). doi: 10.1177/2331216516641055.

Lansbergen, S. and Dreschler, W. A. (**2020**). "Hearing aid feature profiles: the success of rehabilitation," Proc. ISAAR, **7**, 229-236.

Pauler, M., Schuknecht, H. F., and Thornton, A. R. (**1986**). "Correlative studies of cochlear neuronal loss with speech discrimination and pure-tone thresholds," Arch. Otorhinolaryngol., **243**(3), 200-206. doi: 10.1007/bf00470622.

Plomp, R. (**1978**). "Auditory handicap of hearing impairment and the limited benefit of hearing aids," J. Acoust. Soc. Am., **63**(2), 533-549. doi: 10.1121/1.381753.

Sanchez-Lopez, R., Bianchi, F., Fereczkowski, M., Santurette, S., and Dau, T. (**2018**) "Data-Driven Approach for Auditory Profiling and Characterization of Individual Hearing Loss," Trends Hear., **22**. doi: 10.1177/2331216518807400.

Sanchez-Lopez, R., Nielsen, S. G. ,..., and Santurette, S. (**2019**). "Data for: "Auditory tests for characterizing hearing deficits: The BEAR test battery"," (Version v1.0). Zenodo. doi: 10.5281/zenodo.3459580.

Thorup, N., Santurette, S., Jørgensen, S., Kjærbøl, E., Dau, T., and Friis, M. (**2016**). "Auditory profiling and hearing-aid satisfaction in hearing-aid candidates," Dan. Med. J., **63**(10), A5275.

Wu, M., Sanchez-Lopez, R., ..., and Neher, T. (**2020**). "Perceptual evaluation of six hearing-aid processing strategies from the perspective of auditory profiling: Insights from the BEAR project," Proc. ISAAR, **7**, 265-272.

Subjective loudness ratings of vehicle noise with the hearing aid fitting methods NAL-NL2 and trueLOUDNESS

DIRK OETTING[1,3,*], JÖRG-HENDRIK BACH[1,2,3], MELANIE KRUEGER[1,2,3], MATTHIAS VORMANN[2,3], MICHAEL SCHULTE[2,3], AND MARKUS MEIS[1,2,3]

[1] *HörTech gGmbH, D-26129 Oldenburg, Germany*

[2] *Hörzentrum Oldenburg GmbH, D-26129 Oldenburg, Germany*

[3] *Cluster of Excellence Hearing4all, D-26129 Oldenburg, Germany*

Subjects with similar hearing thresholds showed large differences in loudness summation of binaural broadband signals after narrowband loudness compensation. Based on these findings, the fitting method trueLOUDNESS was developed to restore the individual binaural broadband loudness perception. In the present study, the trueLOUDNESS fitting method was compared with NAL-NL2. Loudness judgements of different vehicles' sounds were compared with average judgements by normal-hearing subjects. The loudness judgements with trueLOUDNESS fittings were closer to normal compared to the loudness judgements with NAL-NL2 fittings. This study shows that the lab measurement of binaural broadband loudness perception has validity beyond the laboratory.

INTRODUCTION

Individual loudness perception plays an important role in the fitting of hearing aids. The EuroTrak survey in Germany (EuroTrak DE, 2018) showed that the dimension "comfort with loud sounds" was the most important criterion for the overall satisfaction with hearing aids. In subjects with similar hearing thresholds the loudness summation of binaural broadband signals can differ substantially from the average. The effect has been well described (Oetting et al. 2016, 2018), and the amount of available data for the individual binaural broadband loudness summation of hearing-impaired (HI) subjects continue to increase. Based on the findings concerning the individual differences of binaural broadband loudness perception, the hearing-aid fitting method, trueLOUDNESS, was developed. The fitting rationale of trueLOUDNESS is to restore the binaural broadband loudness perception. The categorical loudness scaling measurements required for the trueLOUDNESS fitting are performed over headphones in the lab. The aim of this study was to show that the lab measurements of loudness scaling are related to real-world loudness perception with hearing aids. Therefore, a user group ("power users") for whom the trueLOUDNESS procedure predicts higher gains than NAL-NL2 was selected. Consequently, we expected that the loudness perception with an NAL-NL2 fitting would be lower than normal in this group. For a "sensitive user" group for whom the trueLOUDNESS procedure predicts lower gains than NAL-NL2, we expected that perceived loudness would be higher than normal with NAL-NL2 because NAL-NL2

*Corresponding author: d.oetting@hoertech.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

gains are above the gains required for restoring the loudness perception according to trueLOUDNESS.

METHOD

Screening

The subjects were selected from a pool of N = 129 subjects for whom the trueLOUDNESS gains from screening measurements were available. The subjects were older adults with an average age of 74 years (standard deviation: 6.5 years) with an averaged pure tone average (PTA) at 500, 1000, 2000 and 4000 Hz of 46.3 dB HL (range: 23.8 to 80.0 dB HL). Slightly more men (56%) than women were in the data set. Sixty-three procent of the subjects were hearing aid wearers. The average trueLOUDNESS gain at 500, 1000, 2000 and 4000 Hz was divided by the PTA for each ear. The values for the left and the right ear were averaged, resulting in a single-number value (average relative gain) for each subject. Fig. 1a shows for all 129 subjects the relative trueLOUDNESS gains over the PTA. The data show a large individual variation in terms of relative gains necessary to restore the loudness for a binaural broadband speech signal at 65 dB SPL even for subjects with similar PTAs.

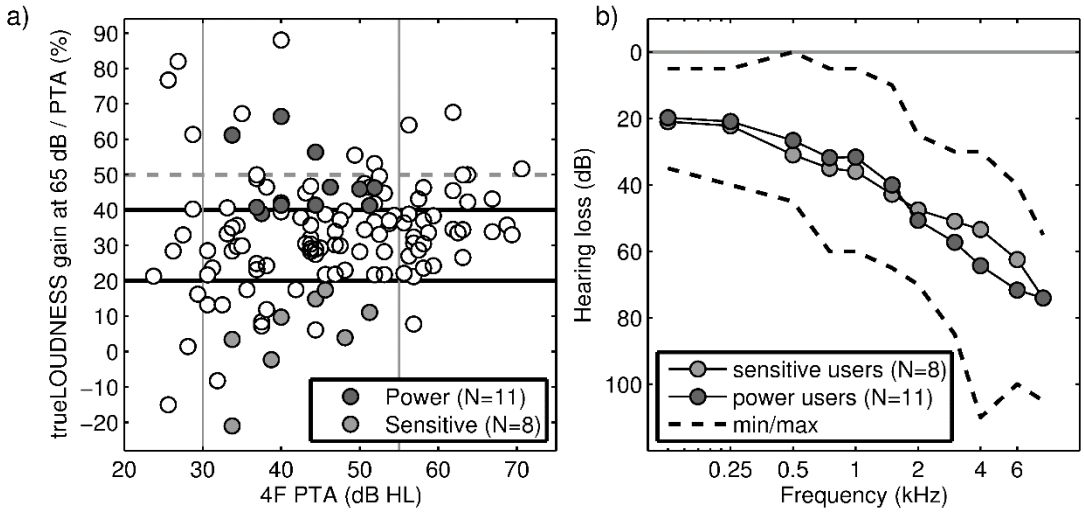


Fig. 1: a) Relative gain according to trueLOUDNESS for a binaural speech signal of 65 dB SPL for N=129 subjects. The average trueLOUDNESS gains at the PTA frequencies were divided by the averaged PTA for both ears. The grey dashed line indicates the half-gain rule at 50% relative gain. Subjects with relative gains below 20% are referred to as "sensitive users", whereas "power users" have more than 40% relative gain. b) Averaged audiograms of the sensitive and power users that participated in this study. The dashed lines indicate the minimum and maximum values of all HI subjects.

For example, subjects with a PTA of around 40 dB HL showed relative gains between 0% and 90%. A value of 0% relative gain leads to an average of 0 dB gain at the four frequencies to restore binaural broadband loudness perception, and a value of 90%

relative gain leads to an average gain of 36 dB at the four frequencies. The difference of the relative gains between the left and the right ear were small, with values below 13.2 percentage points for 90% of the subjects. Five subjects showed differences of the relative gains for the left and the right ear above 20 percentage points, mostly due to asymmetric hearing loss.

Subjects

The subjects that participated in this study were selected to match the following inclusion criteria: 1) PTA averaged across both ears between 30 and 55 dB HL; 2) relative trueLOUDNESS gain at 65 dB SPL averaged across both ears below 20% or above 40%, respectively; 3) absolute difference of the PTA between the left and the right ear below 15 dB.

Subjects were selected with the expectation that for them gain prescription would differ between trueLOUDNESS and NAL-NL2. Relative gains for the standardized audiograms by Bisgaard et al. (2011) with PTAs between 30 and 55 dB HL (first inclusion criteria) were analysed. The relative gains of these audiograms were between 24% and 29%. Therefore, subjects with relative gains between 20% and 40% were excluded from this study. Subjects with relative trueLOUDNESS gains below 20% were assigned to a group referred to as “sensitive users” and subjects with relative trueLOUDNESS gains above 40% were assigned to a group referred to as “power users”.

Eight sensitive users and 10 power users participated in this experiment. Details of both groups are shown in Table 1.

	Gender (f: female, m: male)	Age in years: mean (std.)	PTA in dB HL: mean (std.)	Experience with hearing aids
Sensitive users (N = 8)	5 f, 3 m	75.1 (3.5)	42.0 (6.5)	4 exp, 4 new
Power users (N = 10)	5 f, 5 m	74.7 (2.9)	42.6 (6.1)	6 exp, 4 new
Normal-hearing subjects	9 f, 10 m	50.7 (19.2)	5.2 (4.6)	–

Table 1: Comparison of the details of the subjects in the sensitive user group (below 20% relative gain, cf. Fig. 1) and the power user group (above 40% relative gain, cf. Fig. 1). The mean and standard deviation in age and PTA as well as the gender and experience with hearing aids were similar in both groups.

The maximum difference between the left and right relative trueLOUDNESS gains was 12.1 percentage points for the 18 subjects, indicating symmetric hearing losses. The PTA of the better ear was between 31.3 and 51.3 dB HL, which corresponds to

the WHO criterion for hearing loss of grade 1 (slight impairment) and 2 (moderate impairment), respectively. The average hearing loss for both groups is shown in Fig. 1b. A total of 19 normal-hearing (NH) subjects (9 female, 10 male) participated as a reference group in this experiment. Their better-ear PTA was below 25 dB, which corresponds to the WHO criterion for grade 0 (no impairment). All participants were recruited through the database of Hörzentrum Oldenburg.

Location, vehicles and driving actions

The experiment took place at a former military airport in Oldenburg, where a set of roads was chosen to conduct the experiments with real vehicles and minimum disturbance. Four different vehicles were used: a car (Opel Corsa, 2016), a motorbike (Suzuki VX 800 800cc, 1994), a van (Fort Transit FT100, 1999) and a street sweeper (Kärcher MC 50).

The driving instructions for the first three vehicles were: idling (standing still with the engine on), accelerating, passing by at 30 km/h, passing by at 50 km/h, and breaking until standing still. The street sweeper's actions were standing by, standing by with the brushes switched on, and moving forward with the brushes switched on. Each driving action was repeated twice, once with the vehicle driving on the nearby side of the street with ~3 meters distance to the subjects and once with the vehicle driving on the far side of the street with ~6 meters distance to the subjects. Ten actions per vehicle and 6 actions for the street sweeper were conducted, resulting in 36 test conditions. Details on the driving actions including the recorded sound pressure levels are reported in Llorach et al. (2019). The subjects were placed parallel to the road. Seven positions were available on the left side of a central recording station and seven places on the right side.

Fitting of hearing aids

The HI subjects were equipped with Phonak Audéo B90-312 hearing aids. For the acoustic coupling individual ear moulds (cShells, if available) or domes were used. The type of dome (open, closed, or power dome) was selected according to the recommendation by the Phonak fitting software. Two different programs in the hearing aids were set up. In program 1 the gains were adjusted according to the trueLOUDNESS gain calculations for 50, 65, and 80 dB SPL. The trueLOUDNESS target gain functions were visible in the fitting software. The gains were adjusted manually by an acoustician to achieve a close match between the target trueLOUDNESS functions and the gain functions of the hearing aid. In program 2 the fitting method NAL-NL2 was selected as the gain calculation method in the fitting software. The gender and experience-specific differences were considered during fitting, and the gain level was adjusted to 100% target gain.

Procedure

The subjects reported for each condition their loudness perception on a printed scale of the categorical loudness scaling procedure from "not heard" to "too loud" (Brand and Hohmann, 2002). The vehicle's starting point was in front of the recording

devices. The experimenter showed numbers to indicate the current action. The subjects were asked to rate the loudness and annoyance of that action. Each run of a vehicle took between 7 to 8 minutes to complete all driving actions. The vehicle order was car, motorbike, van, street sweeper. First, program 1 (trueLOUDNESS) was tested with all cars and all driving actions; then program 2 (NAL-NL2) was tested. A break was included between the programs. NH subjects participated and performed the loudness and annoyance ratings at the same time as the HI subjects. Each session lasted in total around 1.5 h. Overall, four sessions on two different days were conducted. On the first day, sessions 1–3 were conducted, and on the second day, session 4 was conducted. The participants per session are shown in Table 2. Calibrated sound recordings have been made to ensure similar test conditions across all sessions, retrospectively. An ethics proposal was approved by the ethical commission of the University of Oldenburg.

Session:	1	2	3	4	Total
HI “sensitive users”	5	1	0	2	8
HI “power users”	1	1	4	4	10
NH	0	10	5	4	19

Table 2: Participants of the different groups over the different sessions of the experiment. Four sessions were conducted. Session 1, 2 and 3 took place on the same day. Session 4 was conducted about one month later.

RESULTS

To compare the NH and HI results, the median loudness responses from the categorical loudness scale in categorical units (CU) were calculated. For the NH subjects, the sensitive users and the power users, the median value for each of the 36 conditions was calculated. Each condition was assigned to the NH loudness category between 0 and 50 CU in 5 CU steps that was nearest to the median NH value. For each category, the median of the assigned condition was calculated for the sensitive and power users with the NAL-NL2 and the trueLOUDNESS fitting. The results are shown in Fig. 2. The abscissa shows the NH loudness category along with the number of conditions assigned to that category. E.g., there were 13 conditions that resulted in a median NH loudness rating of 25 CU (“medium loud”) and 2 conditions that resulted in a median NH loudness rating of 45 CU (“very loud”). If the resulting loudness ratings with hearing aids corresponded to the NH loudness ratings within an error margin of ±2.5 CU, the values would be inside the white diagonal corridor. The resulting median values for the sensitive users with NAL-NL2 fitting are above the diagonal line. For the power users with NAL-NL2 fitting, the median loudness values are below the diagonal line for loudness categories of 30 CU and above. The median values for the sensitive and the power groups with the trueLOUDNESS fitting are close to the diagonal line. That means that for the power users, the trueLOUDNESS fitting predicts higher gains than NAL-NL2, which shifts the resulting loudness ratings towards the diagonal line. For the sensitive users, the trueLOUDNESS fitting

predicts less gain than NAL-NL2, which again shifts the resulting loudness ratings closer to the diagonal line.

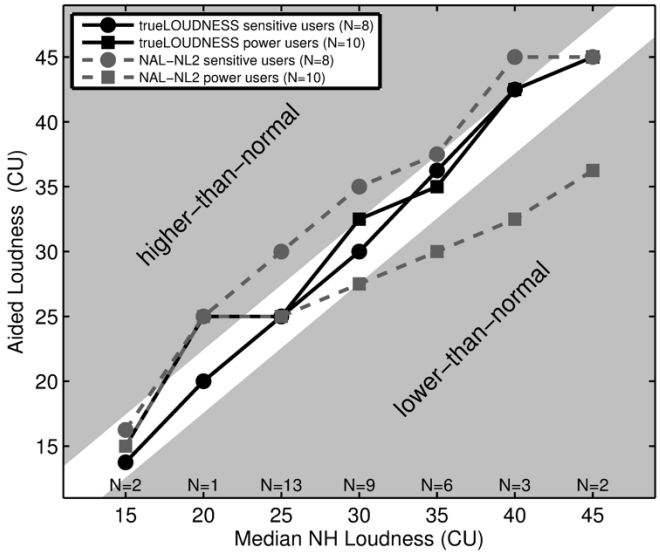


Fig. 2: The test conditions were pooled according to the median NH loudness rating and assigned to the category closest on the loudness scale from 0 to 50 CU in 5 CU steps. The four lines show the median results of the sensitive and power users after fitting with trueLOUDNESS and with NAL-NL2.

For the statistical analysis of the data, the loudness ratings were transformed to the phon scale (interval scale) using the transformation by Heeren et al. (2013) and an interpolation of the tabulated values according to ANSI S3.4 (2007). The difference between the power and sensitive users, and the average NH value was calculated and is referred to as *mean Δloudness re. NH*. These differences for each condition indicate the deviation from the mean NH rating. Fig. 3a shows the *mean Δloudness re. NH* over the *mean NH loudness value in phons* for the trueLOUDNESS fitting.

The conditions with values around 40 phons were the idle conditions of the car (near and far side) rated by the NH subjects on average with 37.6 and 40.1 phons. The regression lines show on average a slightly increased loudness rating in the power users for the trueLOUDNESS fitting. For NH mean values above 75 phons, the deviation according to the regression line was less than 3 phons in both HI groups. The results for the NAL-NL2 fitting are shown in Fig. 3b. The regression line for the sensitive users (grey crosses) show constantly higher loudness ratings for all conditions from low to high NH mean values. The power users with the NAL-NL2 fitting showed higher loudness ratings from 2 phons for the conditions with low NH mean values and −8.8 phons for the conditions with high NH mean values. The *mean Δloudness re. NH* for each group and each fitting were pooled and are shown by the boxplots in Fig. 3c. Whiskers mark values within 1.5 interquartile ranges of the first and third quartile, respectively. Outliers are marked with a cross.

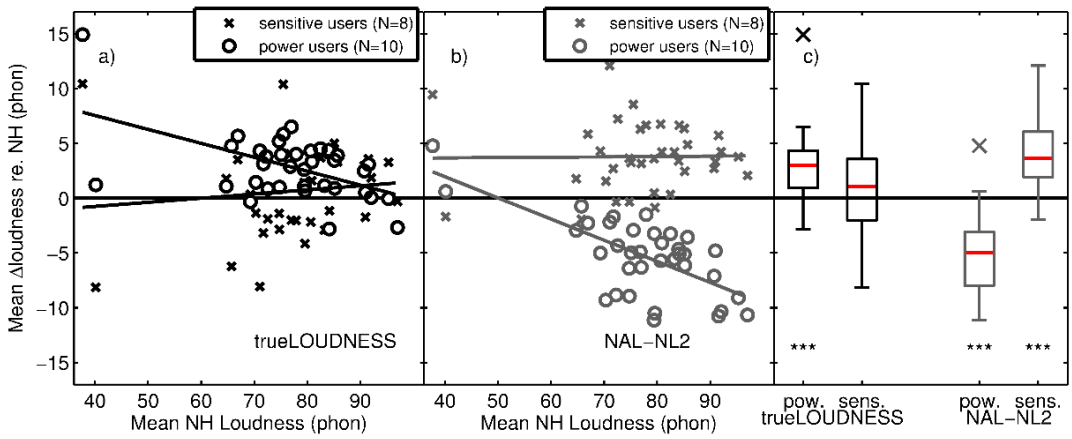


Fig. 3: Distance from the NH mean loudness rating in phons over the mean NH loudness rating in phons for both groups when fitting according to a) trueLOUDNESS and b) NAL-NL2. c) Pooled data over all test conditions. Median values for trueLOUDNESS were closer to zero compared to NAL-NL2 for both groups.

The median values for trueLOUDNESS were 3.0 phons and 1.1 phons for the power users and for the sensitive users, respectively. The median values for NAL-NL2 were −5.0 phons and 3.6 phons for the power users and for the sensitive users, respectively. For each of the four pooled data sets, we performed a Wilcoxon signed rank test with the null hypothesis that the data come from a distribution whose median is zero against the alternative hypothesis that the distribution does not have zero median. Only for the sensitive users with trueLOUDNESS fitting, the test indicated that the null hypothesis cannot be rejected ($p = 0.61$). For the three other conditions, the test indicated a median value different from zero ($p < 0.001$).

DISCUSSION AND CONCLUSION

In this experiment we tested whether the trueLOUDNESS measurements from the lab are related to the real-world loudness perception with hearing aids. We selected a user group for whom the trueLOUDNESS procedure predicts higher gains than NAL-NL2 ("power users"). The results from Fig. 3c showed that in this group the deviation from normal loudness with NAL-NL2 was between −3.1 and −8.0 phons (interquartile range), indicating that loudness was lower than normal in the power user group.

For the sensitive user group, we expected that perceived loudness would be higher than normal with NAL-NL2. The results in Fig. 3c showed a higher-than-normal loudness perception of 2.0 to 6.1 phons (interquartile range) with NAL-NL2. Both groups clearly separate in Fig. 3c, indicating that the grouping according to the trueLOUDNESS measurements in the lab reflects real-world loudness perception with hearing aids. Using the trueLOUDNESS gain prescription, normal loudness perception was achieved in the sensitive user group on average. For the power users, trueLOUDNESS gain predictions lead to higher-than-normal ratings of about 0.9 to 4.3 phons (interquartile range). These results can be used to define a gain reduction

rule for reducing the amount of loudness compensation for power users with the trueLOUDNESS fitting procedure.

The following conclusions can be drawn:

- Binaural broadband loudness scaling results from the lab according to trueLOUDNESS are related to the real-world loudness perception with hearing aids. This indicates that the binaural broadband loudness scaling results are reliable indicators during fitting of hearing aids to avoid under- or over-amplification.
- Hearing aids fitted with NAL-NL2 result in higher-than-normal loudness ratings for real-world vehicle sounds for one group of users (sensitive users), and lower-than-normal ratings for a second group of users (power users). These groups were defined based on their trueLOUDNESS-based relative gains, which shows that trueLOUDNESS measurements from the lab can be used to identify these groups of subjects.
- Prescription rules based solely on the hearing threshold cannot be further tuned towards a better loudness match for sensitive and power users simultaneously. Further information from binaural broadband loudness perception is required to decide if gains should be reduced or increased for an individual listener.
- The trueLOUDNESS fitting leads on average to a normal loudness perception in sensitive users and slightly higher-than-normal perception in power users.

REFERENCES

ANSI S3.4 (**2007**). American national standard: Procedure for the computation of loudness of steady sounds. American National Standards Institute.

Brand, T. and Hohmann, V. (**2002**). "An adaptive procedure for categorical loudness scaling," J. Acoust. Soc. Am., 112, 1597–1604, doi: 10.1121/1.1502902

EuroTrak DE (**2018**). EuroTrak DE 2018. Retrieved from https://www.ehima.com/wp-content/uploads/2018/06/EuroTrak_2018_GERMANY.pdf on 30.08.2019.

Heeren, W., Hohmann, V., Appell, J.-E., and Verhey, J.L. (**2013**). "Relation between loudness in categorical units and loudness in phons and sones," J. Acoust. Soc. Am., 133, 314–319, doi: 10.1121/1.4795217

Llorach, G., Oetting, D., Krueger, M., et al. (**2019**). "Vehicle noise: Loudness ratings, loudness models and future experiments with audiovisual immersive simulations," Proc. Inter-Noise 2019, Madrid, Spain.

Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., and Ewert, S.D. (**2016**). "Spectral and binaural loudness summation for hearing-impaired listeners," Hear. Res., 335, 179–192, doi: 10.1016/j.heares.2016.03.010

Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., and Ewert, S.D. (**2018**). "Restoring perceived loudness for listeners with hearing loss," Ear Hear. 39, 664–678, doi: 10.1097/AUD.0000000000000521

A word elicitation study including the development of scales characterizing aided listening experience

DORTE HAMMERSHØI[1,*], ANNE WOLFF[2], LYKKE J. ANDERSEN[3], RIKKE L. MORTENSEN[3], MADS D. NIELSEN[3], AND STEFANIE A. S. LARSEN[3]

[1] *Department of Electronic Systems, Signals and Information Processing, Aalborg University, Aalborg, Denmark*

[2] *Department of Otolaryngology, Head and Neck Surgery and Audiology, Aalborg University Hospital, Aalborg, Denmark*

[3] *Master program in Engineering Psychology, E-Study Board, Aalborg University, Denmark*

The purpose of the present study was to identify the terms hearing aid professionals and their patients use in the communication about the aided listening experience and develop scales that would help characterize this experience in the domain of corrective actions that a hearing care professional may apply. The study comprised a word elicitation task based on observations and interviews from consultations at the Aalborg University Hospital. The results were analyzed by developing an affinity diagram. The resulting 80 words were then sorted by three hearing professionals in a supervised card sorting session. The resulting attributes were included in a 63-point scale design, which (in a usability test including eight hearing-aid users) were considered easy to survey and use, but also including some redundancy and ambiguities. The results suggest that it is possible to develop scales based on the voluntary statements expressed during actual consultations, but it remains uncertain whether the expressions will be interpreted the same way by other patients and professionals.

INTRODUCTION

Literature suggests (see McCormack and Fortnum (2013) for a review) that many patients fitted with hearing aids don't use them and that it is a complicated process to find the reason for the lack of success. A letter from a frustrated hearing aid (HA) user explains it well (Fig. 1). The user has congenital hearing impairment and now has a cochlear implant (CI) in the left ear after many years of HA experience. A successful fitting experience for her includes good communication with the hearing care professional (HCP), which among other things entails that you understand what each other is saying, and have the same vocabulary for sound and hearing. The user has also observed the frustration among new users, who often find it difficult to describe their experience to the HCP.

*Corresponding author: dh@es.aau.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

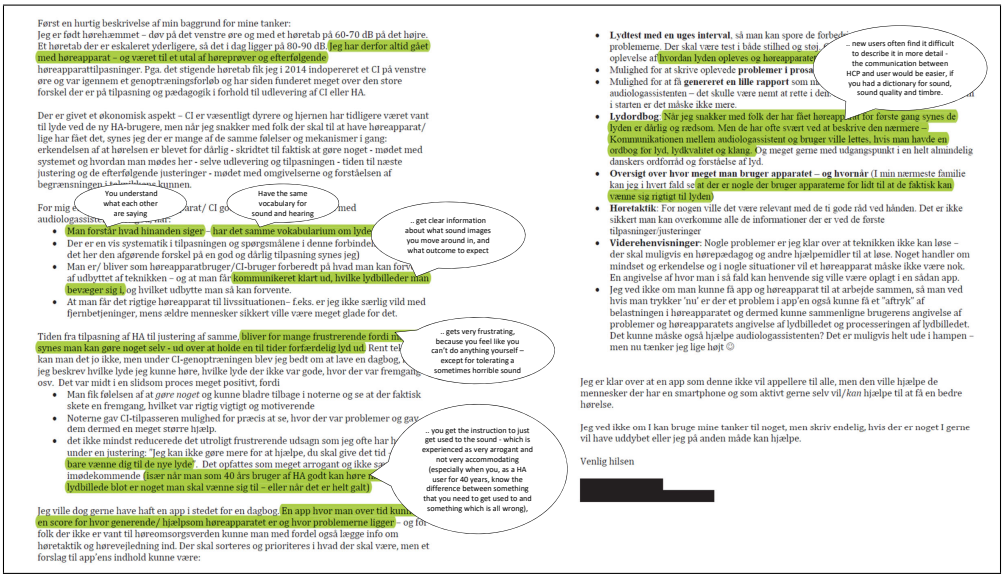


Fig. 1: Letter from frustrated HA user.

The scope of the study was to take a novel approach to 1) identifying the terms HCPs and their patients use in the communication about the aided listening experience, and 2) to develop scales that would help characterize this experience in the domain of corrective actions a HCP may apply.

METHODS

The overall method comprised of two steps, where the first was intended to elicit the words HA users and the HCPs use in the communication. The second step included the development of scales based on this, and a test trial that would provide qualitative experience with the use of such customized scales.

Word Elicitation

Word elicitation as understood by, for example, Francombe *et al.* (2014) was carried out to find suitable attributes for describing the aided listening experience for HA users. The process is depicted in Fig. 2 and includes observations during eight examinations at the audiological department of Aalborg University Hospital, involving fifteen patients, seven audiologists, and two medical doctors. The observations were succeeded by interviews for four HA users, and six interviews with two audiology assistants, which evoked further verbalization of relevant terms for the aided listening experience.

The central words that patients and professionals used for describing the aided listening experience were itemized, noted on cards (394 in total), and analyzed by

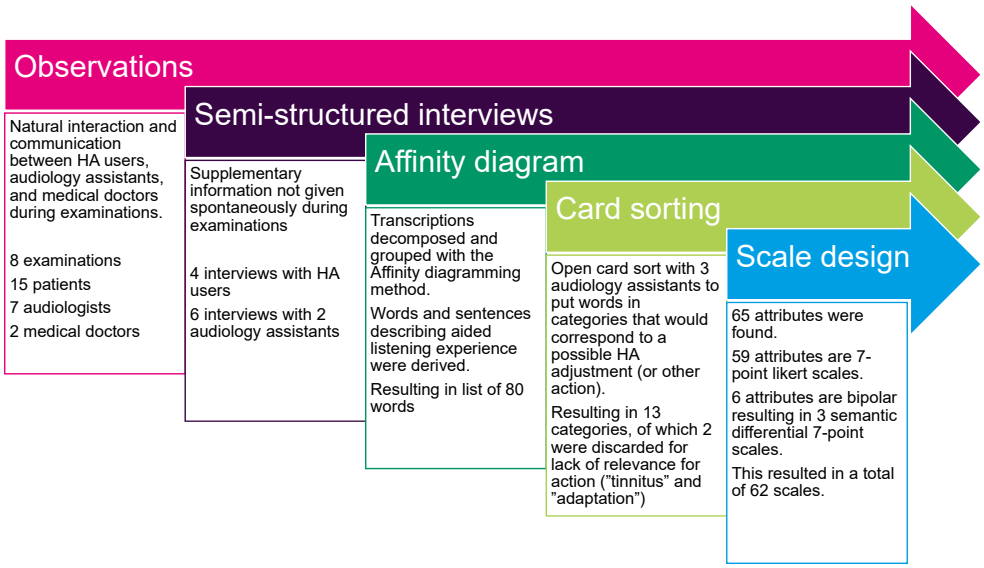


Fig. 2: Process used for word elicitation.

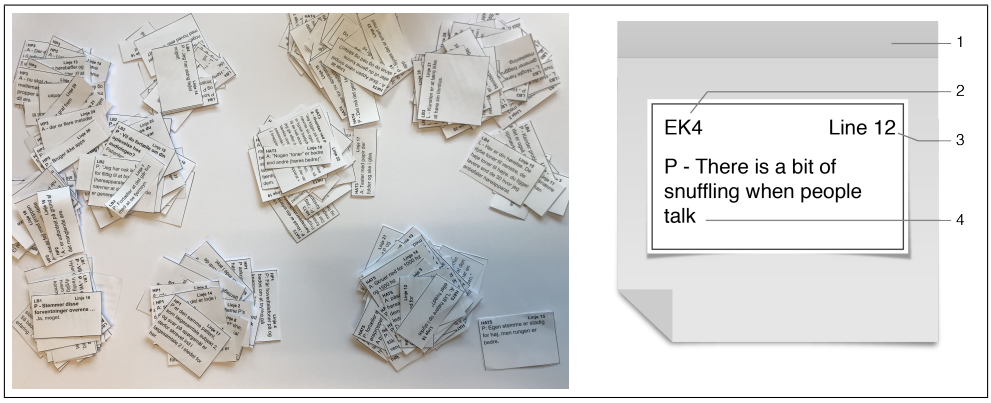


Fig. 3: **Left**: Notes decomposed from transcripts. **Right**: Example note card for the Affinity diagram: 1) post-it note, 2) label showing to which examination the quote was observed, i.e. "EK" represents follow-up ("efterkontrol"), 3) label showing the line in the transcription, where the quotation was found, 4) the quotation.

developing an Affinity diagram using inspiration from Goodwin (2009), Kuniavsky *et al.* (2013), andDan and Siang (2018), see Fig. 3. These notes were sorted by the four junior scientists into 80 relevant words, which were eventually grouped by two HCPs in a supervised open card sorting, according to Albert *et al.* (2013) for example, in groups that were operational in terms of what actions the HCPs would and could perform during consultations. This resulted in 13 categories, of which two were discarded for lack of direct relevance to relevant actions ("tinnitus" and "adaptation").

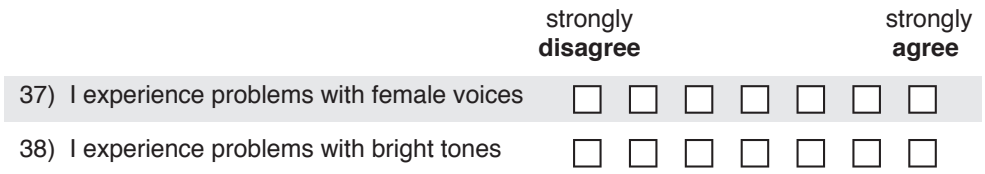


Fig. 4: Example of scale design. An index number is printed to the left, and then the item. The 7-point Likert scale ranging from "strongly disagree" to "strongly agree" is printed to the right.

Scale design and test

All 65 attributes identified in the Affinity diagram were included in a 63-point scale design (including three semantic differential scales). Fig. 4 shows an example of the scale design.

The scales were evaluated in a test with the eight HA users and two HCPs (audiology assistants). The HA users completed the scales, under the instructions following a think-aloud protocol, Mathison (2005). The results were subsequently shown to the HCPs, who were also instructed to follow the think-aloud protocol. Both HA users and HCPs participated in semi-structured interviews after their sessions.

RESULTS

The pooled results for the eight HA users are shown in Fig. 5. Most HA users were positive about using the scales as a basis of the dialogue in the context of examinations. Examples of positive statements are *"easy to survey and use"*, *"quick to fill out"* and *"the check-boxes are good so no self written text is needed"*. Six HA users stated that many of the scales were similar, but five of the six subjects also commented that this wasn't an issue and stated that *"it made sense to ask about the aspects from many different angles"* or *"to assess the validity of the answers"*. One HA user had problems understanding the scales and didn't complete the questionnaire. This HA user stated that binary yes/no questions would make more sense. One HCP also stated that for some of the questions, this would make more sense (i.e., either-or questions).

A qualitative examination of the response distribution reveals that items 1-4 for *Loudness*, 18-20 for *Occlusion*, 32-34 for *Low frequencies* are very similar in their score distribution, and might prove redundant by further examination. These items are also grouped in the same category from the HCP card sorting, whereas other items do not compare well within the assigned category. This is the case, for example, for items 52-56 for *2 kHz*, where the scores vary considerably between items. Since the category represents a given action, which an HCP would exert (e.g., adjustment of the amplification at 2 kHz), it could also suggest that similar actions are made in response to very different patient experiences, and if so, carry options for misunderstandings. This suggests that the grouping of the attributes may not adequately represent relevant actions of adjustments, or that the items do not adequately represent the patient

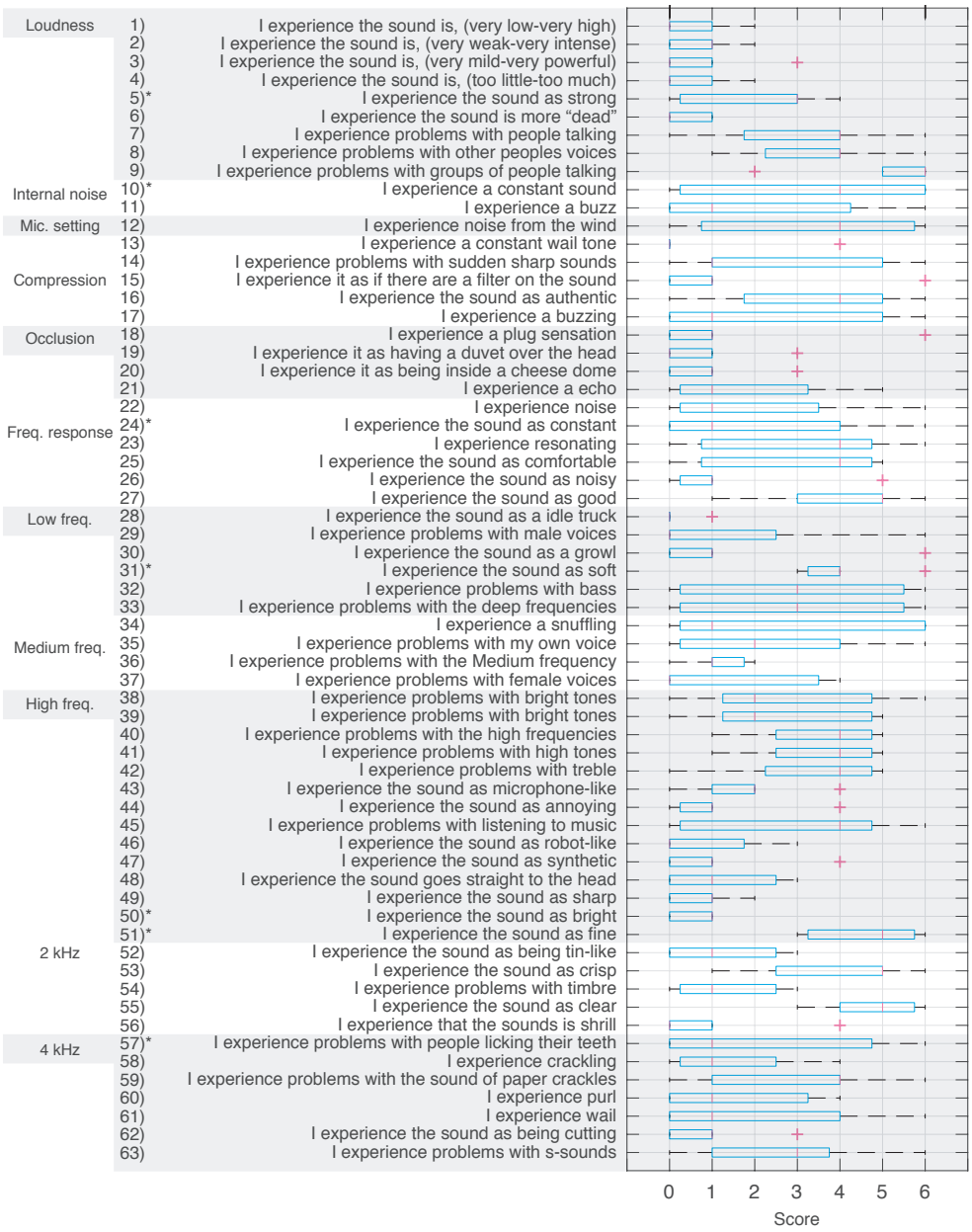


Fig. 5: Data from 8 HA users (numerical scores assigned corresponding to the categories in the 7-point Likert scale, hence “0” corresponds to “strongly disagree”, and “6” to “strongly agree”, except for the top four items, which had the end labels listed in brackets after the item). The categories (our translation) to the left are the result of the card sort. The items (our translation) are shown on the y-axis and the score on x-axis. The items that were found ambiguous or lacking in context in the interview feedback are marked with *.

experience that should result in the given action.

Items 16, 25, 27, 53 and 55 are positive aided listening experiences as opposed to the other, negative aided listening experiences. This means that the positive scores are represented to the right (agreeing to the statement), whereas all other items have positive scores at the left (disagreeing with the statement). Although all scales were presented with appropriate markers for the patient, it can, however, not be ruled out that this may have contributed to the variation between items within the *2 kHz* category.

Items 5, 10, 24, 31, 50 and 51 were found by both HA users and HCPs to be ambiguous or lacking adequate context. This suggests that such statements may only be relevant in particular contexts, and the situation where the HA users responded to the questionnaires did not provide the environment for such assessment. It is possible that establishing a listening scenario, where the patients are stimulated with relevant listening experiences, will improve the reliability of such assessments. If the questionnaires should be filled out in out-of-clinic situations (e.g., prior to a clinical visit), it would seem necessary to develop instructions to accompany some of the items, as they would otherwise also lack context in the out-of-clinic environment.

DISCUSSION

The HCPs pointed out that the HA users would not complete a scale questionnaire with 63 scales. None of the HA users expressed that the amount of scales was too large, only that some of them were similar and stated that *"it is something you fill out believing that it is important for the treatment"*. It is possible that the HA users experience the completion of large questionnaires as satisfying, as long as the activity makes sense to them. In the present study, they completed the questionnaires in the clinic, and the activity provided extra attention.

If the scales are investigated for redundancy this could decrease the sense of similarities and amount of scales. On the other hand, an advantage of the redundancy is the possibility to investigate the validity of the answers given. Two items were by mistake identical (items 38 and 39), which was pointed out to the HA users. Yet the answers differ for one patient. Since scales are inherently only relevant for assessment of perceptive dimensions, which by nature can have a range of options, there will be "noise" in the responses. A test-retest could probe the magnitude of this uncertainty and reveal if some of the dimensions investigated (certain items or categories) are more prone to intrinsic variance. This was not done in the present study.

Another benefit from including all attributes is a stronger probability of including attributes HA users can relate to, as they use a richer vocabulary to describe their experiences (Tab. 1). Further studies would be required to better understand which of the scales best characterize the variance of the aided listening experience, which links efficiently to the actions possible by the HCPs. The observed vocabulary of the HCPs (the audiology assistants in particular), may constitute a desired refinement

Audiology assistant		**Patient**		**Patient (cont.)**	
Bas/base	4	Autentic sound	1	Idle truck	1
Bas amplification	1	Better	4	Lav	1
Bas sound	1	Softer	1	Male voices	1
Pleasant	1	Humming	3	More dead	1
High tones	2	Female voices	2	Like a microphone	1
Bright tones	2	Pillow over head	1	Mild	1
Echo	2	Own voice	4	Music	1
Pork roast with crisp skin	2	Some adults talking	1	(see talking)	
Frequency	1	Filter	2	Paper scratching	1
High frequencies	1	Fine	1	Clotted sensation	3
High tones	1	People talking	2	Radiator turned on	1
Enclosed	1	Powerful	1	Licking teeth	1
Bright	1	Cork	1	Talk	3
Purling	2	Stab	1	Snuffling	2
Robotic	1	Straight to the head	1	Som en vindmølle	1
S'sounds	1	Said in church	1	As a filter (see filter)	
Spoon in glas	1	Bottom	1	Voices	2
Scratching	3	Inside head	1	Strong	1
Scratching	1	Annoying	1	Noise from wind	1
Sharp	1	Smooth	1	Weak	1
Strength	1	(see female voices)		Syntetisk	1
Woom woom	1	Clearer	1	Heavy in head	1
Medical doctor		Click in ear	1	Clear	2
Bas	2	Constant howling	1	Dense filter	1
Deep tones	1	Constant sound	1	Intense	2
High tones	1	Noise	1	be in cheese dome	1
Light tones	1				
Elevate the sound	1				
Mid tones	2				

Table 1: Prevalence of central words used in the description of the aided listening experience during consultation, HA-fitting and follow-up.

guided by previous experience relating to which terms most efficiently excite which reflections in the patient. One phrase "pork roast with crisp skin" ("flæskesteg med sprøde svær") was, for example, deliberately used in the interaction as an instrument to trigger a listening experience, which was considered to reveal sub-optimal fitting, because of a high density of transient and unvoiced consonant combinations. Also the "woom woom" was a self-engineered stimulus to bring attention to the specifics of the qualities of the listening experience at low frequencies.

CONCLUSION

The results suggest that it is possible to develop scales based on the voluntary statements expressed during actual consultations, but that the expressions may not be interpreted the same way by other patients and professionals. The results also suggest that standardized scales (e.g., MUSHRA by ITU-R (2015)) may be interpreted differently by different users.

ACKNOWLEDGEMENTS

Sincere thanks to the HA users and the staff at the audiological department at Aalborg University Hospital for their participation in experiments and interviews. Collaboration and support by Innovation Fund Denmark (Grand Solutions 5164-00011B), Oticon, GN Hearing, Widex Sivantos Audiology and other partners (University of Southern Denmark, Aalborg University, the Technical University of Denmark, Force, and Aalborg, Odense and Copenhagen University Hospitals) is sincerely acknowledged.

REFERENCES

Albert, W., and Tullis, T. (**2013**). "Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics," Ch. 6: Self-reported Metrics; Ch. 7: Behavioral and physiological metrics; Ch. 9: Special Topics in *Interactive Technologies Ser. Elsevier Science & Technology*, second edition.

Dam, R. and Siang, T. (**2018**). "Affinity diagrams. Learn how to cluster and bundle ideas and facts" online at *Interaction Design Foundation*.

Francombe, J., Mason, R., Dewhirst, M., and Bech, S. (**2010**). "Elicitation of attributes for the evaluation of audio-on-audio interference," J. Acoust. Soc. Am., **136**(5), 2630-2641.

Goodwin, K. (**2009**). "Designing for the Digital Age: How to Create Human-Centered Products and Services," Ch. 10: Making Sense of Your Data: Modeling. Wiley Publishing Inc.

ITU-R (**2015**). "RECOMMENDATION ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Geneva, Switzerland.

Kuniavsky, M., Goodman, E., and Moed, A. (**2012**). "Observing the User Experience: A Practitioner's Guide to User Research," Elsevier Science & Technology, second edition, 2012. Ch. 8 More than words: Object-Based technique; Ch. 9 Field Visits: Learning from Observation; Ch. 15 Analyzing qualitative data.

Mathison, S. (**2005**). "Think-Aloud Protocol," In: *Encyclopedia of Evaluation* edited by S. Mathison, by Sage.

McCormack, A., and Fortnum, H. (**2013**). "Why Do People Fitted With Hearing Aids Not Wear Them?," Int. J. Audiol., **52**, 360-368.

Improving robustness of adaptive beamforming for hearing devices

ALASTAIR H. MOORE, PATRICK A. NAYLOR*, AND MIKE BROOKES

Department of Electrical and Electronic Engineering, Imperial College, London, UK

Fixed beamforming for hearing aids is suboptimal due to mismatches in real-world situations between the assumed and encountered sound fields. Adaptive beamforming potentially provides better performance but may degrade it if the characteristics of the signal required by the design procedure are inaccurately estimated. This paper proposes a straightforward but sufficiently rich model for the sound field that can be used to increase the robustness of adaptive beamformer design. A method for estimating the model parameters is also presented. In reverberant acoustic conditions, the proposed method improves performance by > 1 dB even at -16 dB SNR, the lowest signal to noise ratio (SNR) tested. Furthermore, it is shown to be robust in a variety of acoustic conditions which do not conform to the sound field model, and to inaccurate steering of the array.

INTRODUCTION

Current drivers for innovation in microphone array beamforming include the increasing availability of more powerful computational resources, and the increasing significance of several emerging application areas, such as spherical arrays described in Rafaely (2015) and Jarret *et al.* (2017), robot audition as in Tamai *et al.* (2004) and Löllmann *et al.* (2017) and binaural hearing aids discussed in Klasen *et al.* (2007) and Moore *et al.* (2018). The linearly constrained minimum variance (LCMV) family of beamformers are widely used in acoustic beamforming due to their ability to suppress noise without distorting the target signal. The original Capon beamformer (Capon, 1969), or minimum power distortionless response (MPDR) beamformer (van Trees, 2002), minimise the output power given the sample covariance matrix (SCM), whereas the minimum variance distortionless response (MVDR) beamformer design is based on the noise covariance matrix (NCM). Both use a steering vector to set the distortionless constraint on the target signal and, under ideal conditions, they are equivalent. In practice, their sensitivity to errors in the steering vector differs (Cox *et al.*, 1987; Ehrenberg *et al.*, 2010). For the MVDR beamformer, the effect of mis-steering is merely to attenuate the desired signal, whereas for the MPDR signal, cancellation occurs since the distortionless constraint is not matched to the target signal. Furthermore, it is shown in Ehrenberg *et al.* (2010) that an inaccurate estimate of the NCM is preferable to an accurate SCM.

In reverberant environments, even with a perfectly aligned anechoic steering vector,

*Corresponding author: p.naylor@imperial.ac.uk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

coherent reflections originating from the target source cause signal cancellation for the MPDR. Using reverberant relative transfer function (RTF) steering vectors as in Gannot *et al.* (2001), the distortionless constraint preserves the direct path and at least the first few early reflections, reducing the potential for signal cancellation. Effective RTF estimation is an ongoing research problem (Markovich *et al.*, 2009; Markovich-Golan and Gannot, 2015).

Obtaining an estimate of the NCM which is completely uncorrelated with the target speech and is effective in practical applications is difficult to achieve. Informed spatial filtering is a concerted research effort to model the statistics of different signal components from which the total NCM can be obtained (Thiergart and Habets, 2003; Braun and Habets, 2015; Schwartz *et al.*, 2016; Chakrabarty and Habets, 2018; Braun *et al.*, 2018; Moore *et al.*, 2019a). In many cases, estimated hyper-parameters such as the speech presence probability (SPP), coherent to diffuse ratio (CDR), or one or more directions of arrival (DOAs) control when to update each statistic. Inevitably such estimates become less accurate at low signal to noise ratios (SNRs) and in time-varying scenarios which may, for example, lead to target energy leaking into the NCM.

Robust beamformers have been proposed which reduce sensitivity to errors and increase the white noise gain at the expense of reduced directivity, for example in Cox *et al.* (1987) and Li *et al.* (2003). These generally involve diagonal loading of the covariance matrix. Ultimately, to remove all possibility of signal cancellation the conservative approach often adopted in real-world implementations is to design a fixed, super-directive beamformer using an assumed noise model (Bitzer and Simmer, 2001).

In this paper, we propose a simple model of the sound field that is sufficiently rich to describe complex scenes and whose parameters can be estimated at low SNRs. We assume that calibration measurements of the array manifold are available and that the steering direction is known. Using this information, a method for estimating the time-varying parameters of the sound field model is proposed. The adequacy of the proposed model and resulting SCM is evaluated in the specific context of MPDR beamforming for binaural hearing aids (HAs) but it can equally be applied to other filter structures and array geometries. In this case, as is customary, a known target direction is realized by fixing the steering direction towards the front of the head and requiring that the listener turn to face the desired talker.

FORMULATION AND PROPOSED MODEL

The time domain signal received at the m^{th} microphone in an array is denoted

$$y_m(t) = \sum_{l=1}^{L} x_{m,l}(t) + v_m(t) \qquad \text{(Eq. 1)}$$

where t is the time index, l is the source index, $x_{m,l}(t)$ is the signal due to the l^{th} source and $v_m(t)$ is sensor noise. In a reverberant enclosure,

$$x_{m,l}(t) = h_{m,l}(t) * s_l(t) \tag{Eq. 2}$$

where $s_l(t)$ is the signal emitted by the l^{th} source, $h_{m,l}(t)$ is the acoustic impulse response (AIR) from the l^{th} source to the m^{th} microphone, and $*$ denotes convolution. Decomposing $h_{m,l}(t)$ into the direct path, $h^{(d)}_{m,l}(t)$, and reflected components, $h^{(r)}_{m,l}(t)$, Eq. 2 can be rewritten

$$x_{m,l}(t) = (h^{(d)}_{m,l}(t) + h^{(r)}_{m,l}(t)) * s_l(t) \tag{Eq. 3}$$

$$= h^{(d)}_{m,l}(t) * s_l(t) + h^{(r)}_{m,l}(t) * s_l(t) \tag{Eq. 4}$$

$$= x^{(d)}_{m,l}(t) + x^{(r)}_{m,l}(t) \tag{Eq. 5}$$

where $x^{(d)}_{m,l}(t)$ and $x^{(r)}_{m,l}(t)$ are the direct path and reflected components of $x_{m,l}(t)$, respectively.

Combining Eq. 1 and Eq. 5, the microphone signals can be written

$$y_m(t) = \sum_{l=1}^{L} x^{(d)}_{m,l}(t) + \sum_{l=1}^{L} x^{(r)}_{m,l}(t) + v_m(t) \tag{Eq. 6}$$

and can equivalently be expressed in the short time Fourier transform (STFT) domain as

$$Y_m(\nu,\ell) = \sum_{l=1}^{L} X^{(d)}_{m,l}(\nu,\ell) + \sum_{l=1}^{L} X^{(r)}_{m,l}(\nu,\ell) + V_m(\nu,\ell) \tag{Eq. 7}$$

where capitalized letters denote the STFT of the quantities denoted by the corresponding lowercase letters in Eq. 6, and ν and ℓ are the frequency and frame indices respectively. Stacking the signals for all M microphones in an array to give, for example, $\mathbf{y}(\ell) = \begin{bmatrix} Y_1(\ell) & \dots & Y_M(\ell) \end{bmatrix}^T$, Eq. 7 then becomes

$$\mathbf{y}(\ell) = \sum_{l=1}^{L} \mathbf{x}^{(d)}_l(\ell) + \sum_{l=1}^{L} \mathbf{x}^{(r)}_l(\ell) + \mathbf{v}(\ell) \tag{Eq. 8}$$

where $(\cdot)^T$ denotes the transpose, and since all frequency bins are processed independently, the dependence on ν has been dropped for clarity.

The proposed signal model makes four simplifying assumptions: (i) all sources are in the far field, such that $h^{(d)}_{m,l}(t)$ is identical to the response of the array to a plane-wave from the same direction as the l^{th} source, up to a scalar gain and time shift; (ii) the array is sufficiently compact that the RTF to each microphone with respect

to the reference microphone can be represented by a multiplicative constant in the STFT domain (Avargel and Cohen, 2007); (iii) the direct path signals are W-disjoint orthogonal (Yilmaz and Rickard, 2004), such that in each time-frequency bin, a single source is dominant, (iv) the sum of all reflected signals reduces to a diffuse field which, by definition, is isotropic since the incident power from all directions is the same. With these assumptions, Eq. 8 reduces to

$$\dot{\mathbf{y}}(\ell) = \mathbf{a}(\Omega(\ell))\dot{S}_{l(\ell)}(\ell) + \boldsymbol{\gamma}(\ell) + \mathbf{v}(\ell) \quad \text{(Eq. 9)}$$

where $\boldsymbol{\gamma}(\ell)$ is the diffuse noise signal, $l(\ell)$ and $\Omega(\ell)$ are the index and corresponding DOA, respectively, of the dominant source in the ℓ^{th} frame (which may be different at each frequency), $\dot{S}_{l(\ell)}(\ell)$ is the signal due to the dominant source as observed at the arbitrarily selected reference microphone and $\mathbf{a}(\phi)$ is the plane-wave array manifold expressed as the RTF to each microphone with respect to the reference microphone.

The covariance of the microphone signals is

$$\mathbf{R}_{\mathbf{y}}(\ell) = \mathbb{E}\{\mathbf{y}(\ell)\mathbf{y}^H(\ell)\} \quad \text{(Eq. 10)}$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator and $(\cdot)^H$ denotes the conjugate transpose. Using the signal model defined in Eq. 9 and assuming the three terms are uncorrelated

$$\mathbf{R}_{\dot{\mathbf{y}}}(\ell) = \mathbb{E}\{|\dot{S}^2_{l(\ell)}(\ell)|\}\mathbf{a}(\Omega(\ell))\mathbf{a}^H(\Omega(\ell)) + \mathbb{E}\{\boldsymbol{\gamma}(\ell)\boldsymbol{\gamma}^H(\ell)\} + \mathbb{E}\{\mathbf{v}(\ell)\mathbf{v}^H(\ell)\} \quad \text{(Eq. 11)}$$

where $(\cdot)^*$ is the conjugate.

It can now be seen that each term on the right hand side of Eq. 11 can be expressed as the product of a fixed matrix and a scalar parameter. This leads to

$$\mathbf{R}_{\dot{\mathbf{y}}}(\ell) = \sigma_d(\ell)\mathbf{R}_{\mathbf{a}}(\Omega(\ell)) + \sigma_{\gamma}(\ell)\mathbf{R}_{\gamma} + \sigma_v(\ell)\mathbf{R}_{\mathbf{v}} \quad \text{(Eq. 12)}$$

where the covariance is defined by four parameters $\Omega(\ell)$, $\sigma_d(\ell)$, $\sigma_{\gamma}(\ell)$ and $\sigma_v(\ell)$ denoting, respectively, the DOA of the plane-wave component and the powers of the plane-wave, diffuse and sensor noise components.

MODEL PARAMETER ESTIMATION

A method is presented to estimate the parameters of the signal model proposed in Eq. 12, and use them to obtain an estimate of the NCM. The algorithm operates directly in the STFT domain where a recursive estimate, $\hat{\mathbf{R}}_{\mathbf{y}}(\ell)$, of the sample covariance matrix, $\mathbf{R}_{\mathbf{y}}(\ell)$, is obtained as

$$\hat{\mathbf{R}}_{\mathbf{y}}(\ell) = \alpha\hat{\mathbf{R}}_{\mathbf{y}}(\ell-1) + (1-\alpha)\mathbf{y}(\ell)\mathbf{y}^H(\ell) \quad \text{(Eq. 13)}$$

where α defines the time constant.

The model parameters can then be found from the solution to the optimization problem

$$\underset{\Omega(\ell),\sigma_d(\ell),\sigma_{\gamma}(\ell),\sigma_v(\ell)}{\arg\min} \{||\hat{\mathbf{R}}_{\mathbf{y}}(\ell) - \mathbf{R}_{\dot{\mathbf{y}}}(\ell)||_F\} \quad \text{(Eq. 14)}$$

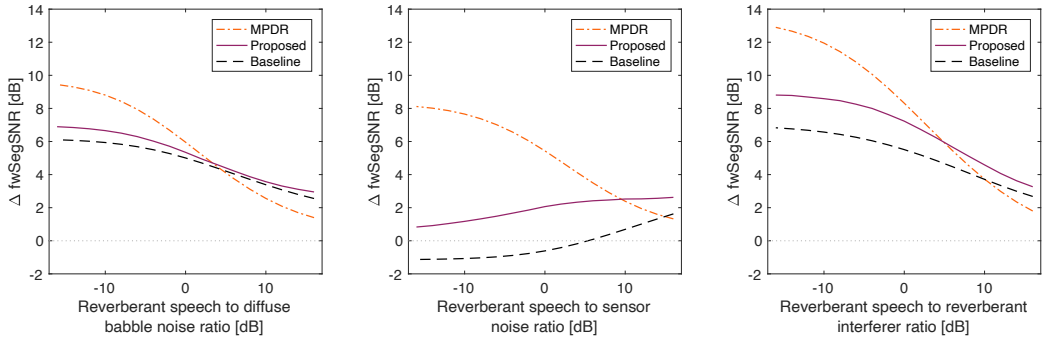


Fig. 1: Improvement in A-weighted segmental SNR as a function of SNR with respect to (a) reverberant babble from 16 directions on circle, (b) sensor noise, (c) interfering speech from -67.5°. In each condition the SNR with respect to the other noise types is 20 dB.

where $||\cdot||_F$ denotes the Frobenius norm. Many approaches to solving Eq. 14 are available. The approach adopted here is to obtain the ordinary least squares solution for $\sigma_d(\ell)$, $\sigma_\gamma(\ell)$ and $\sigma_v(\ell)$ for a candidate set of values of $\Omega(\ell)$, from which the best fit is selected. The final NCM estimate is given by Eq. 12 using the parameter estimates obtained.

SIMULATION EXPERIMENTS

The efficacy of the proposed method in the context of MPDR/MVDR beamforming is evaluated in two experiments. In the first, algorithm performance is assessed as a function of SNR where a single type of noise is dominant. In the second, six different scenarios are considered in which, like real-world acoustic environments, the composition of the sound field is more complicated. Listening examples are available at `https://squaresetsound.com/demos/constrained-covariance-matrix-estimation-2019`.

Microphone signals are simulated for a $7.9 \times 6.0 \times 3.5$ m room with a reverberation time of 250 ms according to Eq. 1 and Eq. 2. Anechoic speech is convolved with hearing aid room impulse responses (HARIRs) measured from a horizontal ring of 16 loudspeakers positioned at azimuth angles, $\phi \in \{0^\circ, 22.5^\circ, \ldots, 337.5^\circ\}$, to a pair of behind the ear (BTE) hearing aids ($M = 4$) worn by a head and torso simulator (HATS) (subject 42; Moore *et al.*, 2019b). Sensor noise is simulated using independent identically distributed Gaussian noise which is filtered to match the spectra of real sensor noise recordings for the microphones used in Moore *et al.* (2019b).

All beamformers are designed based on a repeated set of AIR measurements for the same hearing aids and HATS (as per subject 42; Moore *et al.*, 2019b) but made on a different day, after complete removal and replacement of the hearing aids from the

mannequin, and the mannequin from the measurement room. These hearing aid head-related impulse responses (HAHRIRs) (subject s28; Moore *et al.*, 2019b) are truncated to remove reflections from the room and so contain only direct-path propagation. The steering vector, $\mathbf{d} = \mathbf{a}(0)$, is defined here as the RTF with respect to the front right microphone for a plane-wave arriving from $\phi = 0$.

The covariance matrix for a diffuse field is assumed to be cylindrically isotropic, since real rooms tend to have more absorption in the floor and/or ceiling compared to the walls (Schwarz *et al.*, 2015). It is computed by discretising

$$\mathbf{R}_\gamma = \int_{\Omega=0}^{2\pi} \mathbf{h}^{(d)}(\Omega)\mathbf{h}^{(d)H}(\Omega)d\Omega \qquad \text{(Eq. 15)}$$

to the 7.5° resolution of the HAHRIRs.

Beamformer weights are calculated according to

$$\mathbf{w} = {\mathbf{R}_\varepsilon}^{-1}\mathbf{d}\left[\mathbf{d}^H{\mathbf{R}_\varepsilon}^{-1}\mathbf{d}\right]^{-1} \qquad \text{(Eq. 16)}$$

where $\mathbf{R}_\varepsilon = \mathbf{R} + \varepsilon\mathbf{I}$, $\mathbf{I}$ is the identity matrix and $\varepsilon \geq 0$ is set to limit the condition number of $\mathbf{R}_\varepsilon$ to $\leq$100. The baseline method assumes cylindrically isotropic noise (i.e., $\mathbf{R} = \mathbf{R}_\gamma$). The proposed method uses the estimated covariance matrix from Eq. 12 (i.e., $\mathbf{R} = \mathbf{R}_{\dot{\mathbf{y}}}(\ell)$ with the parameters from Eq. 14). The robust MPDR method uses the estimated sample covariance matrix from Eq. 13 (i.e., $\mathbf{R} = \hat{\mathbf{R}}_{\mathbf{y}}(\ell)$). It should be noted that the baseline method is signal independent (fixed), whereas the proposed method and MPDR method are adaptive with, respectively, 4 and $M(M+1)/2 = 10$ estimated parameters per time-frequency cell.

Signals are processed at a sample rate of 20 kHz in the STFT domain with 16 ms frames overlapping by 50 %. The time constant for recursive estimation of $\hat{\mathbf{R}}_{\mathbf{y}}(\ell)$ in Eq. 13 is chosen to be 50 ms in the following experiments.

Experiment 1

The spatial arrangement of sound sources is fixed throughout Experiment 1. The desired source is male speech from $\phi = 0°$, and there are three noise sources: (a) an interferer (male speech) at $\phi = -67.5°$ (to the listener's right); (b) babble noise from sixteen equally-spaced azimuths on the horizontal plane, such that powers of the direct path signals arriving from all azimuth directions are the same; (c) sensor noise.

The levels of the target and interferer speech sources are measured as the average active level in dB of the reverberant signals at the two front microphones, when each sound source is presented from $\phi = 0°$, as defined in ITU-T (1993) and Brookes (1997). Sound presentation from other angles (i.e., interferers) therefore includes the effect of the natural directivity of the head/array geometry. The levels of the noise signals are measured as the average power at the front two microphones.

The level of the desired source is fixed, and in each of three test cases, the effect of varying the level of one, dominant, noise source is assessed whilst keeping the other

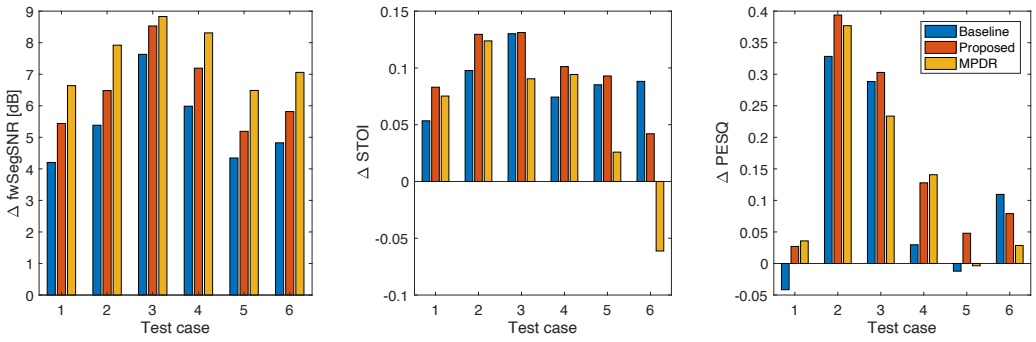


Fig. 2: Improvement in (a) A-weighted segmental SNR (b) STOI and (c) PESQ for the 6 test cases defined in Table 1.

two fixed at −20 dBr with respect to the desired source. Performance is evaluated in terms of the improvement in frequency-weighted segmental SNR (fwSegSNR) at the reference microphone where the clean target is the direct path component of the desired speech. The improvement in fwSegSNR is an appropriate metric as it quantifies the noise reduction only during periods of target speech activity and changes can be easily interpreted (in dBs) regardless of the specific listening situation.

Figure 1shows that in all cases the proposed method outperforms the baseline. In Figure 1(a), where the dominant noise is babble, the diffuse noise model of the baseline is a reasonably good approximation, and the benefit of the proposed method is smallest. In Figure 1(b), where the dominant noise is uncorrelated between sensors (i.e., spatially white) the baseline method actually reduces the fwSegSNR, which is consistent with the well known trade-off between directivity and white noise gain. In Figure 1(c), where the dominant noise source is reverberant interfering speech, the benefit of the proposed method is most clearly seen with about 1 dB improvement over a wide range of signal to interference ratios (SIRs).

In all cases the MPDR beamformer performs the best at low SNRs but even worse than the baseline at high SNRs. At low SNRs the estimated sample covariance matrix is dominated by noise and so good noise reduction is achieved. In contrast, at high SNRs the estimated sample covariance matrix contains the direct path target and coherent reflections which leads to target cancellation. Experiment 2 investigates the robustness to model violations and employs additional metrics which further highlight the degradation caused by the MPDR method.

Experiment 2

As is well-known, sound fields in real-world situations do not normally conform to the idealised situation of having a single dominant noise type. To evaluate the effect

	Interferer Male @ −67.5°	Interferer Female @ 67.5°	Babble directions	Steering error
1	✓		all	
2	✓	✓	all	
3			22.5°... 157.5°	
4	✓		22.5°... 157.5°	
5	✓		all	7.5°
6	✓		all	15°

Table 1: Test case definitions.

of model violation in complex scenarios, the following additional noise sources are defined: (a) an interferer (female speech) at $\phi = 67.5^\circ$ (to the listener's left) and (b) babble noise only from the seven DOAs to the listener's left (i.e., $22.5^\circ \leq \phi \leq 157.5^\circ$). In this experiment, the levels of the target and all active noise sources are equal, except sensor noise, which is always present at −20 dB with respect to the target. Table 1 defines which noise sources are active in each test case. Test Case 1 has the same spatial arrangement as in Experiment 1, but with the levels of intefering speech and babble being equal. Test Case 2 adds a second interferer. Test Case 3 has non-isotropic babble with no interferers, and Test Case 4 reinstates the male interferer to the right. Test Cases 5 and 6 are the same as Test Case 1 but consider the effect of missteering, where the listener's head is not directly facing the desired source.

In addition to the improvement in fwSegSNR, we also consider the improvements in short-time objective intelligibility measure (STOI) (Taal *et al.*, 2011) and PESQ (ITU-T, 2003).

Figure 2 shows that in Test Cases 1 to 5, all metrics suggest that the proposed method outperforms the baseline. Only in Test Case 6, where the steering misalignment is 15°, do the STOI and PESQ metrics suggest that performance of the proposed method is degraded. Whilst the MPDR method is effective at reducing the noise, as indicated by its superlative improvement in fwSegSNR, both the STOI metric and informal listening suggest that there is also signal degradation. Consistent with the literature (Li *et al.*, 2003; Ehrenberg *et al.*, 2010), the MPDR beamformer is particularly sensitive to steering errors as seen in Test Cases 5 and 6.

DISCUSSION AND CONCLUSIONS

The proposed model of the sound field as the weighted sum of three idealised components allows a wide range of real-world sound fields to be approximated. By constraining the allowed DOA of the plane-wave component to a fixed set of candidates, the potential for signal cancellation during desired speech activity is minimised. When the desired speech is dominant, provided the steering error is not

too large, it is likely that the DOA coinciding with the look direction is selected, even in the presence of reflections, and so the MVDR's distortionless constraint ensures that no cancellation of the direct path wave-front occurs. The remaining components of the covariance matrix are a combination of diffuse and spatially white noise and so are as benign as a fixed beamformer. When the desired speech is not dominant or absent, the contribution of the plane-wave component allows the estimated covariance matrix to adapt, at least to some extent, to the irregularities of the encountered sound field, improving the attenuation compared to an ideal model of the noise distribution. By continuously adapting the estimated covariance matrix, the method can respond immediately to changes in the acoustic scene. Combining the proposed method with head-tracker informed beam-steering as in Moore *et al.* (2018), it is feasible to relax the requirement for the user to face the target source. Simulation experiments using measured reverberant impulse responses and challenging levels of realistic noise show that the proposed method outperforms a fixed beamformer by $\geq$1 dB over a range of acoustic scenarios and is more robust than a conventional, diagonally loaded MPDR beamformer.

ACKNOWLEDGEMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council [grant number EP/M026698/1].

REFERENCES

Avargel, Y., and Cohen, I. (**2007**), "On multiplicative transfer function approximation in the short-time Fourier transform domain," IEEE Signal Process. Lett., **14**(5), 337-340.

Bitzer, J., and Simmer, K.U. (**2001**), "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, 19-38.

Braun, S., and Habets, E.A.P (**2015**), "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," EURASIP J. Audio Speech Music Process., vol. 2015, no. 1, p. 34.

Braun, S., Kuklasiński, A., Schwartz, O., Thiergart, O., Habets, E.A.P., Gannot, S., Doclo, S., and Jensen, J. (**2018**), "Evaluation and comparison of late reverberation power spectral density estimators," IEEE/ACM Trans. Audio Speech Lang. Process., **26**(6), 1056-1071.

Brookes, D.M. (**1997**), "VOICEBOX: A speech processing toolbox for MATLAB," 1997–2016. [Online]. Available: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

Capon, J. (**1969**), "High resolution frequency-wavenumber spectrum analysis," Proc. IEEE, **57**, 1408-1418.

Chakrabarty, S., and Habets, E.A.P. (**2018**), "A Bayesian approach to informed spatial filtering with robustness against DOA estimation errors," IEEE/ACM Trans.

Audio Speech Lang. Process., **26**(1), 145-160.

Cox, H., Zeskind, R.M., and Owen, M.M. (**1987**), "Robust adaptive beamforming," IEEE Trans. Acoust. Speech Signal Process., **35**(10), 1365-1376.

Ehrenberg, L., Gannot, S., Leshem, A., and Zehavi, E. (**2010**), "Sensitivity analysis of MVDR and MPDR beamformers," Proc. IEEE Conv. Electrical and Electronics Engineers, 416-420.

Gannot, S., Burshtein, D., and Weinstein, E. (**2001**), "Signal enhancement using beamforming and nonstationarity with applications to speech," IEEE Trans. Signal Process., **49**(8), 1614-1626.

ITU-T (**1993**), "Objective measurement of active speech level," Intl. Telecommunications Union (ITU-T), Recommendation P.56, Mar. 1993.

ITU-T (**2003**), "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Intl. Telecommunications Union (ITU-T), Recommendation P.862, Nov. 2003.

Jarrett, D.P., Habets, E.A.P., and Naylor, P.A. (**2017**), *Theory and Applications of Spherical Microphone Array Processing*, ser. Springer Topics in Signal Processing. Springer International Publishing, 2017.

Klasen, T.J., Bogaert, T.V. den, Moonen, M., and Wouters, J. (**2007**), "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," IEEE Trans. Signal Process., **55**(4), 1579-1585.

Li, J., Stoica, P., and Wang, Z. (**2003**), "On robust Capon beamforming and diagonal loading," IEEE Trans. Signal Process., **51**(7), 1702-1715.

Löllmann, H.W., Moore, A.H., Naylor, P.A., Rafaely, B., Horaud, R., Mazel, A., and Kellermann, W. (**2017**), "Microphone array signal processing for robot audition," Proc. HSCMA, 51–55.

Markovich, S., Gannot, S., and Cohen, I. (**2009**), "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," IEEE Trans. Audio, Speech, Lang. Process., **17**(6), 1071-1086.

Markovich-Golan, S. and Gannot, S. (**2015**), "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," Proc. ICASSP, 544–548.

Moore, A.H., Lightburn, L., Xue, W., Naylor, P.A., and Brookes, M. (**2018**), "Binaural mask-informed speech enhancement for hearing aids with head tracking," Proc. IWAENC.

Moore, A.H., Xue, W., Naylor, P.A., and Brookes, M. (**2019**), "Noise covariance matrix estimation for rotating microphone arrays," IEEE/ACM Trans. Audio Speech Lang. Process., **27**(3), 519-530.

Moore, A.H., de Haan, J.M., Pedersen, M.S., Naylor, P.A., Brookes, M., and Jensen, J. (**2019**), "Personalized signal-independent beamforming for binaural hearing aids," J. Acoust. Soc. Am., **145**, 971–2981.

Rafaely, B. (**2015**), *Fundamentals of Spherical Array Processing*, ser. Springer Topics in Signal Processing. Berlin Heidelberg: Springer-Verlag, 2015.

Schwartz, O., Gannot, S., and Habets, E.A.P (**2016**), "Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm," Proc. EUSIPCO, 1123–1127.

Schwarz, A., and Kellermann, W. (**2015**), "Coherent-to-diffuse power ratio estimation for dereverberation," IEEE/ACM Trans. Audio Speech Lang. Process., **23**(6), 1006–1018.

Taal, C.H., Hendriks, R.C., Heusdens, R., and Jensen, J. (**2011**), "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process., **19**(7), 2125-2136.

Tamai, Y., Kagami, S., Amemiya, Y., Sasaki, Y., Mizoguchi, H., and Takano, T. (**2004**), "Circular microphone array for robot's audition," Proc. IEEE Sensors, 565–570.

Thiergart, O., and Habets, E.A.P. (**2013**), "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," Proc. ICASSP, 659–663.

van Trees, H.L. (**2002**), *Optimum Array Processing*, ser. Detection, Estimation and Modulation Theory. John Wiley & Sons, Inc., 2002.

Yilmaz, O., and Rickard, S. (**2004**), "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., **52**(7), 1830-1847.

How to compare hearing-aid processing of real speech and a speech-modified stimulus for objective validation of hearing-aid fittings?

SØREN LAUGESEN[1,*]

[1] *Interacoustics Research Unit, Kgs. Lyngby, Denmark, DK-2800 Lyngby, Denmark*

A method is proposed to evaluate whether a modern hearing aid with all automatic features enabled processes a speech-modified stimulus for objective validation of hearing-aid fittings as speech. The proposed method measures short-term coupler gains from brief snippets of steady-state probe noise, crossfaded into either the International Speech Test Signal (ISTS) or the speech-modified stimulus, which thus act as conditioning signals. For reference, the method is also applied to a steady-steady noise signal, which drives the hearing aids into noise mode. Results for a selection of hearing aids show that the method classifies the hearing aids' mode of processing according to expectations, with all three conditioning signals.

INTRODUCTION

For the purpose of validating hearing-aid fittings in prelingual infants, an objective assessment based on the auditory steady-state response (ASSR) is considered. One important aspect of an appropriate assessment is to ensure that speech-relevant gain and signal-processing features are activated in the hearing aids during the measurement. To avoid modifying the hearing-aid settings for the validation measurement, a family of speech-modified ASSR stimuli has been devised. The preferred member of this stimulus family consists of three bandlimited CE-Chirps® (Elberling and Don, 2010) presented at different repetition rates, individually modified by frequency-band specific envelopes derived from the International Speech Test Signal (ISTS; Holube *et al.*, 2010), and scaled in level to match the long-term ISTS band levels as described by Laugesen *et al.* (2018).

Prior to testing, it needs to be verified that the speech-modified ASSR stimulus in fact drives the hearing aid into speech mode. As a benign example, the Genie fitting software for e.g. the Oticon Alta and Sensei hearing aids offers 'Live Demonstration' of the current classification of the incoming soundscape, which in tests clearly indicates that the speech-modified ASSR stimuli are processed as if they were speech, whereas CE-Chirp stimuli without speech-modifications are classified as noise. However, most hearing aids' fitting tools do not offer such feedback about mode of processing and therefore a 'black-box' measurement method is proposed in this paper. Under the assumption that the ISTS properties applied to the speech-modified ASSR stimuli are used by the hearing aids for detecting speech, the method does not require any specialist hearing-aid brand knowledge and can therefore be used broadly to verify correct processing of speech-modified ASSR stimuli also in the clinic.

*Corresponding author: slau@iru.interacoustics.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

MATERIAL

The hearing aids tested were Oticon Alta Pro, Sensei Pro Power, and Opn 1; GN Resound LiNX 3D 9 and Quattro; and Phonak Sky B90-M and B90-P. They were all programmed according to standard audiograms from (Bisgaard *et al.*, 2010), either N3 (moderate hearing loss) or N5 (severe hearing loss) as reported below, using the respective fitting software's suggested prescription, except that WhistleBlock was enabled and SoundRecover2 was disabled for the Sky aids (initially). As already mentioned, the Genie fitting software for the Alta and Sensei aids allows for feedback about the processing mode; this was not available for the other aids.

All measurements were taken in an Interacoustics TBS25 test box. The input to the hearing aid was recorded with a G.R.A.S. 40BL ¼" microphone, located next to the hearing aid's microphone inlets at the reference position in the test box. The output was recorded with a G.R.A.S. RA0045-S1 ear simulator to which the hearing aid was connected as appropriate for the type of aid (Sensei, LiNX 3D, and Sky behind-the-ear with sound tube; Alta, Opn 1, and Quattro receiver-in-the-ear with closed 10-mm domes). Playback and recording were accomplished through an RME Fireface UC soundcard connected to a standard laptop running custom Matlab software. Microphone sensitivities were calibrated relative to a G.R.A.S. 42AB Sound Calibrator prior to recording. Furthermore, the playback chain frequency response (considering both magnitude and phase) was equalised with a 16384-tap FIR filter.

The specific stimuli used were the ISTS, a steady-state noise (SSN_{ISTS}) delivered together with the ISTS having the same spectrum, and the 3B ISTS-modified CE-Chirp®. The latter consists of a two-octave wide low-frequency chirp centred at 707 Hz, and two one-octave wide chirps centred at 2000 and 4000 Hz. The repetition rates for the three chirps were 38.1, 68.4, and 69.3 Hz, respectively.

METHOD

The family of speech-modified ASSR stimuli is based on different combinations of the narrow-band (NB) CE-Chirps® (Elberling and Don, 2010), which are one-octave wide chirps centred at 500, 1000, 2000, and 4000 Hz. Accordingly, the current method of hearing-aid evaluation compares the reference gain applied to the ISTS with the gain applied to the speech-modified ASSR stimulus in matching one-octave analysis bands. It is thus assumed that all hearing aids will classify the ISTS as speech in agreement with current standards (IEC 60118-15, 2012; IEC 61669, 2015).

To determine the time evolution of gain for each signal (and in each frequency band), inspiration was taken from Naylor and Johannesson (2009) who determined hearing-aid gain trajectories from time-aligned input and output signals in 10-ms time windows. The short windows were selected to capture the fastest gain-modifying behaviour of the hearing aids: wide dynamic-range compression or output limiting. Naylor and Johannesson compared time-averaged gain values across different settings of hearing-aid compression for speech and noise mixed at different signal-to-noise ratios. Accordingly, the original idea of the present investigation was to compare such gain trajectories measured with either the ISTS or the speech-modified ASSR stimulus

passed through the hearing aid, assuming they would be very similar if the hearing aid was in speech mode. However, because the detailed waveforms of the two signals are markedly different when evaluated in 10-ms time windows, as can be seen from Figure 1 below, a direct comparison turned out to be difficult, even when the (Oticon Alta) hearing aid was in speech mode for both signals, according to the fitting software.

Evaluation by probe snippets

Instead, short probe snippets of the SSN_{ISTS} are crossfaded into the two signals. Thus, the ISTS and the speech-modified ASSR stimulus are merely used as conditioning signals, whereas the actual gain comparison is based on the probe snippets. This approach assumes that the probe snippets are short enough and occur rarely enough that the hearing aid remains in speech mode throughout the recording. This assumption is justified as all modern hearing-aid noise-reduction systems have a built-in activation sluggishness to avoid overly rapid switching back and forth between processing modes. For the same reason, the conditioning signal will be allowed to run without probe snippets for a while to ensure the hearing aid has settled into a stable processing mode. Thus, the method is characterised by four parameters, where T_S is the settling time (with unmodified conditioning signal), T_P is the duration of the probe snippets, and T_I is the time interval between successive probes, see Figure 1. The fourth parameter, N_T, denotes the total number of probe snippets included. Raised-cosine gates with 1-ms rise and fall times were used to crossfade between conditioning signal and probe snippets. At each probe interval, the exact same probe snippet is used with each conditioning signal, whereas different snippets are used at successive intervals. The selection of the parameter values (T_S, T_P, T_I, and N_T) is a compromise among (i) obtaining enough probe-signal recording time for a reliable gain evaluation, (ii) ensuring that the hearing aid remains in speech mode irrespective of the probe snippets, and (iii) total measurement time. This compromise will be explored below.

For further reference, measurements were also made with the SSN_{ISTS} as the conditioning signal. It was verified that this signal drove the Oticon Alta and Sensei aids into noise mode, allowing the effects on hearing-aid gain to be observed.

Time alignment

In contrast to Naylor and Johannesson (2009), who examined mild non-linearities related to dynamic-range compression, most modern hearing aids pose an additional challenge as they comprise highly non-linear signal processing such as frequency-shifting for suppressing acoustic feedback. This means that classical linear cross-correlation methods for time alignment of input and output signals from the hearing aid are useless, typically from around 1000 Hz and upwards in frequency where the frequency-shifting is employed. Consequently, the time alignment for the present method is based solely on the octave-band filtered signals centred at 500 Hz. In practice, the time alignment is done between the digital stimulus and the recorded input and output from the hearing aid, respectively. In this way, the probe snippets can be isolated from both the recorded input and output signals for analysis.

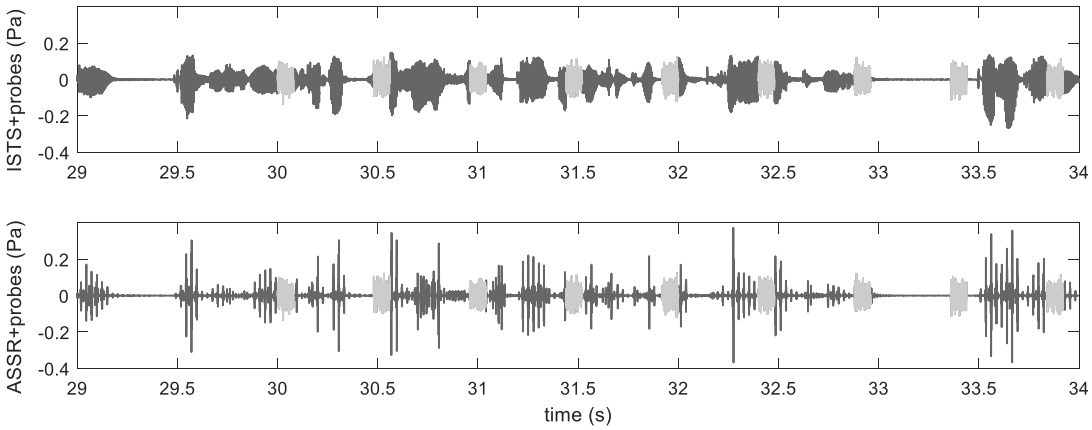


Fig. 1: Measurement signal excerpts with probe snippets indicated by light grey crossfaded into the ISTS (top) and the 3B ISTS-modified CE-Chirp ASSR stimulus (bottom panel). The first probe snippet occurs at T_S = 30 s, the probe duration is T_P = 80 ms, and the between-probe interval is T_I = 0.4 s.

Gain comparison metric

The respective sequences of probe snippets in each analysis frequency band and at the input and output of the hearing aid are first concatenated. Then, coupler-gain trajectories are computed as successive ratios of output and input RMS values taken across 10-ms rectangular windows in each frequency band, see Figure 2 for examples measured in the Oticon Alta and the Phonak Sky B90-M aids. Note that the examples in Figure 2 deliberately were measured with the settling time set to zero and a high number of probes in order also illustrate the course of noise-reduction activation. In addition, an extreme setting with noise-reduction parameters set to maximum values was chosen for the Sky aid to illustrate the possible range of gain reduction.

There are several important observations to make from Figure 2. First, the gain trajectories for the ISTS and the speech-modified ASSR stimulus overlap in all cases, as expected. The gain trajectories for the SSN_{ISTS} signal shows rather different activation profiles for the two aids: the Alta aid's noise reduction is fully activated after n = 160 analysis windows, which translates to 9.6 s (8 analysis windows per 80-ms probe and 0.4 s between-probe intervals, $\frac{160}{8}(0.08\text{ s}+0.4\text{ s})=9.6\text{ s}$), while the Sky noise reduction is activated in two stages with full activation after n = 200 ~ 12 s. When fully activated, the Alta aid provides about 5 dB of gain reduction at 500 Hz and almost none at 4000 Hz, while the Sky aid provides about 17 and 10 dB gain reduction at the two analysis frequencies. The exact numbers will possibly depend on the precise orientation of the hearing aid in the test box, since adaptive directionality may be activated by the SSN_{ISTS} signal. (This was, however, not the case for the Alta aid as observed through the fitting software.)

To form a single-valued metric of gain comparison between two signals A and B (with added probe snippets and appropriate settling time), median coupler gain values in dB are determined in each frequency band and for each of the gain trajectories, $g_{Ak}(n)$ and

$g_{Bk}(n)$, where n denotes the analysis time-window index and k denotes the frequency-band index. Thus, the proposed gain comparison metric is the maximum absolute difference in median gain across the four frequency bands:

$$C = \max_k |\mathrm{med}\{g_{Ak}(n)\} - \mathrm{med}\{g_{Bk}(n)\}|. \qquad \text{(Eq. 1)}$$

Inspired by the results presented in Tables 1 and 2 below, the proposed criterion value is C_{crit} = 1 dB. That is, if C is below 1 dB, the two signals A and B are considered as being processed similarly and *vice versa*. A gain error of 1 dB is small compared with the expected variation from practical sound-field measurements.

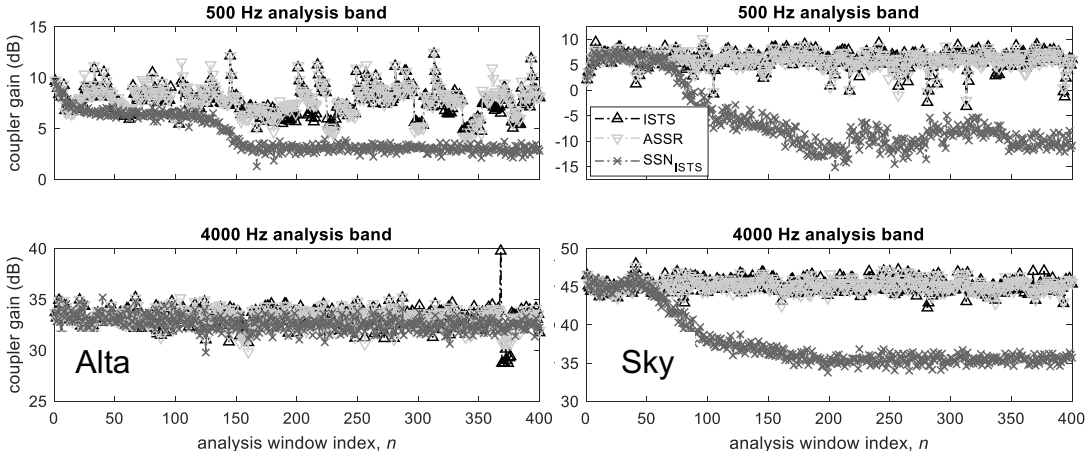


Fig. 2: Gain trajectories for the 500 and 4000-Hz octave bands and 10-ms analysis windows indexed by n, measured in the Oticon Alta (left panels, N3 hearing loss) and Phonak Sky B90-M (right panels, N3 loss). Probe parameters as in Figure 1, except the settling time which was set to T_S = 0. Note the different ordinate-scale ranges in all panels.

RESULTS

The results in Figure 2 as well as data from Bentler and Chiou (2006) suggest that a fail-safe choice of settling time is T_S = 30 s. In this way, the proposed test method will be able to fulfil its purpose, by allowing the hearing aid under test enough time to fully activate noise reduction, if – contrary to expectations – the speech-modified ASSR stimulus is classified as noise by the hearing aid. Next, the selection of the remaining analysis parameters is considered. The results across different parameter selections are compiled in Table 1, in terms of the gain comparison metric, C from eq. 1, for comparisons between ISTS and the 3B ISTS-modified CE-Chirp® ASSR stimulus, as well as between the ISTS and SSN_{ISTS}. In addition, the total recording time for one signal is stated. The measurements for Table 1 were all done with the Oticon Alta aid, which allowed the ground truth speech/noise classification to be examined through the fitting software, as described above. The results in Table 1 show remarkable robustness of the method towards variation in the analysis parameters. With all parameter combinations, the classification of the ASSR and SSN_{ISTS} signals is correct according to the suggested 1-dB criterion value. Varying the probe length, T_P, from

40 to 160 ms has no effect, which is to be expected from the results in Figure 2 where the onset of noise reduction is observable only from about $n = 40 \sim 2.4$ s. This assumes that the switching mechanism is reset as soon as the probe snippet stops. This seems to be the case for the Alta aid, since lowering the probe interval, T_I, from 4 s down to as little as 40 ms had only marginal effect on the results in Table 1. For the two conditions with $T_I = 40$ ms, the fitting software would very occasionally and briefly change the classification from 'Speech' to 'Speech in noise' during the ISTS and ASSR recordings. In all the other conditions the classification remained stable at 'Speech'. The SSN_{ISTS} was always classified as 'Noise'. Finally, the number of probe snippets, N_P, was varied with no observable effect to the results. Further reduction of N_P was not considered, since the total measurement time was already dominated by the 30-s settling time with $N_P = 30$. The results in Table 1 allows for selecting parameters with a broad safety margin in consideration of hearing aids potentially keener to switch to noise mode. Thus, probe duration was set to $T_P = 80$ ms, not exceeding typical syllable duration, and the between-probe interval was set to $T_I = 0.4$ s, (as used in Figures 1 and 2). Using these parameters, additional hearing aids and alternative settings were tested, with the results shown in Table 2.

T_P	T_I	N_P	T_{total}	C (ISTS vs. ASSR)	C (ISTS vs. SSN_{ISTS})
40 ms	4 s	120	511 s	0.8 dB	4.2 dB
80 ms	1 s	60	94 s	0.5 dB	4.1 dB
80 ms	0.4 s	30	44 s	0.4 dB	4.5 dB
80 ms	40 ms	60	37 s	0.4 dB	3.2 dB
160 ms	40 ms	30	36 s	0.2 dB	3.2 dB

Table 1: Total recording time for one signal, T_{total}, and the gain comparison metric, C, for two signal comparisons and various combinations of probe-method parameters. All measurements were done with the Oticon Alta hearing aid and a settling time of $T_S = 30$ s.

The results in Table 2 show the expected classification in all cases considered, that is, the gain differences between the ISTS and the speech-modified ASSR stimulus is below criterion value in all cases, whereas gain reductions above criterion value are observed for SSN_{ISTS} in all cases. The Opn 1, LiNX 3D, and Sky aids allow considerable user-defined changes to the noise-reduction (NR) parameters. Thus, in addition to measuring with the default settings described above, the NR parameters were set to maximum values, denoted NR max in Table 2. For the Opn 1 aid, settings were additionally brought way beyond what was achievable through the Genie 2 fitting software (Zaar *et al.*, 2020); this setting is denoted NR++. The grossly non-linear frequency-shifting features, Speech Rescue (SR) in Opn1, Sound Shaper (SS) in LiNX 3D, and SoundRecover2 (SR2) in Sky, were also enabled for separate

measurements. In all these extreme cases, the classification was correct according to the results in Table 2. Finally, repeat recordings were made for the Sky B90-M aid showing very minor variations in *C*-values within ±0.1 dB, which further testifies to the robustness of the method (results not shown in Table 2).

Hearing aid	**Setting**	***C*** (ISTS vs. ASSR)	***C*** (ISTS vs. SSN_{ISTS})
Oticon Sensei	N5	0.1 dB	4.7 dB
Oticon Opn 1	N3	0.5 dB	5.3 dB
Oticon Opn 1	N3, NR max	0.9 dB	13.2 dB
Oticon Opn 1	N3, NR++	0.6 dB	10.9 dB
Oticon Opn 1	N3, SR on	0.4 dB	1.5 dB
GN Resound LiNX 3D	N5	0.1 dB	4.3 dB
GN Resound LiNX 3D	N5, NR max	0.4 dB	16.6 dB
GN Resound LiNX 3D	N5, SS on	0.1 dB	4.2 dB
GN Resound Quattro	N3	0.2 dB	12.3 dB
GN Resound Quattro	N5	0.1 dB	11.0 dB
Phonak Sky B90-P	N5	0.3 dB	4.6 dB
Phonak Sky B90-P	N5, NR max	0.3 dB	10.0 dB
Phonak Sky B90-P	N5, SR2 on	0.3 dB	4.6 dB
Phonak Sky B90-M	N3	0.2 dB	7.3 dB
Phonak Sky B90-M	N3, NR max	0.3 dB	16.5 dB

Table 2: The two gain comparison metrics measured with the proposed method for a selection of hearing aids and feature settings, see text for details.

GENERAL DISCUSSION

The evidence presented above indicates that the proposed method for evaluating whether a speech-modified ASSR stimulus is processed as speech by a hearing aid works robustly and as intended, across a selection of modern hearing aids, and with no need for prior knowledge about the settings of the hearing aid. In this way, it will potentially be possible to assess in the clinic whether a given hearing aid can be used with an aided ASSR protocol for hearing-aid validation in infants, without any modifications to the hearing aid's settings. This will ensure clinical expedience and add to the face validity of the aided ASSR test.

The proposed method will continue to be verified against so far untested hearing-aid brands as well as new hearing-aid models brought to the market. One critical factor is expected to be the time-alignment, which currently assumes quasi-linear processing within the 500-Hz octave band. In this regard it should be noted that the frequency-warped filter-bank design used in the GN Resound aids created no problems despite its varying delay across frequency. Another potential challenge is future generations of hearing aids using soundscape classification based on deep neural networks, for which the methods of detection will be opaque.

Besides demonstrating the robustness of the proposed classification method, as well as the effectiveness of the speech-modified ASSR stimulus in driving the tested hearing aids into speech mode, the results in Figure 2 and Table 2 serve to illustrate the potential gain-measurement errors if due care is not taken to bring the hearing aid into speech mode for an aided ASSR recording. Thus, gain errors up to 16.5 dB were observed, which would seriously confound a measurement intending to validate a hearing-aid fitting. According to the results presented, this can be avoided by using the speech-like 3B ISTS-modified CE-Chirp® stimulus for aided ASSR.

REFERENCES

Bentler, R. and Chiou, L.-K. (**2006**). "Digital noise reduction: an overview," Trends Amplif., **10**, 67–82.

Bisgaard, N., Vlaming, M. S. M. G., and Dahlquist, M. (**2010**). "Standard audiograms for the IEC 60118-15 measurement procedure," Trends Amplif., **14**, 113–120.

Elberling, C. and Don, M. (**2010**). "A direct approach for the design of chirp stimuli used for the recording of auditory brainstem responses," J. Acoust. Soc. Am., **128**, 2955-2964.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). "Development and analysis of an International Speech Test Signal (ISTS)," Int. J. Aud., **49**, 891–903.

IEC 60118-15. (**2012**). "Electroacoustics – Hearing aids – Part 15: Methods for characterising signal processing in hearing aids with a speech-like signal," IEC International Standard.

IEC 61669. (**2015**). "Electroacoustics – Measurement of real-ear acoustical performance characteristics of hearing aids," IEC International Standard.

Laugesen, S., Rieck, J. E., Elberling, C., Dau, T., and Harte, J. M. (**2018**). "On the cost of introducing speech-like properties to a stimulus for auditory steady-state response measurements," Trends Hear., **22**, 1-11.

Naylor, G. and Johannesson, R. B. (**2009**). "Long-term signal-to-noise ratio at the input and output of amplitude-compression systems," J. Am. Acad. Audiol., **20**, 161–171.

Zaar, J., Simonsen, L. B., Behrens, T., Dau, T., and Laugesen, S. (**2020**). "Investigating the relationship between spectro-temporal modulation detection, aided speech perception, and directional noise reduction preference in hearing-impaired listeners," Proc. ISAAR, **7**, 181-188.

Benefit from different beamforming schemes in bilateral hearing aid users: Do binaural hearing abilities matter?

MATTHIAS LATZEL[1,*], KIRSTEN C. WAGENER[2], MATTHIAS VORMANN[2] AND TOBIAS NEHER[3]

[1] *Sonova AG, CH-8712 Stäfa, Switzerland*

[2] *Hörzentrum Oldenburg, D-26129 Oldenburg, Germany*

[3] *Institute of Clinical Research, University of Southern Denmark, DK-5230 Odense, Denmark*

Using a hearing aid simulator and virtual acoustics, Neher *et al*. (2017) recently showed that binaural hearing abilities influence speech-in-noise reception through different bilateral directional processing schemes. The current study aimed to extend this finding to real acoustic environments and commercial devices. Three beamforming schemes were tested – they differed in signal-to-noise ratio (SNR) improvement and binaural cue preservation. The participants were 38 elderly experienced hearing aid users. Speech understanding and localisation performance were measured. Binaural hearing abilities were assessed using the binaural intelligibility level difference (BILD). The analyses revealed a clear effect of the BILD on speech understanding in noise, but no interaction with the beamformer conditions. Greater SNR improvement was generally beneficial. In contrast, localisation of static and dynamic stimuli was more accurate when low-frequency binaural cues were preserved. Furthermore, the interaction with the BILD was marginally significant for dynamic stimuli (p = 0.054). Altogether, these results suggest that when selecting directional processing schemes in bilateral hearing aid fittings both speech understanding and aspects of spatial awareness perception should be considered.

INTRODUCTION

Almost all hearing aids (HAs) comprise directional microphones as directionality is the only feature that improves speech intelligibility (Dillon, 2012). Although considerable effort is dedicated to parameterizing these systems to provide the optimal benefit to the intended target population, the acceptance and benefit of the so-called "FirstFit" vary remarkably across the individual users (Gatehouse *et al.,* 2003; Lunner, 2003). Thus, it is of interest to investigate how HA settings can be better tailored to the individual needs and medical/audiological parameters of the user. There have been a number of investigations looking into several factors and their ability to explain individual differences in HA outcomes, but recommendations for translating this information into a meaningful prescribed fitting are rather rare.

*Corresponding author: matthias.latzel@sonova.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

In a recent study, Neher *et al.* (2017) focused on individual binaural hearing abilities determined through the binaural intelligibility level difference (BILD; Kollmeier, 1996). They investigated how individual binaural hearing abilities play a role to understand speech-in-noise and how this information correlates with the setting of directional microphone systems in HAs. They found that the BILD was correlated with speech perception in situations with lateral interferers, while in spatially diffuse situations the speech perception was driven by the signal-to-noise ratio (SNR) improvement (or directivity index) provided by the different beamformer technologies independent of the BILD values. Together, these findings provide a base for adapting directional processing to the HA user, its binaural hearing abilities, and the acoustic scenario.

The current study investigated the extent to which the results of Neher *et al.* (2017) can be transferred to directional processing strategies used in commercially available hearing aids. Thereby, real acoustical coupling, real acoustical scenes and the possibility of head movements are of relevance. In addition to speech reception in noise, it was investigated how binaural cue preservation in the different directional processing schemes affects aspects related to spatial awareness.

METHODS

Participants

In the current study, 38 experienced HA users (16 women) with an average age of 74.7 yrs (range: 63-82 yrs) and moderate-to-severe bilateral hearing losses participated. All of them had participated in the Neher *et al.* (2017) study. The participants were divided into two groups according to their binaural hearing abilities as assessed using the BILD measure: BILD < 2.5 dB ('BILD–'; $N = 18$) and BILD ≥ 2.5 ('BILD+'; $N = 20$). The individual BILD values were equally distributed between the minimum value of -0.4 dB and maximum value of 5,2 dB. The two groups were balanced in terms of four-frequency pure-tone average hearing loss (55 dB HL and 51 dB HL, respectively). All 38 participants completed a set of speech-in-noise measurements, similar to those performed by Neher *et al*. (2017). A subset of 26 paricipants (9 women) with the same mean age, hearing loss and distribution of BILD values completed a set of additional spatial awareness measurements (see below).

HA conditions

The participants were fitted with Phonak Audéo V90-312 devices using the xP-receiver with closed sShells and a flat real-ear-to-coupler difference to maximize the acoustic differences between the different beamformer settings. For gain prescription, a modified version of the NAL-NL1 fitting rule (Dillon, 1999) was used with linear amplification based on the gain prescribed for 65 dB input level of NAL-NL1. To ensure adequate audibility and comparability with the results of Neher *et al.* (2017), a minimum gain of 6 dB was defined in the frequency range from 250 Hz to 500 Hz. This was verified using real-ear measurements. The beamformer settings were all

steered in the 0° direction and set at non-adaptive. Five settings were tested (see polar patterns in Figure 1):

(1) Real Ear Sound (RES): A commercially available beamformer setting simulating the pinna effect (Latzel, 2013) of the outer ear with a small degree of directivity (mean directivity index, DI: -1,0 dB) > 1 kHz. Output: Dichotic stimulus with binaural cues preservation over the entire frequency range.
(2) UltraZoom (UZ): A commercially available unilateral beamformer setting (Latzel, 2013) providing SNR improvement over the whole frequency range (mean DI: 2,3 dB). Output: Dichotic stimulus with binaural cue preservation over the entire frequency range.
(3) StereoZoom (SZ): A commercially available bilateral beamformer setting (Latzel, 2013) providing SNR improvement at frequencies < 2 kHz (mean DI: 4.7 dB). Output: predominantly diotic stimulus < 2 kHz and dichotic stimulus above.
(4) StereoZoom INV (SZ-inv): An experimental beamformer setting based on SZ that provides SNR improvement > 800 Hz (mean DI: 4.2 dB). Output: Dichotic stimulus < 800 Hz and diotic stimulus above.
(5) FullBeam (FB): An experimental beamformer setting based on SZ that provides SNR improvement over the whole frequency range (mean DI: 4.9 dB). Output: Diotic stimulus over the entire frequency range (no binaural cue preservation).

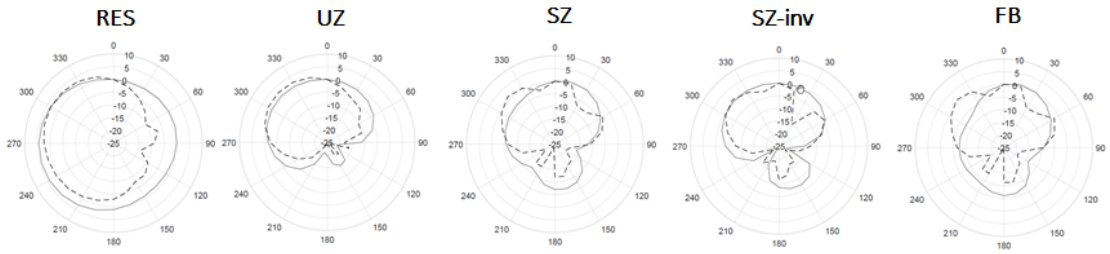


Fig. 1: Polar patterns of the five beamformer settings (left ear) calculated in octave bands with centre frequencies of 517 (low frequencies, solid line) and 1981 (high frequencies, dashed line). The azimuth is in degrees and the gain in dB.

Acoustic scenarios and speech-in-noise measurements

The different beamformer conditions were tested in two different acoustic scenarios. In both cases, the Oldenburg sentence test (OLSA; Wagener *et al.*, 1999) was performed with the target speech presented from 0° and 1-m distance. The participants' task was to repeat as many of the five words per sentence as possible. For the background noise, two different masker scenarios were implemented:

(1) *Lateral interferers*: 10 sentences of a male speaker of an alternatively recorded OLSA (Hochmuth *et al.*, 2015) were concatenated without any pauses and presented from two loudspeakers placed at ±60°. To ensure that different sentences were played from both speakers, an offset of about 9s between the speakers was applied.

(2) *Diffuse interferer:* Recording made in a large cafeteria (T_{60} = 1.25 s) during a busy lunch hour (Kayser *et al.*, 2009) presented through 11 loudspeakers placed around the participant in 30° steps (excluding 0°).

For each combination of acoustic scenario and beamformer condition, two speech reception threshold (SRT) measurements were determined per participant. A correlation analysis revealed high test-retest reliability (r = .81, p < .0001). For the statistical analyses, the average SRT per condition was used.

Spatial awareness measurements

In addition to the speech-in-noise measurements, sound localization was assessed using both static and dynamic stimuli. A traffic-junction scene was simulated using TASCAR (Toolbox for Acoustic Scene Creation and Rendering) (Grimm *et al.*, 2015). The overall level of the noise scenario was 60.2 dB SPL at the listening position. The total length of the traffic-junction scene was 360 s. Within this scene, different target stimuli were presented at random time intervals:

(1) *Static localization*: Barking dog 1 (length: ~3.7 s; overall level: 68.4 dB SPL) or barking dog 2 (length: ~3.0 s; overall level: 68.5 dB SPL), placed at different angles (0°, ±45°, ±60°, ±75° or ±90°), presented twice per angle.
(2) *Dynamic localization*: Ambulance (length: ~4.5 s; angular velocity: ~13°/s; overall level: 68.3 dB SPL) or car horn (length: ~4.4 s; angular velocity: ~13°/s; overall level: 68.5 dB SPL), moving on a circle around the subject (-90°➔-30°; -30°➔+30°; +30°➔+90°; +90°➔+30°; +30°➔-30°; -30°➔-90°), with constant velocity. Three measurements per target movement were conducted.

The task of the participant was to pay attention to the different signals by turning the head into the direction of the source (static signals) or following the sources by moving the head synchronously (moving signals). The trajectories were recorded using a head tracker. The order of the presentation angles/trajectories was varied randomly. The order of the two localization tasks/stimuli was randomized across participants.

RESULTS

Speech reception in noise

Do binaural hearing abilities and/or the masker scenarios correlate with speech intelligibility in noise? To answer this question the data were grouped either according to BILD+ or BILD- or to the masker scenario. Figure 2 shows the average SRTs for all beamformer conditions either pooled according to *masker scenario* (Figure 2 right) or according to the *BILD* (Figure 2 left). Figure 2 left shows a clear difference in speech intelligibility between BILD+ and BILD-. Figure 2 right shows a different pattern of speech intelligibility performance depending on the different beamformer conditions. When the noise scenario was diffuse the binaural beamformer conditions were better than for the noise scenario with lateral interferers: the better the DI the better the speech intelligibility in a diffuse noise scenario. A three factor ANOVA revealed a significant main effect of the *BILD* (p < .000), *masker scenario* (p < .005)

and the *beamformer conditions* ($p < 0.05$). A posthoc test (Bonferroni corrected) showed that SZ is statistically significantly better than all other beamformer conditions regardless of the binaural hearing abilities (SZ⇔ RES ($p < .01$), SZ⇔ UZ($p < .01$), SZ⇔ SZ-inv ($p < .05$) SZ⇔ FB ($p < .05$)). In Figure 3 the individual SRT data are visualized according to the BILD values with a regression line plotted for each beamformer condition. For the *diffuse interferer* scenario, the regression lines are spread for small BILD values but much narrower for large BILD values. This suggests that for participants with good binaural hearing abilities the selection of the beamformer is not relevant for speech intelligibility in noise and should be individually selected based on other parameters (see section environmental awareness test). For participants with poor binaural hearing abilities, the beamformer providing the highest DI values should be selected as it allows for better speech intelligibility performance. Almost the opposite trend can be observed for the *lateral interferer* scenario: the choice of beamformer is not relevant for participants with poor binaural hearing abilities, as the listeners do not benefit from binaural cues anyway. Therefore, other parameters are potentially more relevant to select the most effective individual beamformer condition. For participants with good binaural hearing abilities, the beamformer condition that preserves most binaural cues showed the best speech intelligibility.

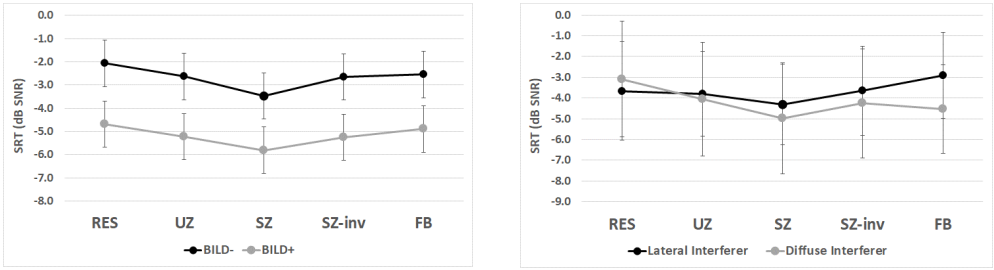


Fig. 2: Left panel: Average (standard deviation) speech reception thresholds (SRT) in noise for beamformer conditions grouped according to binaural hearing abilities: BILD < 2,5 dB (BILD-, black), BILD ≥ 2,5 dB (BILD+, grey). Right panel: Average (standard deviation) speech reception thresholds (SRT) in noise for beamformer conditions grouped according to masker scenarios: Lateral interferer (black), diffuse interferer (grey).

Spatial awareness measurements

Static localization: The localization ratings for both static signals were averaged for analyses since no statistically significant differences was detected between the ratings for a given presentation angle. To condense the results, the localization errors were analysed depending on the angle deviation from the front, independent of hemisphere since no statistically significant differences were detected for the same angle deflections in either left or right direction from the front. Figure 4 (left) shows the distribution of localization errors for all beamformer conditions after this data condensation.

Static localization: The localization error was analysed by a repeated measures two factor ANOVA with the factors *beamformer condition* and *presentation angle*. Analyses revealed significant main effects of *beamformer condition* ($p < .001$) and *presentation angle* ($p < .001$). In addition, a significant interaction of *beamformer condition* and *presentation angle* was detected ($p < .001$). A post hoc test (Bonferroni corrected) revealed that all *beamformer conditions* were significantly different from SZ and FB where SZ also performed significantly different from FB (all $p < .05$). RES, UZ, and SZ-inv were not significantly different. The post hoc analysis for *presentation angle* revealed that 0° and 45° both significantly differed from 75° and 90°, as well as that 60° significantly differed from 90°. All other data were not significantly different. The between-subject factor *BILD* was not statistically significant.

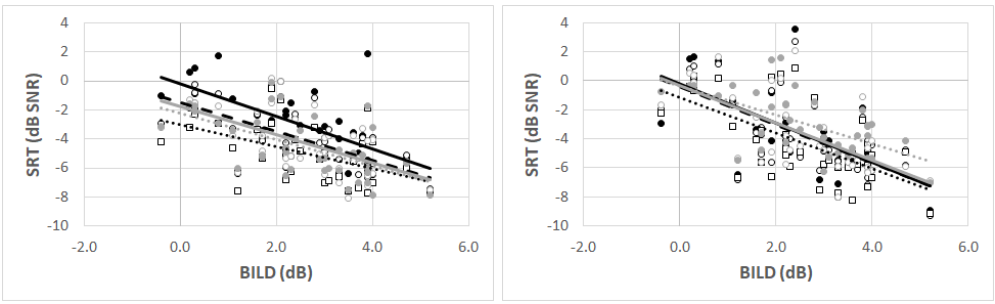


Fig. 3: Scatterplot of the BILD and SRT data. Left panel: diffuse interferer, right panel: lateral interferer. Least square regression lines corresponding to RES (black solid line, filled black circles), UZ (dashed black line, unfilled black circles), SZ (dotted black line, unfilled black squares), SZ-inv (solid grey line, unfilled grey circles) and FB (dotted grey line, filled grey circles). SRT is SNR in dB

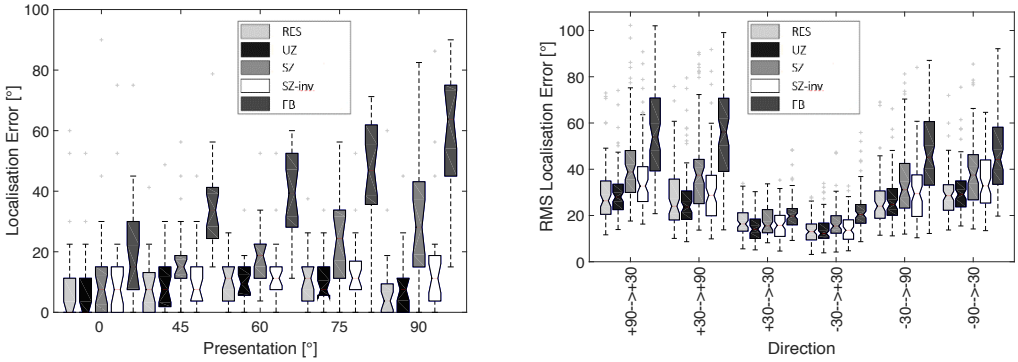


Fig. 4: Left panel: Box plots of the static localization errors in % for beamformer conditions and presentation angles. Right panel: Box plots of the RMS dynamic localization errors in % for beamformer conditions and trajectories.

Dynamic localization: The head movements of the participants were recorded as individual trajectories per participant and per measurement condition. In case of any discontinuities or missing data these were replaced by continuous completion. In addition, the individual trajectories were resampled (125Hz). From these trajectories the differences between the perceived angles and target angles were calculated resulting in the outcome measure RMS localization error (see Figure 4 right). For further analysis, the three measurement repetitions per target movement were averaged and analysed by a repeated measures two factor ANOVA with the factors *beamformer condition* and *direction* and individual *BILD* as covariate. Statistically significant main effects were revealed for *beamformer condition* ($p < .001$) and *target movement* ($p < .001$). In addition, an interaction trend was detected between *beamformer condition* and *BILD* ($p = 0.054$). A post hoc test for *beamformer condition* revealed that RES and UZ did not differ in RMS localization error but both showed significantly lower RMS localization errors than SZ-inv. SZ-inv yielded significantly lower RMS localization errors than SZ, that yield significantly lower RMS localization errors than FB (all $p < .05$). The post hoc analysis for *direction* revealed that the internal (-30°➔+30, +30°➔-30°) and external (all other directions) target movements significantly differed in terms of RMS localization error.

In summary the data showed that beamformer conditions with preservation of binaural cues at low frequencies provided better localisation performance for static and dynamic objects.

SUMMARY AND CONCLUSIONS

We investigated the influence of binaural hearing abilities and acoustic scenarios on speech intelligibility in noise and on spatial awareness for five different beamformer approaches that are available in commercial hearing devices. The main differences between the beamformer approaches were the preservation of binaural cues, especially at low frequencies, and the DI, the ability to emphasize the target source from the front.

The analyses revealed that speech intelligibility in noise depends on binaural hearing abilities, masker scenario and beamformer conditions. Listeners with poor binaural hearing abilities have worse speech perception in noise compared with listeners with good binaural hearing abilities. An interaction effect between masker scenario and beamformer was demonstrated as well, but there was no interaction effect between binaural hearing abilities and beamformer condition. A post hoc analysis revealed that the commercially available beamformer SZ outperformed all other beamformers, independent of masking scenario and binaural hearing abilities. This means that we could only partly replicate the results of Neher *et al.* (2017), a study that included the same participants. This may be due to some differences in the set-up of the study, such as allowing for real head movements, real acoustic scenarios, and real acoustic couplings. Additionally, the algorithms differed slightly from those in the Neher *et al.* study as the systems used here were already fine-tuned to be effective under real life conditions. Therefore, we could not emanate from a clear distinction between diotic and dichotic output of the hearing devices which was the case in the Neher *et al.*

study's experimental setup. This can explain the higher variances in this study. The additional measurements of environmental awareness revealed a clear advantage of algorithms that preserve the binaural cues at low frequencies when localizing static or moving sound sources in a noisy environment.

Together, these findings of the study provide a basis for adapting beamformer (settings) in commercial hearing devices to the individual binaural hearing abilities and the noise situation meaning that both speech understanding and aspects of spatial awareness perception should be considered.

REFERENCES

Dillon, H. (**1999**) "NAL-NL1: A prescriptive fitting procedure for non-linear hearing aids," The Hear J. **58**(10): 10-16.

Dillon, H. (**2012**). "Hearing Aids," 2nd ed., Boomerang Press, Sydney, Australia.

Gatehouse, S., Naylor, G., and Elberling, C. (**2003**). "Benefits from hearings aids in relation to the interaction between the user and the environment,' Int. J. Audiol. **42**(suppl.1). 77-85.

Grimm, G., Luberadzka, J., Herzke, T., and Hohmann, V. (**2015**). "Toolbox for acoustic scene creation and rendering (tascar): Render methods and research applications," In F. Neumann, editor, Proceedings of the Linux Audio Conference, Mainz, Germany, 2015. Johannes-Gutenberg Universitat Mainz.

Hochmuth, S., Jürgens, T., Brand, T., and Kollmeier, B. (**2015**). " Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation?," Int. J. Audiol. **54**:sup2, 23-34.

Kayser, H., Ewert, S.D., Annemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B. (**2009**). "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse response," EURASIP J. Adv. Signal Process., **298605**. 1-10.

Kollmeier, B. (**1996**). "Computer-controlled speech audiometric techniques for the assessment of hearing loss and the evaluation of hearing aids," In; Kollmeier, B. (Ed.) Psychoacoustics, Speech and Hearings Aids. World Scientific, Singapore, pp. 57-68.

Latzel, M. (**2013**). "StereoZoom and auto StereoZoom," Phonak compendium.

Lunner, T. (**2003**). "Cognitive function in relation to hearing aid use,' Int. J. Audiol. **42**(suppl.1), 49-58.

Neher, T., Wagener, K.C., and Latzel, M. (**2017**) "Speech reception with different bilateral directional processing schemes: Influence of binaural hearing, audiometric asymmetry, and acoustic scenario," Hear Res., **353**: 36-48.

Wagener, K., Brand, T., Kühnel, V., and Kollmeier, B. (**1999**) "Development and evaluation of a sentence test for the German language I-III: Design, optimization and evaluation of the Oldenburg sentence test," Z. Audiol. (Audiol. Acoustics). **38**. 4-15, 44-56, 86-95.

Feature-based audiovisual speech integration of multiple streams

JUAN CAMILO GIL-CARVAJAL[1,2,*], JEAN-LUC SCHWARTZ[3], TORSTEN DAU[2] AND TOBIAS SØREN ANDERSEN[1]

[1] *Cognitive Systems, DTU Compute, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Hearing Systems, DTU Health Tech, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[3] *GIPSA-lab, Univ, Grenoble Alpes, CNRS, Grenoble INP*

Speech perception often involves the integration of auditory and visual information. This is shown in the McGurk effect, in which a visual utterance, e.g., /ipi/, dubbed onto an acoustic utterance, e.g., /iki/, produces a combination percept, e.g., /ipki/. However, it is still unclear how phonetic features are integrated audiovisually. Here, we studied audiovisual speech perception by decomposing the auditory component of McGurk combinations into two streams. We show that auditory /i_i/, where the underscore indicates an intersyllabic silence, dubbed onto visual /ipi/ produce a strong illusion of hearing /ipi/. We also show that adding an acoustic release burst to /i_i/ creates a percept of /iki/. An auditory continuum was created with stepwise temporal alignments of the release burst and /i_i/. When dubbed onto /ipi/, this continuum was perceived mostly as a visually driven response /ipi/ when the burst overlapped with either acoustic vowel. Other temporal alignments frequently produced combination responses. Mostly /ikpi/ combinations were obtained when the burst was closer to the initial vowel, and reverse /ipki/ responses when it was closer to the final vowel. These results are indicative of feature-based audiovisual integration where burst and aspiration are sufficient cues for the consonant /k/, while the perception of /p/ depends on place information in the visual stream.

INTRODUCTION

The visible facial gestures accompanying the voice of the talker in face-to-face conversations facilitate speech perception (Sumby and Pollack, 1954). This is particularly advantageous in noisy listening situations (Binnie *et al.*, 1974). However, it is still unclear how phonetic information is integrated. The McGurk effect demonstrates phonetic integration for speech comprehension (McGurk and McDonald, 1976). Here we study the McGurk *combinations*, in which the audiovisual pairing of an auditory non-labial consonant (e.g., /iki/) and a visual labial consonant (e.g., /ipi/) leads to a cluster percept in which both consonants are represented (e.g., /ipki/ or /ikpi/). A typical finding in many studies that have reported the perceived consonant order of the combination response is that the labial consonant leads the non-

*Corresponding author: juac@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

labial (e.g., Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009), although not always (Hampson *et al.*, 2003). It has also been suggested that the combination responses occurred more frequently with unvoiced consonants (Colin *et al.*, 2002). This could be due to the strong consonantal burst and aspiration that have been shown to increase the frequency of the combination responses (Green and Norrix, 1997). However, the role of these acoustic features on the perceived consonant order has remained unclear.

A few studies (e.g., Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009) varied the timing in combination illusions by altering the audiovisual stimulus onset asynchrony (SOA). This approach could be suboptimal for studying the perceived consonant order, since the cross-modal asynchrony could influence the strength of the integration in addition to the perceived consonant order. To minimize the effect of asynchrony on audiovisual integration, we chose to vary only the timing of the consonants such that the vowels were still synchronous across modalities. To do so, we isolated the burst and aspiration from the vowels of the auditory component. An auditory continuum was then created by pairing the vowels and the burst at nine SOAs. To test the effect of varying the timing of the visual articulatory gestures, we paired the auditory continuum with two visual contexts. This resulted in two audiovisual continua in which the visual consonant was pronounced either in the offset of the initial syllable or the onset of the final syllable in a vowel-consonant-vowel (VCV) context. Our hypothesis was that varying the timing of the consonantal burst and aspiration could systematically change the perceived consonant order of the combination response. If so, the combination responses could provide information about the temporal organization of the stimulus features integrated across modalities.

METHODS

Participants

The test subjects were 14 native French speakers (mean age 25, five female). All subjects reported to have normal hearing and normal or corrected-to-normal vision. Before the testing, all participants provided written consent and all experiments were approved by the Comité d'Ethique pour les Recherches Non Interventionnelles (CERNI), reference number IRB00010290.

Speech material

The speech material recorded consisted of the disyllables /i_i/, /ip_i/, /i_pi/, /i_ki/, /ikpi/, and /ipki/ articulated by a native French speaker. The underscore represents an intersyllabic silence, which corresponded to 300, 114, and 189 ms for /i_i/, /ip_i/ and /i_pi/, respectively. The speaker pronounced each syllable synchronously with two consecutive beats of a metronome, which was delivered through EarPods and set at 130 beats per minute. The idea was to provide an auditory reference that could enable the production of speech utterances with similar speaking rate and duration.

To create stimuli in which only the timing of the burst and aspiration was affected, we used two separate auditory streams. One stream contained only the vowels and

consisted of the recorded sound of /i_i/. The second stream contained only the consonantal burst and aspiration with a duration of 100 ms, which were extracted from the recorded articulation of /i_ki/. An auditory continuum was then generated offline by pairing the two streams at nine different SOAs with a step size of 50 ms. At one end (–200 ms), the waveform of the initial vowel fully overlapped with the burst, and at the other end (200 ms), the final vowel fully overlapped with the burst. At 0 ms (arbitrarily defined), the burst was in the middle of the two vowels. Two audiovisual continua were also created by pairing the auditory continuum with the visual articulatory gestures for /ip_i/ (consonant offset) or /i_pi/ (consonant onset).

In addition to the continua, we tested the recorded cluster articulations /ikpi/ and /ipki/, and the articulation of the vowels /i_i/, which were presented to the subjects in the auditory, visual and congruent audiovisual conditions. The articulations /ip_i/ and /i_pi/ were only tested in the visual and the congruent audiovisual conditions. In total, 40 stimuli were tested in the experiment. The audio of the speech material had a sampling rate of 48 kHz, and a resolution of 24 bits. The video had a resolution of 720 x 576 and frame rate of 25 Hz. The total duration of each stimulus was one second. All visual articulations started and ended with a neutral expression of the speaker with the mouth closed, which lasted at least two video frames.

Experiments

The experiments were conducted in a sound-proof booth. The subjects were seated 80 cm in front of a 21.5-inch Dell monitor, from which the videos were reproduced. The sound was played back monaurally at 65 dBA from a loudspeaker positioned above the computer monitor. Prior to the experiment, the subjects received written instructions in French and performed a training session that lasted one trial. The experiment was conducted in two separate blocks of ten trials each. Within a trial, all stimuli were presented in random order. The subjects were asked to report what they heard in each trial, or what they saw in the case of the visual trials. The response options were labelled on a computer keyboard and corresponded to /k/, /p/, /kp/, /pk/, and “no consonant” (n.c.). The subjects took a five-minute break between the experimental blocks. The total duration of the experiment session was 60 minutes.

Data analysis

Data analysis of responses was carried out in three steps: (1) Assessment of the auditory consonant identification by contrasting all percepts containing /k/ with pure /p/ percepts; (2) assessment of the visual influence by contrasting all percepts containing /p/ with “no consonant” or pure /k/ percepts; and (3) assessment of the perceived consonant order by defining the PK-index. The index was estimated as the proportion of /pk/ responses out of the total combination responses, and hence, represents the conditional probability of obtaining a /pk/ response given the occurrence of a combination response. For all analyses, the effects were evaluated using linear-mixed model ANOVAs with subject as a random factor and SOA and stimulus type as fixed factors. Post-hoc multiple comparisons were performed with Tukey’s HSD test. Analyses involved the mean proportions of percepts with /k/ (step

1), the mean proportion of percepts with /p/ (step 2) or the mean PK-index (step 3). A significance level of 0.05 was considered in all analyses.

RESULTS

Assessment of the auditory consonant identification

To study the effect of SOA and articulatory gestures on the perception of the auditory stream, we analyzed the responses containing /k/, which indicate that the burst was indeed perceived as a consonant. Figure 1 shows the response percentages obtained for the auditory and audiovisual continua with the two visual contexts as a function of SOA. A linear-mixed model ANOVA with fixed factors for visual context (three levels: no visual, /ip_i/, and /i_pi/) and SOA, (nine levels) and subject as a random factor was performed on the sum of /k/, /kp/ and /pk/ response percentages. The outcome of the test revealed a significant main effect of SOA [$F(8,338) = 35.62$, $p < 0.0001$] and visual context [$F(2,338) = 58.10$, $p < 0.0001$] on the perception of /k/, whereas their interaction was not significant [$F(16,338) = 1.25$, $p = 0.228$].

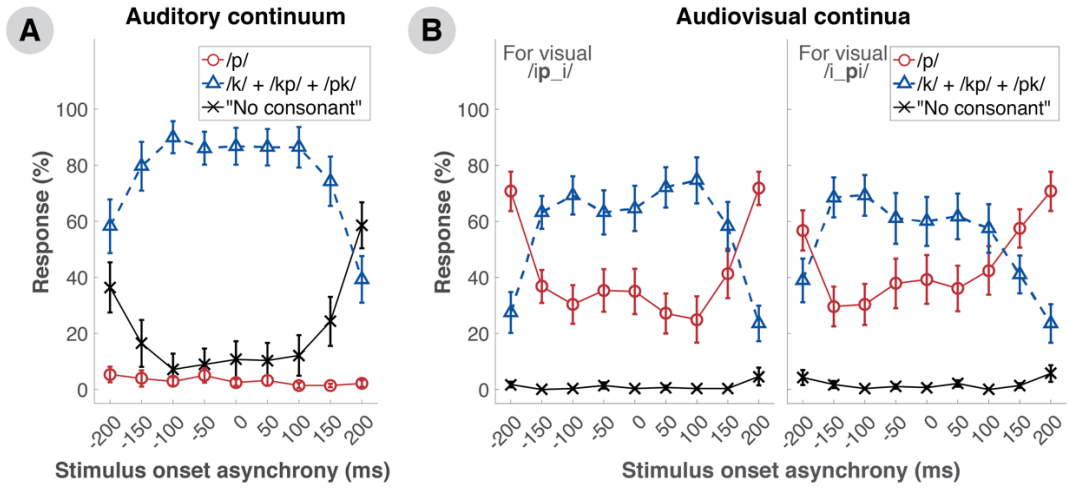


Figure 1: Mean response percentages obtained for the auditory continuum **(A)** and the audiovisual continua **(B)** for the two visual articulatory contexts as a function of SOA. The error bars show the standard error of the mean across subjects.

The main effect of the visual context on the perception of the auditory stream is reflected in the lower percentage of responses containing /k/ in the audiovisual continua than in the auditory continuum. Post-hoc multiple comparisons using Tukey's HSD test confirmed the significant differences in the responses containing /k/ between the auditory continuum and each of the two audiovisual continua ($p < 0.0001$), but not between the two audiovisual continua ($p = 0.211$). This indicates that both audiovisual continua decreased the perception of /k/ across SOA in a similar way, which could be due to a "visual capture" effect reflected in the larger proportion of /p/ responses obtained in the audiovisual continua than in the auditory continuum.

Varying SOA also affected the perception of /k/ responses. The effect was more pronounced at the extreme SOAs for which the percentage of responses with /k/ decreased for all continua. The post-hoc multiple comparisons using Tukey's HSD test showed significant differences between –200 ms and all other SOAs ($p < 0.001$), except for 200 ms ($p = 0.04$), and 200 ms contrasted with all other SOAs ($p < 0.001$). Interestingly, for all comparisons between –150 ms and 100 ms, no significant differences were found ($p = 0.84$), revealing a "plateau" region in which /k/ was similarly perceived across SOAs.

The responses obtained for the auditory continuum indicate that the two auditory streams in the continuum (burst and vowels) were perceived as /iki/ across most SOAs despite the lack of formant transitions. Importantly, /k/ was not clearly perceived for all of the SOAs, since at –200 and 200 ms the "no consonant" responses increased at the expense of the responses containing /k/. This suggests that it is not the burst alone that is perceived as a VCV, which is further supported by the fact that the subjects correctly perceived the auditory stimulus /i_i/ (mean response percentage of 95%).

Assessment of the visual influence

The effect of SOA and visual articulatory gestures on the visual influence was studied by analyzing the sum of /p/, /kp/ and /pk/ responses, in which the visual /p/ influenced speech perception. Figure 2 shows the response percentages obtained for the audiovisual continua in the two visual articulatory contexts as a function of SOA, and for the five unimodal visual articulations. A linear-mixed model ANOVA was fitted to the sum of responses containing /p/. The subject was treated as a random factor, whereas the visual context (two levels: /ip_i/ and /i_pi/) and SOA (nine levels) were treated as fixed factors. The test revealed a significant effect of SOA [$F(8,221) = 8.88$, $p < 0.0001$], while the visual context [$F(1,221) = 2.05$, $p = 0.153$], and the interaction of visual context and SOA [$F(8,221) = 0.73$, $p = 0.66$] were insignificant.

The main effect of SOA on the perception of the visual stream is reflected by the decreased responses containing /p/ for the median SOAs (between –50 and 50 ms). Post-hoc multiple comparisons using Tukey's HSD test confirmed the significant differences between –50 ms contrasted with all other SOAs ($p < 0.01$), except for 0 and 50 ms ($p = 0.78$), and for 50 ms compared to all other SOAs ($p < 0.05$), except for –50 and 50 ms ($p = 0.78$). In contrast, no significant differences were found for all comparisons in the range from –200 to –100 ms ($p = 0.99$), nor in the range from 100 to 200 ms ($p = 0.99$). These results suggest that audiovisual integration occurred more frequently when the burst was closer to the vowels.

Across SOAs, the response percentages containing /p/ were independent of the visual context, as both visual /ip_i/ and /i_pi/ produced similar responses. For these two visual stimuli, the response almost always contained /p/, as reflected in 94% and 97% of responses for /ip_i/ and /i_pi/, respectively. However, cluster percepts were also frequently obtained due to the difficulty in detecting whether the visual articulation contained /k/ in addition to /p/. The two visual cluster articulations presented perceptual confusions and were perceived correctly in 63% and 54% of the trials for

visual /ikpi/ and /ipki/, respectively. Also, the subjects were remarkably successful in recognizing when the visual articulation did not contain a consonant, as indicated by 99% correct identifications of /i_i/.

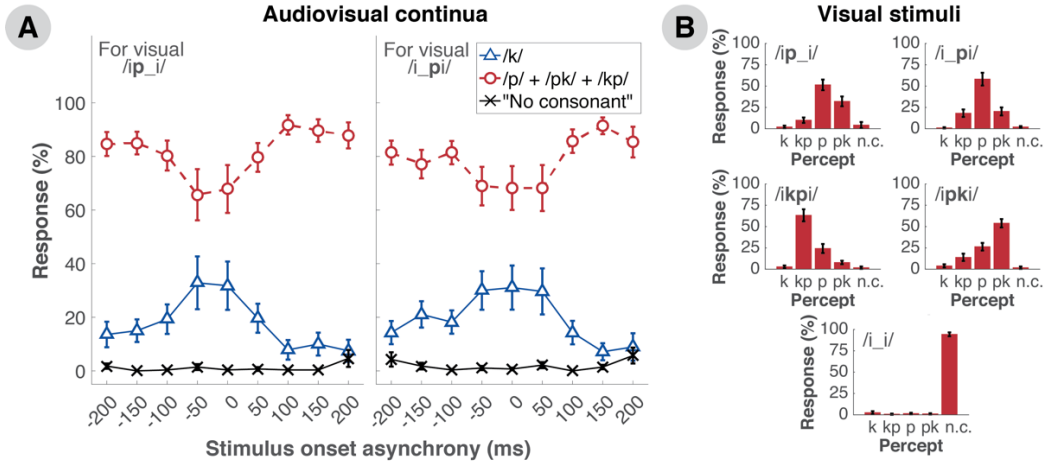


Figure 2: Mean response percentages obtained for the audiovisual continua in the two visual articulatory contexts as a function of SOA **(A)** and for the five visual stimuli tested **(B)**. The error bars show the standard error of the mean across subjects.

Assessment of the perceived consonant order

The PK-index was used to assess the perceived consonant order of the combination responses. Figure 3 shows the PK-index estimated for the two visual contexts as a function of SOA. A linear mixed-effects model was fitted to the mean PK-index. The visual context and SOA were taken as fixed factors and the subjects as a random factor. The outcome of the test revealed a significant main effect of visual context [$F(1,181.05) = 16.05$ $p < 0.0001$] and SOA [$F(8,180.92) = 60.71$, $p < 0.0001$], and an insignificant two-way interaction [$F(8,178.96) = 1.21$, $p = 0.29$].

The effect of SOA on the PK-index for the two audiovisual continua can be seen in the two distinct regions found with different combination responses. One region with a small PK-index, from –200 to –100 ms, for which the asynchrony produced mostly /kp/ responses, and another region with a high PK-index, from 50 to 200 ms, where mostly /pk/ responses were perceived. Tukey's HSD test showed significant differences in the PK-index for all possible pairwise comparisons across regions ($p <$ 0.0001). No significant differences were found in the PK-index for any of the pairwise comparisons within the region with small PK-index ($p > 0.96$), nor within the region with high PK-index ($p > 0.89$). These results suggest that, for the two audiovisual continua the subjects perceived one order of consonants when the burst was closer to the initial vowel, and the reverse consonant order when the burst was closer to the final vowel.

For each audiovisual continuum, the consonant order reversal occurred at different SOAs. For the continuum paired with visual /ip_i/ the reversal occurred (earlier) at

–50 ms, and for the continuum paired with visual /i_pi/ the reverse percept occurred (later) at 0 ms. This is consistent with the fact that, in the case of the articulation of /ip_i/, the lips are closed earlier to produce the bilabial consonant than in the case of the articulation of /i_pi/. Post hoc Tukey's HSD test confirmed the significant differences between the two visual contexts on the PK-index ($p < 0.0001$). This indicates that the perceived consonant order depends on the timing of the acoustic burst and the visual articulatory gestures.

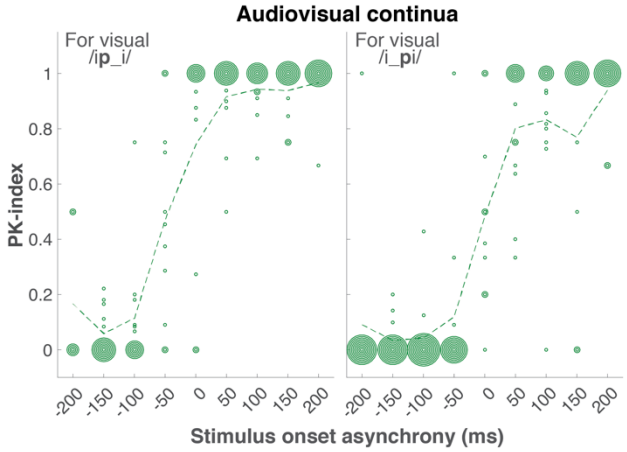


Figure 3: PK-index obtained for the audiovisual continua in the two visual articulatory contexts as a function of SOA. The concentric circles indicate the individual indices and the dotted lines the estimated mean across subjects.

DISCUSSION

The main result of the current study is that the perceived consonant order in McGurk combinations can be reversed consistently by varying the timing of the burst and aspiration of the auditory component. Importantly, we show that the timing at which the reversal in the perceived consonant order occurs seems to depend on the temporal alignment of the burst relative to the articulatory mouth gestures of the speaker. These results support the hypothesis that the burst and aspiration are important cues for audiovisual speech perception, affecting the perceived consonant order for McGurk combinations, and not only the strength of the integration of the cluster percept as has been shown earlier (Green and Norrix, 1997).

Our results are surprising in light of previous findings in which mostly the combination response with the visual labial consonant leading was found (e.g., Massaro and Cohen, 1993; Soto-Faraco and Alsius, 2009). While in previous studies the researchers varied the cross-modal timing of consonant-vowel (CV) stimuli, in the present study only the timing of the burst and aspiration was altered, while the vowels were kept synchronous. Our results are consistent with the prior reports in that we mostly found cluster responses with the labial consonant leading when the burst was closer to the onset of the vowel, as would be the case for CV stimuli. Also, since we found reverse combination responses when the acoustic burst was closer to the offset

of the vowel, one could expect cluster responses with the acoustic non-labial consonant leading for vowel-consonant (VC) stimuli. This is partly what Hampson and colleagues reported when testing VC combinations (Hampson *et al.*, 2003). They found several responses with the reverse order of consonants, although these percepts only exceeded the most common order of consonants (with the labial leading) when the visual component lagged the auditory component. It thus appears that the position of the consonant within the stimulus, either consonant offset or onset, influences the perceived consonant order, and that such effect seems to be driven by the timing of the burst and aspiration relative to the articulatory gestures of the mouth, as seen in our results.

Our findings also suggest that the cluster percept in McGurk combination provides information on how the individual stimulus features are integrated. While the burst and aspiration seem to be sufficient cues for the perception of the consonant /k/ at most SOAs, the perception of the bilabial /p/ seems to depend on the place information in the visual stream. A combination response then arises when both /k/ and /p/ are perceived, whereas the order of the consonants in the cluster percept depends on the temporal organization of these acoustic features (burst and aspiration) and the mouth closing gestures. Finally, the experimental paradigm in this study further revealed the robustness of audiovisual speech perception, as the phonetic features were split into different streams across a range of temporal asynchronies and were yet integrated.

REFERENCES

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (**1974**). "Auditory and visual contributions to the perception of consonants." J. Speech, Lang. Hear. R., **17**, 619-630.

Colin, C., Radeau, M., Deltenre, P., Demolin, D., and Soquet, A. (**2002**). "The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations." Eur. J. Cogn. Psychol., **14**, 475-491.

Green, K. P., and Norrix, L. W. (**1997**). "Acoustic cues to place of articulation and the McGurk effect: the role of release bursts, aspiration, and formant transitions." J. Speech Lang. Hear. R., **40**, 646-665.

Hampson, M., Guenther, F. H., Cohen, M. A., and Nieto-Castanon, A. (**2003**). "Changes in the McGurk Effect across phonetic contexts." Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems.

Massaro, D. W., and Cohen, M. M. (**1993**). "Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables." Speech Commun., **13**, 127-134.

McGurk, H., & MacDonald, J. (**1976**). "Hearing lips and seeing voices". Nature, **264**, 746-748.

Soto-Faraco, S., and Alsius, A. (**2009**). "Deconstructing the McGurk–MacDonald illusion." J. Exp. Psychol. Hum. Percept. Perform., **35(2)**, 580.

Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise." J. Acoust. Soc. Am., **26**, 212-215.

Auditory adaptation in real and virtual rooms

FLORIAN KLEIN[1,*], STEPHAN WERNER[1], GEORG GÖTZ[1,2] AND KARLHEINZ BRANDENBURG[1]

[1] *Electronic Media Technology Group, Technische Universität Ilmenau, D-98693 Ilmenau, Germany*

[2] *Department of Signal Processing and Acoustics, Aalto University, FI-00076 Aalto, Finland*

Walking from room to room in real listening conditions is a natural process in our everyday life and there is no obvious challenge for our auditory system to cope with. However, in experiments with virtual acoustic environments switching the virtual room or switching from real to virtual rooms can result in auditory confusions which can lead to in-head localization. This effect is known as the room divergence effect. A series of listening tests were conducted to verify this effect under different conditions as well as experiments which studied the effect of prior sound exposure and the time variant behaviour of it. In this paper two of these experiments are described and discussed. The first experiment shows that the extent of the room divergence effect depends on the room acoustics we have just learned. That indicates, that the room divergence effect is diminished during ongoing exposure to a specific room acoustic condition. The second listening test shows further evidence of this time-variant effect and we show that it can be suppressed by interrupting with the adaptation process. These tests raise the question why switching virtual rooms leads to temporary confusions but doing so with real rooms is unproblematic. Different theories are discussed in this publication.

INTRODUCTION AND MOTIVATION

Auditory adaptation effects are well known in a broad range of research areas like neurosciences as outlined in King (2008) and hearing rehabilitation as described by Moore *et al.* (2009). Listening training is important for hearing impaired people to familiarize with their newly fitted hearing aid or cochlear-implants. However, such adaptation effects are rarely taken into account during the evaluation of binaural synthesis systems or other spatial audio reproduction techniques. Previous research in this field has shown, that auditory adaptation to altered localization cues as shown by Mendonça (2014) can improve localisation. Also, adaptation to changing room acoustic situations were observed by Keen and Freyman (2009) as well as Seeber *et al.* (2016).

The ability for spatial hearing is not only based on signal driven processing but also

*Corresponding author: florian.klein@tu-ilmenau.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

on listener experience and expectations. Research in the domain of quality perception suggests that expectations about sound serve as an internal reference for the listener in order to rate the perceived quality (see Raake and Blauert (2013)). These expectations can change depending on prior sound exposure. Based on this concept, it might be possible that listeners learn how to interpret spatial cues and room reflections for localization tasks.
Virtual acoustic environments aim to place the listeners in different acoustical environments and therefore forces them to adapt to these situations. This publication presents research which shows such adaptive processes in different listening tests and it discusses differences between the perception of real and virtual rooms.

STATE OF THE ART

The auditory precedence effect describes the prioritization of the first sound waves of a sound event arriving at a listener in the perception of the sound event as described by Wallach *et al.* (1949). Sound waves arriving after this are assigned to the first sound until an echo threshold is reached. This effect refers in particular to the localization and directional assignment of auditory events in an environment affected by sound reflections. The temporal range of the precedence or fusion depends on the spectral composition of the sound waves arriving later in relation to direct sound and on adaptation to the spatial temporal patterns of direct sound and sound arriving later.

The temporal order of magnitude of the echo thresholds underlies a build-up process. The build-up of the precedence effect has been intensively investigated in experiments by Clifton and Freyman (1989); Freyman *et al.* (1991) as well as Clifton *et al.* (1994). A repetition of the same patterns of direct sound and reflections led to an increase in the echo threshold. This indicates an adaptation and learning process that is less dependent on the length of time and more on the number of comparable reflection patterns (see Djelani and Blauert (2001)). In conclusion, this means that in a changing acoustic environment there is no sudden collapse (or rebuilding) of the precedent effect. A significant extension of the precedence effect is the one proposed by Clifton (1987) and Litovsky *et al.* (1999) on the spatial variation of the pattern of direct sound and reflection. A change of the pattern leads to a reduction of the echo threshold and thus to a collapse of the precedence effect. After the change a new precedence effect is established. This effect is commonly referred to as the Clifton effect.

The room divergence effect (RDE, see also Werner *et al.* (2016)) describes the influence of the acoustical differences between an virtually auralized room (for example via binaural synthesis) and real room. If such a divergence is present, the perceived externalization of the auditory event is reduced. The reason for this effect lies in a cognitive disproportion between the expected auditory event and the actual perceived auditory event. A basic approach to the explanation can be found in the precedence effect and its extension, the Clifton effect. The patterns stored in the auditory system for recognizing the room and the audio scene do not correspond to those derived from the synthesis. If these deviations are sufficiently large, the brain

is no longer able to produce perceptive fusion between virtual and real room. The assimilation of the currently perceived event to a stored schema/pattern fails. The term externalization describes the perception of the location of an auditory event outside the head. The counterpart to this is the in-head localization (IHL). The perception of auditory events outside the head is regarded as a mandatory quality feature of a binaural headphone system for the generation of a plausible auditory illusion. In studies by Toole (1970) and Plenge (1972) it becomes clear that the effect of the IHL is not necessarily dependent on the use of a headphone system. In his experiments, Toole (1970) was able to show that IHL also occurs when loudspeakers are used in environments with low reverberation. In a further study on the emergence of the IHL through Plenge (1972), the hypothesis is put forward that the IHL arises through a lack of adaptation or an inadequate learning process. The learning process includes the short-term learning of properties of the sound source and the listening room. Further experiments show that a smooth transition between an out-of-head localization and an in-head localization in loudspeaker reproduction in a low-reflection room cannot be clearly established. Even small changes in the test signals lead either to IHL or to the perception of externalization. The perception of externalization can be understood as a dichotomous quality feature based on the results of this experiment. Plenge (1972) states that an in or at the head localization occurs when there is a "missing, inadequate or incorrect sound source and sound field knowledge and/or the signals and thus the stimuli are of such a nature that they cannot be assigned to any stimulus pattern contained in the long-term memory". The results suggest that the quality feature externalization is influenced by the context of the playback and the listening situation. To acquire a deeper understanding how prior-listening experience alters the perception of externalization the following listening tests were conducted. This way it may be possible to find methods to estimate the role of context parameters in quality perception. The results could help to design listening tests which are closer to real-life experiences of auditory augmented or virtual realities.

LISTENING TESTS AND RESULTS

This section outlines two studies which indicate an effect of prior room exposure on the perception of externalization.

Externalization rating

The evaluation of externalization of an auditory event is performed by selecting a inner, middle or outer region on rating sheet similar to Figure 1. The following definitions are used for the individual areas: a) mid-point: "The auditory event is completely in my head and very diffuse."; b) inner circle: "The auditory event is completely in my head and easy to locate."; c) mid circle: "The auditory event is external but very close to my ears or head."; d) outer circle: "The auditory event is external and easy to locate."; e) outer point: "The auditory event is external and very diffuse." According to the definition, externalization occurs when the auditory event is perceived outside the geometric extension of the head. The position of the

perceptive decision point varies from person to person. For example, one person may already have an unambiguous degree of externalization if the auditory event is located very close to the head, while another person may still have an unambiguous degree of in-the-head localization. On the basis of the individual decision, externalization is evaluated on a multi-point scale. In the current study, the externalization index is calculated by dividing the number of external ratings d) and e) by the number of all ratings.

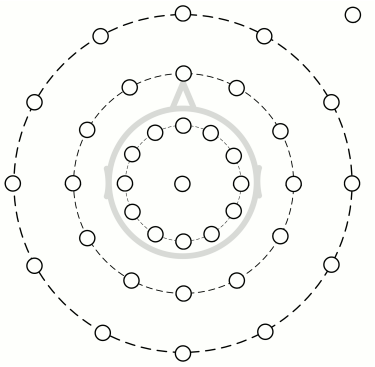


Fig. 1: Graphical user interface (GUI) for externalization and localization rating. Concentric rings are declared as in-head localization (1st ring), near but outside the head localization (2nd ring), and outside the head localization (3rd ring).

Listening test design

In order to provoke the room-divergence effect in the experiments discussed here, two rooms of similar size but strongly differing reverberation time and direct-to-reverberant ratios were chosen. The seminar room (SR) has a reverberation time at 1 kHz of 2 s and the listening lab (LL) has a reverberation time at 1 kHz of 0.339 s. For both listening tests, individual binaural room impulse responses (BRIRs) and headphone compensation filters were recorded in both rooms. The first listening test was conducted with a static binaural synthesis system. For the second test a dynamic binaural synthesis was realized by using the Smyth Realizer (Smyth *et al.*, 2008). For both studies speech and saxophone signals were used.

Excerpt from study no. 1

For the first listening test 31 participants were randomly separated into two groups, each trained to one of the rooms. The “convergent group” was trained to the SR room by listening to real loudspeakers in this room (LS) and a binaural synthesized stimuli of these loudspeakers (Synth SR). The “divergent group” was trained by listening to binaural synthesized stimuli of loudspeakers from the LL room (Synth LL). After the training, both groups were faced with the familiar and unfamiliar room condition to

measure how the training sessions would influence the externalization ratings. The training was designed as a simple localization task accompanied with a judgment on the perceived level of externalization. Next to the rating task, visual feedback was provided at the correct source position by visually highlighting a loudspeaker model. After training, the listeners had to rate externalization of stimuli from the familiar and unfamiliar room.

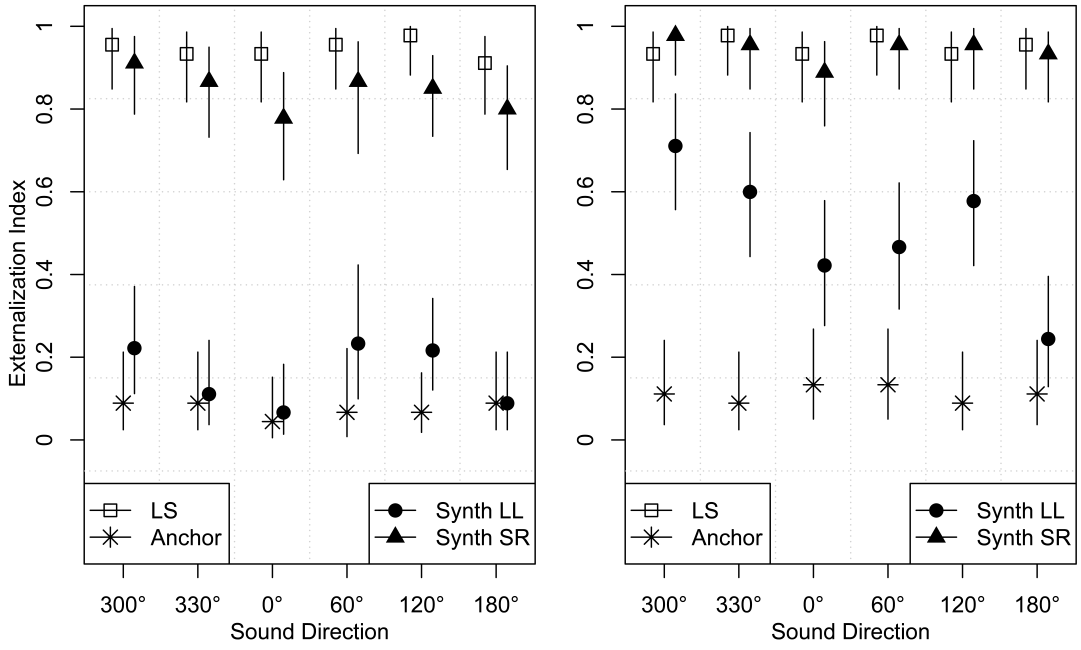


Fig. 2: Results for the externalization ratings of **left**, the convergent group and, **right**, the divergent group. Ratings are separated according to the direction of the presented sound. **Synth SR**, synthesis of the actual listening room (SR), **LS**, real loudspeakers in the listening room (SR), **Synth LL**, synthesis of acoustically dry listening room (LL) (Klein and Werner, 2017).

Figure 2 shows an excerpt of the test results. Ratings are separated according to the direction of the presented sound, because externalization is often direction dependent (see Werner *et al.* (2016)). The listening tests were conducted in the SR room. The low-quality anchor was measured with an omnidirectional microphone in both rooms and is aimed to provoke in-head localization. The results show high ratings for the actual room (LS and Synth SR) regarding to the externalization. The ratings for Synth SR of the convergent group is in tendency a bit lower than for the divergent group. Since this group listened to the Synth SR signals during training they might have discovered flaws of the binaural synthesis system. The rating of room LL is very different between the groups. A difference between Synth LL and Synth SR is clearly visible for both groups and relates to the room-divergence effect. It is particularly

strong when the synthesis of an acoustically dry room is presented in a reverberant room as it was the case here. The ratings of Synth LL are significant higher for the divergent group than for the convergent group. This shows that the perceived externalization can shift according to the previous training session. In other words, the room-divergence effect highly depends on the listeners' acoustic experience.

Excerpt from study no. 2

The main aim of this study was to measure effects of head movements on the externalization (Werner *et al.*, 2017). In the original study several playlists were presented subsequently with and without head tracking enabled. Because of this test design, adaptation effects regarding externalization were expected. To avoid a mixup between the effects of head tracking and adaptation on the externalization rating, the room related adaptation was interrupted on purpose. Overall 36 participants were divided into a room convergent group which rated the synthesis of the actual listening test room (SR room), and a divergent group which rated the LL room while sitting in the SR room.

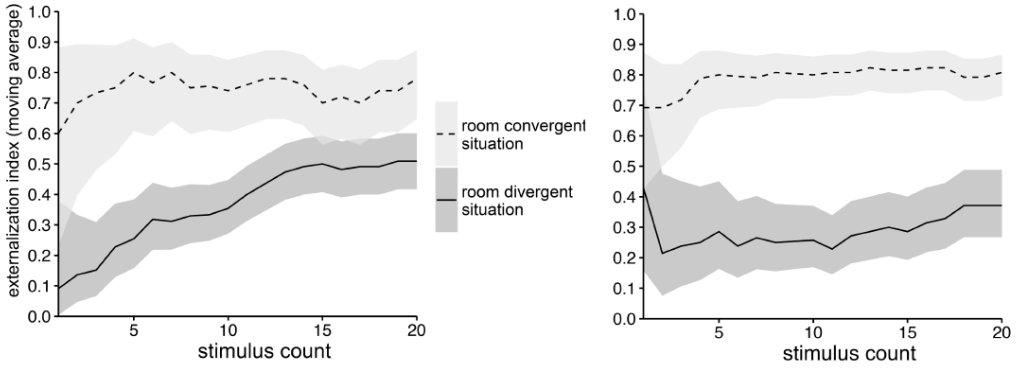


Fig. 3: Moving average for the externalization index over stimulus count and for two room situations. Gray area corresponds to 95% binomial confidence interval. Between first test (left) and second test (right) adaptation was interrupted by the presentation of 36 stimuli from loudspeakers.

In the first test each group had to rate the externalization of 20 test stimuli of Synth LL or Synth SR depending on which group they belong to. The ratings from the first test are shown on the left side of Figure 3. During this test session the ratings of the room divergent situation increases clearly with stimulus count. Before the second test run was conducted, 36 stimuli from the real loudspeakers in the SR room were presented in order to interrupt this adaptation behaviour. The second test run was identical to first one and the results are shown on the right hand side of Figure 3. Until stimulus five there are again some variations, but the ratings of the divergent situation never reach the values of the first test. Also, the increase over stimulus count is much less

then in the first test. The results of this study show that a rapid room related adaptation can occur in such tests (at least regarding externalization) and that it is important to control this adaptation. It may be necessary to suppress this time-variant effect in order to study other quality elements such as the effect of specific system components (for example the benefit of head-tracking) on the perception of externalization.

CONCLUSION AND QUESTIONS

The listening tests show the effect of training to a specific room situation. The results indicate that externalization is influenced by prior knowledge and expectations about room acoustics. These results fit to the early findings of Toole (1970) and Plenge (1972). The experiments have shown that altering prior experience affects the perception of externalization. Listening in virtual acoustic environments and quality ratings thereof are strongly influenced by auditory adaptation effects. At this point the question arises whether there is a difference between the perception of real and virtual rooms. The studies on the precedence and Clifton effect indicate that the human auditory system needs time to interpret direct sound and its reflections in order to perceive a distinct position of a sound source. The relevant time frames in which an adaptation occurs is in order of a few milliseconds to seconds in these studies while in our studies the adaptation happens over the time span of several minutes.
Based on the listening experience in real rooms most people have probably never experienced a familiarization phase accompanied with in-head localization. So why is it the case in virtual acoustics environments? Experiments like those presented, create situations which normally do not exist. For example, rooms are switched with the press of a button while in reality there is always a transition phase when switching rooms. Furthermore, we hypothesize that an adaptation to new acoustic environments happens all the time. In reality a vast amount of acoustic information is available because mostly there are several sound sources at once and the listener also emits or creates sounds by walking in the room for example. In addition, all other senses also provide coherent information: visual and acoustic sound source positions match, the listener movement is translated into a change of the acoustic signal and so on. In laboratory experiments the amount of information which is provided to understand an acoustic scene is mostly limited. These limitations possibly require a longer time for our brain to understand the scene and as long the scene is not understood, perceptual errors like in-head localization are likely.

REFERENCES

Clifton, R. K. (**1987**), “Breakdown of echo suppression in the precedence effect,” J. Acoust. Soc. Am., **82**(5), 1834–1835.

Clifton, R. K. and Freyman, R. L. (**1989**), “Effect of click rate and delay on breakdown of the precedence effect,” Percept. Psychophys., **2**(49), 139–145.

Clifton, R. K., Freyman, R. L., Litovsky, R. Y., and McCall, D. (**1994**), “Listeners’ expectations about echos can raise or lower echo threshold,” J. Acoust. Soc. Am., **95**(3), 1525–1533.

Djelani, T. and Blauert, J. (**2001**), "Investigations into the Build-up and Breakdown of the Precedence Effect," Acta Acust united Ac., **87**(2), 253–261.

Freyman, R. L., Clifton, R. K., and Litovsky, R. Y. (**1991**), "Dynamic Processes in the Precedence Effect," J. Acoust. Soc. Am., **90**(2), 874–884.

Keen, R. and Freyman, R. L. (**2009**), "Release and re-buildup of listeners' models of auditory space," J. Acoust. Soc. Am., **125**(5), 3243 – 3252.

King, A. J. (**2008**), "Visual influences on auditory spatial learning," Philos. Trans. R. Soc. Lond., B, Biol. Sci., **364**(1515), 331–339.

Klein, F. and Werner, S. (**2017**), "Influences of training on externalization in binaural synthesis in situations of room divergence," J. Audio Eng. Soc., **65**(3).

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. (**1999**), "The precedence effect," J. Acoust. Soc. Am., **106**(4), 1633–1654.

Mendonça, C. (**2014**), "A review on auditory space adaptations to altered head-related cues," Front. Neurosci., **8**(219), 1–14.

Moore, D. R., Halliday, L. F., and Amitay, S. (**2009**), "Use of auditory learning to manage listening problems in children," Philos. Trans. R. Soc. Lond., B, Biol. Sci., **364**, 409 – 420.

Plenge, G. (**1972**), "Über das Problem der Im-Kopf-Lokalisation [On the Problem of In Head Localization]," Acustica, **26**(5), 241–252.

Raake, A. and Blauert, J. (**2013**), "Comprehensive modeling of the formation process of sound-quality," Proc. Quality of Multimedia Experience (QoMEX).

Seeber, B. U., Müller, M., and Menzer, F. (**2016**), "Does learning a room's reflections aid spatial hearing?" Proc. ICA.

Smyth, S., Smyth, M., and Cheung, S. (**2008**), "Smyth SVS headphone surround monitoring for studios," 23rd UK Conference of the Audio Engineering Society.

Toole, F. E. (**1970**), "In-Head Localization of Acoustic Images," J. Acoust. Soc. Am., **48**(4B), 943–949.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**), "The Precedence Effect in Sound Localization," Am. J. Psychol., **62**(3), 315–336.

Werner, S., Götz, G., and Klein, F. (**2017**), "Influence of Head Tracking on the Externalization of Auditory Events at Divergence between Synthesized and Listening Room Using a Binaural Headphone System," Audio Engineering Society Convention 142.

Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K. (**2016**), "A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events," Proc. Quality of Multimedia Experience (QoMEX).

Audio-visual sound localization in virtual reality

THIRSA HUISMAN[1,*], TOBIAS PIECHOWIAK[2], TORSTEN DAU[1], AND EWEN MACDONALD[1]

[1] *Centre for Applied Hearing Research, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *GN Hearing, GN ReSound, Region Hovedsteden, Denmark*

Virtual reality (VR) can be a strong research tool in audio-visual (AV) experiments. It allows us to investigate AV integration in complex and realistic settings. Here, using a VR setup-up in combination with a loudspeaker array, 16 normal-hearing participants were tested on their sound localization abilities. The virtual environment consisted of a 1:1 model of the experimental environment except with the loudspeaker array replaced by a ring. This ring indicated the height, but not the position of the loudspeakers. The visual component of the stimuli consisted of a ball falling and then bouncing once on the ring after which it disappeared. As the ball collided with the ring, an impact sound was played from a loudspeaker. Participants were asked to indicate the apparent sound origin, for both congruent and incongruent visual and audio spatial positions ranging from -30 to 30 degrees. The VR visual stimuli in combination with real auditory stimuli were capable of inducing AV integration. The range of this integration extended, for several participants, over large ranges of AV disparity compared to some earlier studies.

INTRODUCTION

The integration of information from senses is a vital and well-studied topic, with often-cited studies dating back to 1950's. While in earlier studies, visual cues were thought to 'capture' the auditory cues, nowadays audio and visual cues are assumed to be integrated following bayesian causal inference (BCI). In BCI, the probability of a common cause is assessed. If there is no common cause, auditory and visual cues are processed independently. On the other hand, if there is a common cause, the cues are integrated optimally, such that they are weighted relative to their reliability. The actual perceived location of the integrated stimulus can then still vary depending on the decision-making strategies (Wozny *et al.*, 2010).

The assessment of a common cause is an important step in the model as audio-visual (AV) integration has been shown to be influenced by the timing and distance between the audio and visual stimuli, with increased disparity in either time or space reducing the probability of visual capture. Additionally, realism or 'compellingness' is often hypothesized to facilitate the ventriloquist effect. Despite this common assumption that more ecologically-valid stimuli might be integrated more strongly, experiments often use simpler stimuli, such as a light flash and a noise burst. The preference for these stimuli is most likely due to their ease of use. This is where virtual reality (VR) can be a valuable tool. VR allows us to simulate complex, ecologically-valid, and yet

*Corresponding author: thuis@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

controlled scenarios. The current study was intended to function as a normal hearing baseline for a later comparison with hearing-impaired listeners, using VR as a tool to produce ecologically valid stimuli. We hypothesize that our VR baseline will match results of earlier studies on AV integration.

METHODS

Participants

Seven females and 9 males (average age 29.5 ± 13 years) were recruited from the DTU community. All had normal hearing thresholds and normal or corrected-to-normal vision. The procedure was approved by De Videnskabsetiske Komitéer for Region Hovedstaden (H-16036391) and all participants gave informed consent. The participants were compensated with an hourly rate of 122 DKK.

Apparatus

The experiment took place in the Audio-Visual-Immersion-Lab (AVIL) of the Technical University of Denmark. 5 loudspeakers were used to present the auditory cues. These loudspeakers were 2.4 m from the participant in an arc ranging from -30° to 30° azimuth with 15° separation between loudspeakers. Participants were seated in a height adjustable chair at the center. This chair was raised such that the participants' ears were at the height of the loudspeakers.

The visual cues were presented using an HTC VIVE VR headset. The virtual environment was a 1:1 model of the experimental room. However, the loudspeaker array was replaced by a gray ring. This ring, which was 5 cm in height, indicated the elevation and distance, but not the exact azimuth, of the loudspeakers. Only in the final condition of the experiment was the loudspeaker array shown. Just below the loudspeaker ring, at 0° azimuth, was a small white screen, with a visual angle (VA) of 10°. This was the focus point before and during the trials.

Participants could proceed through the experiment and record their judgements using a handheld HTC VIVE controller. In VR, a thin red rod was attached to the end of the controller so that it appeared to have a laser pointer. Participants pointed this "laser" at the location where they perceived the auditory stimuli and pressed a button to record their judgement.

Stimuli

The auditory stimulus was a 20 ms recording of the impact of a handball landing on a carpeted floor, presented at 65 dB peSPL. The visual stimulus consisted of an 8° VA ball. At the start of a trial, the ball appeared at a location above the ring, fell for half a second, bounced once on the ring and then, 20 ms after bouncing, disappeared. The audible impact of the ball on the loudspeaker ring was, on average, delayed by 105±15 ms relative to the visual impact of the ball.

Procedure

The experiment consisted of 4 blocks: unimodal audio, bimodal, unimodal visual and a pointing task. The blocks were presented in this fixed order. In the unimodal conditions and in the pointing task, 2 additional loudspeakers, at ± 45° azimuth, were included. These were not included in the bimodal conditions as these positions were near the limits of the field of view of the VR headset.

In the unimodal conditions, a stimulus was presented, randomly, at one of the 7 loudspeakers. Per position, the measurement was repeated 5 times for the auditory condition and 3 times for the visual condition. The AV block consisted of 322 trials. For each of the 5 loudspeakers used to present the auditory stimuli, visual stimuli were presented at the 7 loudspeaker positions and in a range of 30° around the loudspeaker position, using a 3° step size. This 30° range around the loudspeaker was limited to ± 45° for the speakers positioned at ± 30°, due, again, to limitations of the field of view of the VR headset. All combinations were repeated 3 times. In terms of AV separation, a maximum separation of ± 75° was tested, with the densest sampling occurring in the range of ± 15° disparity.

Trial

Participants were instructed to look forward at the focus point while the stimuli were presented. Once it was verified that their head was oriented towards the focus point, participants could press a button on the controller to start a trial. After the stimuli were presented, participants were asked to indicate where they heard the sound came from. In trials where there were no auditory stimuli, participants were asked instead to indicate where they saw the stimulus came from. Participants were allowed to move and look freely when pointing. Responses were restricted to the ring, such that the elevation and distance was fixed. Participants were, however, allowed to use the entire ring to answer, allowing for front-back confusions.

As eye movements have been shown to influence AV integration, an additional task was used to ensure that, at the moment of collision, participants were looking at the focus point straight ahead (rather than at the ball). As the ball collided with the ring, a letter appeared for 200 ms at the focus point. This letter was recognizable only when looking at the focus point. After performing the spatial localization task, a matrix of 16 different letters was presented and participants were asked to select the letter that had appeared during the trial. If an incorrect letter was selected, the trial was considered invalid and repeated again at a later random position.

The pointing condition was included to obtain the motor error in the pointing. Thus, it did not follow the above described structure. Instead, at the start of this condition, the ring was replaced by the model of the loudspeaker array. On each trial, participants were shown a number and were then instructed to point at the center of the loudspeaker labelled with that number. The loudspeakers were continuously visible and no auditory stimuli were presented in this condition. Hence, this condition estimated how well participants could point at a specific target.

RESULTS

All invalid trials were disregarded in the analysis. Due to a logging error in the pointing condition (where very fast responses could be logged as responses to the previous rather than current target), results with an error over 15 degrees azimuth (in this pointing condition only) were considered invalid.

First, the unimodal conditions were analyzed to predict the visual bias in the bimodal condition and to correct for localization biases found in earlier studies (e.g., Odegaard *et al.*, 2015; Ahrens *et al.*, 2019).

Figure 1 shows the average localization error of the unimodal conditions. As expected, the auditory localization error (max. 13.0° ± 17.6) and especially the variance was greatly increased, compared to the visual localization (max. 4.5° ± 2.9). The effects of the pointing method itself were very small, with the largest pointing error being about a single degree (max. 0.97° ± 0.4).

As in earlier studies (e.g., Odegaard *et al.*, 2015; Freeman *et al.*, 2018; Ahrens *et al.*, 2019), a bias in the unimodal conditions was found. For visual localization, a centralized bias, where visual stimuli are perceived more towards the center, was found (*t*-test, $p < 0.01$ for all non-zero locations). For auditory localization, an externalized bias can be seen for most, but not all locations (*t*-test, $p < 0.01$, for all but 15° and 30° azimuth). Surprisingly, a small externalizing bias, not centralized as might be expected in this visual task, was found ($p < 0.05$ for all paired Welch tests between adjacent angles).

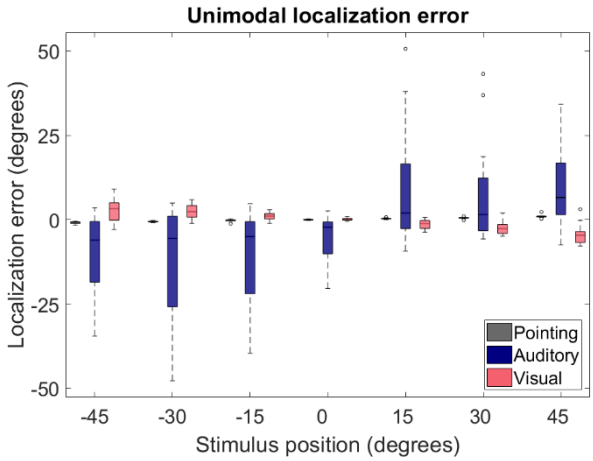


Figure 1: The localization error per modality for 7 stimulus positions. Stimuli presented at negative azimuths occurred in the left hemisphere and stimuli presented at positive azimuth occurred in the right hemisphere. Similarly, a negative localization error indicates that the stimulus was perceived leftwards of the stimulus positions, whereas a positive localization error indicates that the stimulus was perceived to the right of the stimulus position.

The bimodal data were clustered per participant using a Gaussian mixture models clustering (MATLAB, 2017b). The clustering was run with a maximum of 3 clusters to allow for audio, AV and visual clusters. The optimal number of clusters was then chosen using the BIC criteria (Schwarz, 1978). Figure 2 shows the cluster results for 3 participants. The clustering was run 20 times, after which the most prevalent clustering was used in the analysis.

As the task was to localize where the sound came from, we would expect that the localization error in Figure 2 would be around 0 (with some deviation due to the localization bias). Thus, we would expect clustering around the horizontal line in Figure 2. Indeed, at least one cluster with this property appeared for most participants. However, most participants also showed additional clusters, which were consistent with judgments being influenced by the position of the visual stimuli.

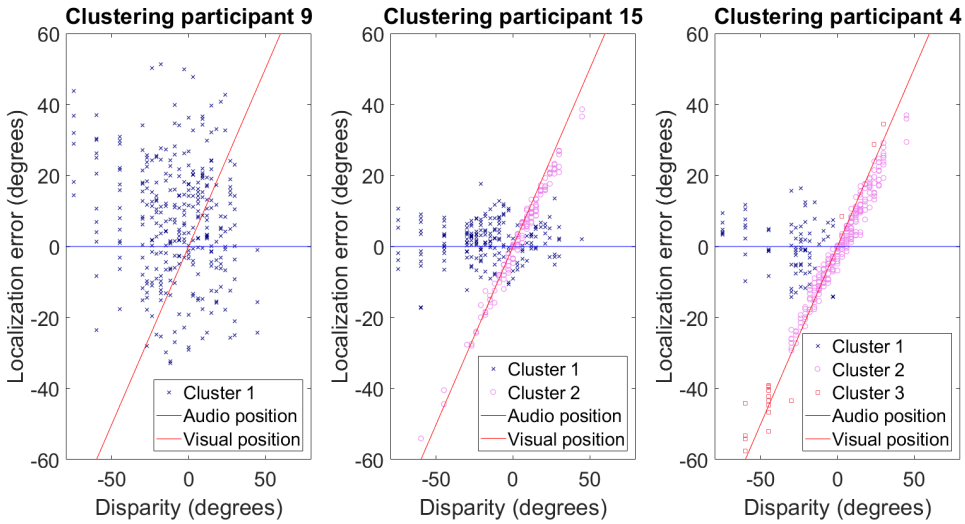


Figure 2: The panels show the cluster results for 3 participants, where, respectively, 1, 2 and 3 clusters were found. Each point is a single measurement, where the localization error is shown as a function of the disparity between the position of the auditory and visual stimuli. The results for the left hemisphere are mirrored such that a negative disparity indicates that the visual stimulus occurred closer to the midline. Additionally, the relative location of the auditory and visual stimulus is shown. The clusters are indicated with different symbols.

Through each cluster, a linear regression was fitted. The slope of each cluster was compared with 0, 1 and the predicted bias to see if it could be explained by either auditory localization, visual localization or AV localization. The predicted bias was based on the variance in the unimodal data. It weights the visual and auditory cues relative to the inverse of the localization variance for each modality, thereby predicting "optimal bimodal integration" (Alais and Burr, 2004). To ensure that the localization bias did not affect the categorization, the comparisons were also run with a correction for the auditory and visual bias. The results are shown in Table 1.

	Sub-categories	Number of clusters	Predicted visual bias	Visual bias	Range	
					Min	*Max*
Audio	*Average*	15	0.94	-0.02	-74.0	36.6
Visual*	*Average*	4	0.98	0.96	-67.5	41.3
Audio-visual	*Average*	16	0.92	0.57	-48.2	39.4
	Larger than 1	1/16	0.99	1.37	-27.0	30.0
	As predicted	2/16	0.87	0.86	-42.0	36.0
	Smaller than predicted	11/16	0.92	0.57	-43.9	39.3
	Smaller than 0	2/16	0.98	-0.11	-75.0	45.0

Table 1: 35 clusters were categorized as audio, visual* or AV based on the slope of the fitted linear regression curve. Audio had a slope that was consistent with auditory localization. Visual* was consistent both with visual and AV localization, but is assumed to be the result of visual only localization. All other clusters, where visual cues influenced, but did not dominate, auditory localization, were considered AV. The columns show, respectively, the number of clusters per category, the average predicted visual bias for these clusters, the average measured visual bias and the average range in degrees (min, max) over which these clusters occurred.

The distribution of the data points in each of the main categories of clusters listed in Table 1 is shown in Figure 3. As expected, the probability of AV responses is largest when the disparity is small. An asymmetry was found, in that the 'optimal' point for AV integration was shifted, such that the probability of integration is largest when the visual stimulus occurs more outwards compared to the auditory stimulus. Overall disparities with visual stimuli being presented from an eccentric position, relative to the position of the auditory stimulus (i.e. positive disparities), were more likely to result in AV integration.

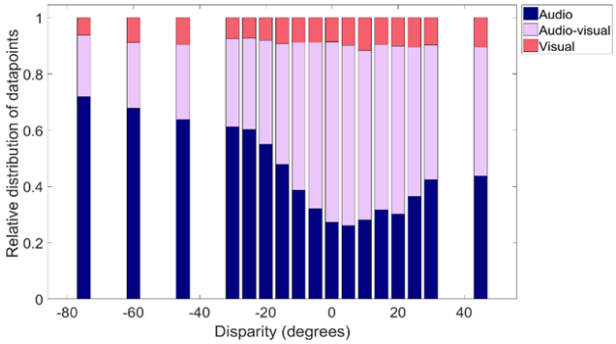


Figure 3: The relative distribution of data points in each cluster across the testing range, averaged over all participants.

DISCUSSION

While we found integration behavior similar to experiments using real world stimuli, there were some differences in our results. First, likely due to the delay between the audio and visual stimuli, we found many audio-only responses. For those, who did show integration behavior we found that the predicted bias generally overestimated the visual bias. This prediction relied on the unimodal results, where the audio-only condition showed a quite high variance on average, which would result in a very high predicted visual bias. Indeed, the predicted visual biases in this study were generally larger than the visual biases found by Battaglia *et al.* (2003). However, the average slope of AV clusters with an overestimated visual bias, is much lower than would be predicted by the visual biases found by Battaglia *et al.*

An alternative explanation might be in a combination of decision-making strategies. As mentioned earlier, the assumed decision-making strategy in most experiments with BCI is model selection. While the distribution of results in the present study could be explained with model selection, the stochastic manner in which the decision-making strategy switches within participants, sometimes from trial to trial, is more consistent with probability matching, which has been found to be the dominant strategy of most participants in a previous study (Wozny *et al.*, 2010). This probability matching by itself cannot explain a decrease in the visual bias. However, earlier studies (Battaglia *et al.*, 2003; Meijer *et al.*, 2019) also found a deviation from optimal integration. Their results showed an increased influence of visual position compared to what optimal integration would predict. They explained this by a model averaging strategy, where the AV and unimodal responses are weighted based on the probability of the underlying causal structure. As this experiment used mostly incongruent stimuli and also covered a wide range of incongruencies, the probabilities of the causal structures were strongly biased towards separate causes. Thus, model averaging could also explain the reduction in the visual influence observed here. As neither the probability matching nor model averaging strategy can account for the results by themselves, it appears that some of our participants combined different decision-making strategies.

Additionally, compared to recent studies (e.g., Bosen *et al.*, 2016), the AV results were found to extend over a surprisingly large range, with stimuli 75° (5 loudspeakers) apart being biased towards the visual stimulus. However, similar results have been found in much older studies (e.g., Jackson, 1953). Potentially, the use of more ecologically valid stimuli (which were also used in older studies) and/or the immersion of VR extends the range of spatial separation over which integration occurs.

These deviations from previous experiments, in particular the integration range, mean that the setup of the current experiment is of limited value for comparing the behaviour of normal-hearing and hearing-impaired listeners as the expected effects in the slope would be too small and the results of normal-hearing listeners already exhibit some visual influence over the full field of view available in VR. Potentially, when VR offers a larger field of view or when an alternative reproduction method is used for the visual stimuli, this method would be more applicable.

CONCLUSION

The VR ecologically valid stimuli used in the present study were capable of inducing integration showing that VR can be used for AV integration experiments. We found a deviation from the expected decision-making strategies. Some of the results here can only be explained by a combination of decision-making strategies (i.e., model averaging and probability matching). As integration occurred over a surprisingly large range, the current paradigm is not applicable for a comparison study between normal-hearing and hearing-impaired listeners with this setup.

ACKNOWLEDGEMENTS

This research was supported by the Centre for Applied Hearing research (CAHR) through a research consortium agreement with GN Resound, Oticon, and Widex.

REFERENCES

Ahrens, A., Lund, K. D., Marschall, M., and Dau, T. (**2019**). "Sound source localization with varying amount of visual information in virtual reality," PLoS ONE, **14**(3), 1–19. doi: 10.1371/journal.pone.0214603

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (**2003**). "Bayesian integration of visual and auditory signals for spatial localization," J. Opt. Soc. Am. A., **20**(7), 1391. doi: 10.1364/JOSAA.20.001391

Bosen, A. K., Fleming, J. T., Brown, S. E., Allen, P. D., O'Neill, W. E., and Paige, G. D. (**2016**). "Comparison of congruence judgment and auditory localization tasks for assessing the spatial limits of visual capture," Biol. Cybern., **110**(6), 455–471. doi: 10.1007/s00422-016-0706-6

Freeman, L. C. A., Wood, K. C., and Bizley, J. K. (**2018**). "Multisensory stimuli improve relative localisation judgments compared to unisensory auditory or visual stimuli," J. Acoust. Soc. Am., **143**(6). doi: 10.1121/1.5042759

Jackson, C. V. (**1953**). "Visual Factors in Auditory Localization," Q. J. Exp. Psychol, **5**(2), 52–65. doi: 10.1080/17470215308416626

Meijer, D., Veselič, S., Calafiore, C., and Noppeney, U. (**2019**). "Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation," Cortex, **119**, 74–88. doi: 10.1016/j.cortex.2019.03.026

Odegaard, B., Wozny, D. R., and Shams, L. (**2015**). "Biases in visual, auditory, and audiovisual perception of space," PLoS Comput. Biol., **11**(12), 1–23. doi: 10.1371/journal.pcbi.1004649

Schwarz, G. (**1978**). "Estimating the dimension of a model," Ann. Stat., **6**(2), 461–464.

Wozny, D. R., Beierholm, U. R., and Shams, L. (**2010**). "Probability matching as a computational strategy used in perception," PLoS Comput. Biol., **6**(8). doi: 10.1371/journal.pcbi.1000871

A method for evaluating audio-visual scene analysis in multi-talker environments

KASPER D. LUND[1]*, AXEL AHRENS[1] AND TORSTEN DAU[1]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark*

In cocktail-party environments, listeners are able to comprehend and localize multiple simultaneous talkers. With current virtual reality (VR) technology and virtual acoustics it has become possible to present an audio-visual cocktail-party in a controlled laboratory environment. A new continuous speech corpus with ten monologues from five female and five male talkers was designed and recorded. Each monologue contained a substantially different topic. Using an egocentric interaction method in VR, subjects were asked to label perceived talkers according to source position and content of speech, while varying the number of simultaneously presented talkers. With an increasing number of talkers, the subjects' accuracy in performing this task was found to decrease. When more than six talkers were in a scene, the number of talkers was underestimated and the azimuth localization error increased. With this method, a new approach is presented to gauge listeners' ability to analyze complex audio-visual scenes.

INTRODUCTION

Normal-hearing listeners are able to localize and understand multiple talkers in complex listening environments, also referred to as 'cocktail-party' scenarios (Bronkhorst, 2000). The ability of the auditory system to analyze such complex scenes is often referred to as "auditory scene analysis". Previous studies have employed signal patterns with varying degrees of spectral or temporal differences to investigate how the auditory system analyses scenes (Bregman, 1994). Other studies have used more speech-like stimuli to increase the ecological validity. However, these test paradigms might not reflect perception in more realistic complex acoustic scenes.

Kopčo *et al.* (2019) asked subjects to identify the location of a female talker in a mixture of male talkers and showed a reduction in localization accuracy relative to a condition without interferers. Weller *et al.* (2016) simulated a more realistic auditory scene with up to six simultaneous continuous speech sources in a reverberant room and asked subjects to identify the location and the gender of the talkers. The subjects were provided a top-down view of the room on a touchscreen. Weller *et al.* (2016) found that normal-hearing subjects were able to accurately analyze scenes with up to four talkers.

*Corresponding author: kdue@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Even though the realism of the paradigms investigating multi-talker scene analysis has increased, some factors have not been considered. For example, most studies focused on audio-only settings, or used allocentric interfaces, where subjects do not have a first-person view of the scene. Thus, potential influences of visual cues and egocentric perception have not been considered.

With recent advances in virtual reality (VR) technology, it is possible to present visual content via a head-mounted display (HMD) in a controlled environment. In the present study, we propose a novel method for investigating complex scene analysis in a realistic audio-visual setup, by combining VR technology, virtual acoustics, and by utilizing an egocentric interface.

METHODS

Speech stimuli

To create the auditory scenes a speech material corpus was designed. The corpus was established by recording continuous Danish speech material, consisting of monologues on substantially different topics. Ten monologues with easy readability and unique terms and words to maximize their distinguishability were composed. Ten non-professional native Danish speakers (five females, five males) were individually recorded while reading each of the ten stories. The fundamental frequencies of the talkers ranged from 116 to 204 Hz. The talkers were recorded with a Neumann TLM 102 large diaphragm condenser microhpone (Neumann GmbH, Berlin, Germany) in a sound-proof listening booth. The text was presented on a virtual teleprompter on an HTC Vive Pro VR system (HTC Corporation, New Taipei City, Taiwan) to avoid the noise from paper or acoustic reflections from a computer screen. Using the VR controller, talkers could scroll through the text in their own pace. For optimal readability the virtual teleprompter was adjustable in distance (size) and height. Each monologue recording was equalized to the same root-mean-square level.

Acoustic setup

The experiment was conducted in an anechoic room containing a 64-channel loudspeaker array (see Ahrens *et al.*, 2019a, for details). The dry speech recordings were spatialized in a simulated reverberant room created with Odeon (Odeon A/S, Kgs. Lyngby, Denmark). The loudspeaker signals were generated employing a nearest loudspeaker mapping method using the loudspeaker auralization toolbox (LoRA Favrot and Buchholz, 2010). The room had an average reverberation time of $\sim$0.4 s. Fig. 1 (left panel) shows an overview of the room and the 21 possible talker positions. The positions were all at ear level and in the frontal hemisphere from -90° (left) to 90° (right), in 30° steps. Three distances were considered for all azimuth directions, 1.4m, 2.4m and 3.4m, where the 2.4m distance coincided with the radius of the loudspeaker array.

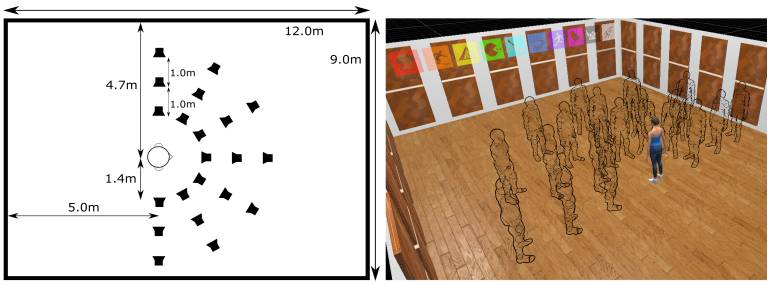


Fig. 1: (Left) Overview of the acoustic scene setup. The room was 2.8m high. The 21 simulated talker positions are indicated by loudspeaker symbols. (Right) Overview of the visual scene setup visually. Semi-transparent humanoid bodies were positioned at locations corresponding to the acoustic source positions.

Visual setup

The visual scene was presented on a HMD (HTC Vive system, HTC Corporation, New Taipei City, Taiwan). Fig. 1 (right panel) shows the virtual visual scene. It contained a room that visually matched the simulated acoustic room in terms of size and surface materials as well as 21 semi-transparent unisex humanoid bodies that were displayed at positions corresponding to the acoustic source positions. At the back wall of the the virtual room, a list of coloured icons was shown, representing the topics of all stories.

The visual scene was rendered using Unity software (Unity Technologies, San Francisco, California, USA) with the SteamVR plugin (Valve Corporation, Bellevue, Washington, USA). To ensure the spatial alignment between the acoustic and the visual scenes, a calibration method using VR trackers was employed (Ahrens *et al.*, 2019b).

Subjects and procedure

Six young (24,3 years old on average), self-reported normal-hearing, native Danish speaking subjects participated in the experiment. Prior to their participation all subjects gave their written consent to the ethics agreement approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). Three repetitions of scenes for each number of simultaneous talkers were run - 27 trials in total. Each subject completed the experiment within two hours and where allowed breaks after each trial.

On each trial, between two and ten talkers and stories were randomly chosen and simultaneously presented from random locations. Duplicates for talkers, stories and positions were not allowed. On each trial, the subjects were asked to identify the stories present in the scene and to change the color of the virtual laser pointer (as seen in Fig. 2) with a button press on a handheld VR controller to match the color

of an icon representing the perceived story. Another button was assigned to change the distance of the laser-pointer to mark sources at different distances. After choosing the color and the distance of the laser-pointer, the subjects could label the perceived talker location by choosing an avatar. After the selection, the avatar changed the color according to the color of the icon/laser-pointer. The audio was presented for 120 s. The time for the subjects' responses was not restricted, but finalized with a button press on the controller. Each acoustic talker was presented at a sound pressure level (SPL) of 55 dB and no feedback was provided to the subjects.

Before the test session, each subject participated in a familiarization session. Each of the ten stories were separately presented once to the subject. A random talker and location was assigned in each trial and the subject was asked to do the task as described above. The audio signals were presented for up to 60 s. After the response, feedback was provided to the subjects by indicating the correct story and position of the talker.

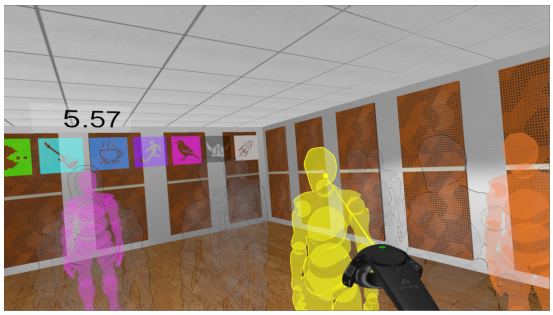


Fig. 2: The visual scene as seen from the point-of-view of the listener. Using the virtual laser pointer subjects' task was to analyze the acoustic scenes and label the perceived positions of a talker according with the perceived story.

RESULTS AND DISCUSSION

Fig. 3 shows how often each talker (top panel) and each story (bottom panel) was correctly identified. This measure evaluates the overall identification difficulty of talkers and stories in the collected response data. The top panel is split into female ('f', left) and male ('m', right) talkers. The analysis of a linear model showed no significant difference in talker identification difficulty ($F(8,50) = 1.37, p = 0.23$). However, an average difference between male and female talker identification accuracy of 9.3%-points was found ($F(1,50) = 10.01, p = 0.0026$), indicating that it was more difficult to identify the male compared to the female talkers. Previous studies showed similar trends (Bradlow *et al.*, 1996). The bottom panel of Fig. 3 shows the story identification difficulty. The analysis of the linear mixed model showed no significant difference across the stories ($F(9,45) = 1.25, p = 0.29$).

Fig. 4 shows the number of perceived talkers as a function of the number of presented talkers in the scene. The black squares represent the mean across subjects and the grey lines show the individual subject data.

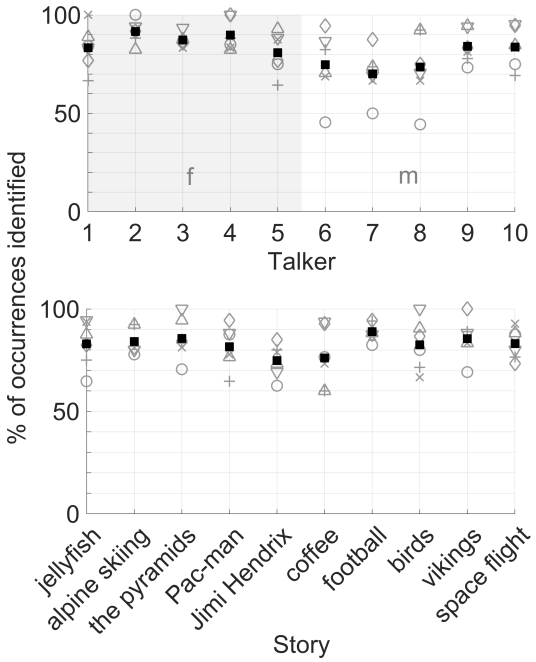


Fig. 3: Overall identification difficulty of talkers (top) and stories (bottom), as % of the occurrences identified. Open grey symbols represent individual subjects, averaged over three repetitions, and black squares represent the mean over subjects. Genders are indicated for talkers ('m' or 'f').

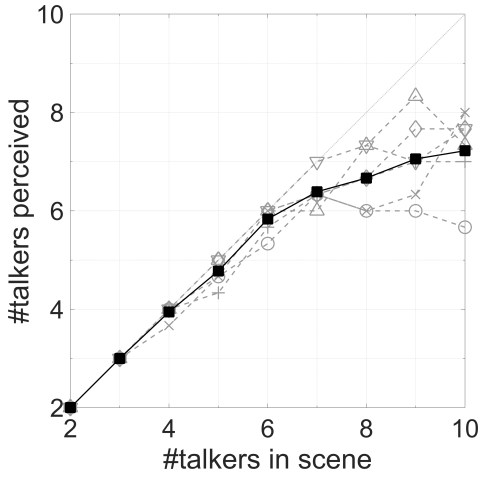


Fig. 4: Number of perceived talkers over number of presented talkers in the scene. Open grey symbols represent individual subjects, averaged over three repetitions, and black squares represent the mean over subjects.

For scenarios with up to six simultaneous talkers, the majority of the subjects were able to correctly identify the number of talkers. In scenarios with more than six

simultaneous talkers, this ability decreased gradually and, on average, the subjects underestimated the number of talkers in the scene.

Compared to the results from Weller *et al.* (2016), the number of correctly identified simultaneous talkers was higher in the present study. While the task in Weller *et al.* (2016) was similar, they only presented auditory stimuli (i.e. no visual information), used an allocentric interface for the response collection, and subjects were only given 45 s response time. Whether the observed larger number of identified talkers resulted from the visual gain, the egocentric interface or the increased time limit, still needs to be clarified. In Weller *et al.* (2016) the subjects needed to judge the gender of the talkers, while in the current study the content of the story needed to be identified. The identification of the content is likely to be more difficult and is expected to reduce the number of correctly identified talkers, which has not been observed in the current study.

Fig. 5 shows the localization accuracy of the sources as a function of the number of simultaneous talkers. The left panel shows the root mean squared (RMS) error in azimuth and the right panel shows the RMS error in distance. Individual subject responses are indicated by the grey symbols whereas the average results across subjects are shown as black squares. For up to five simultaneous talkers, all talker azimuth positions were correctly identified. For up to seven simultaneous talkers, the error did not significantly increase ($p < 0.0001$) as indicated by the analysis of a linear mixed model. For eight and more talkers, the error increased gradually. The distance error (right panel) was found to be independent of the number of the talkers ($F(8,148) = 0.79, p = 0.62$). The average RMS distance error was about 0.57m and thus below the chance level of 1.15m. The chance level was calculated as the RMS across all possible errors at all three distances.

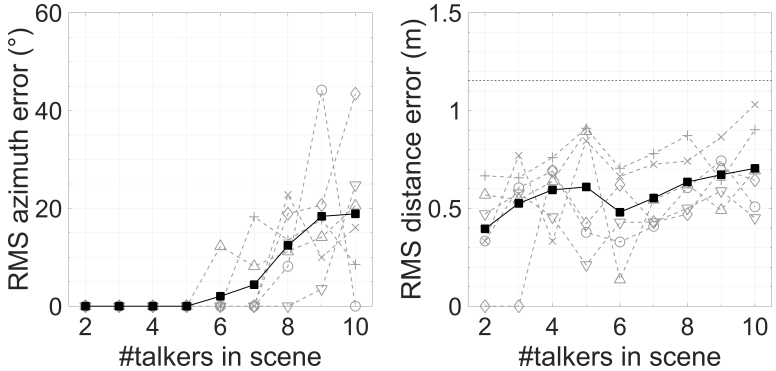


Fig. 5: Localization accuracy, in azimuth (left) and distance (right) RMS errors over number of talkers in scene. Open grey symbols represent individual subjects, averaged over three repetitions, and black squares represent the mean over subjects. The dotted, horizontal line represents the chance level.

Compared to Weller *et al.* (2016), the present study showed a higher localization accuracy, potentially resulting from the smaller spatial range of the response options. The additional visual information and/or the egocentric interface might also have improved the localization accuracy.

Fig. 6 shows the story identification ability of the subjects with respect to the number of simultaneous talkers. The story identification ability is represented as the percentage of scenarios where all stories were recognized. The analysis of a linear mixed model showed a significant effect of the number of talkers ($F(8,148) = 31.4, p < 0.0001$). The ability to identify the correct story decreased gradually with an increasing number of simultaneous talkers. On average, the subjects could identify stories correctly for up to five simultaneous talkers, whereas for eight or more talkers, none of the subjects were able to analyze any scene correctly. Thus, while the subjects were able to accurately identify the number of talkers up to six talkers, the speech recognition ability was only accurate up to five talkers.

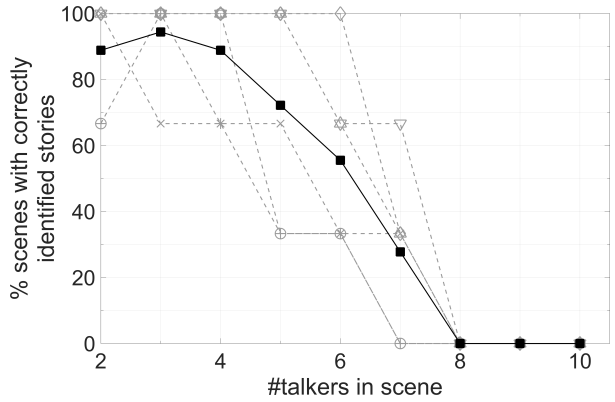


Fig. 6: Percentage of scenes analyzed correctly according to presented stories over number of talkers in scene. Open grey symbols represent individual subjects, averaged over three repetitions, and black squares represent the mean over subjects.

SUMMARY AND CONCLUSION

In the current study, a novel method for evaluating a subject's auditory scene analysis ability in realistic multi-talker scenes was proposed. The method allows to measure sound source identification and localization perception in an audio-visual environment using a loudspeaker-based virtual sound environment and a virtual reality headset. Compared to traditional sentence-based audio-only approaches, this method allows for testing in a more realistic environment and could possibly be used as a tool to evaluate hearing instruments and algorithms.

It was shown that subjects were able to identify the number of talkers in scenes with up to six simultaneous talkers. Furthermore, the localization ability was found to remain

unaffected for scenes with up to seven simultaneous talkers, while the perception of distance did not depend on the number of simultaneous talkers in the scene. The speech recognition ability was found to be worse than the identification of the number of simultaneous talkers.

The VR-based audio-visual method presented in the current study results in improved response accuracy and talker number identification ability compared to previous studies. Future investigations could address how different listening conditions additionally affect motion behavior, such as head rotation and eye gaze.

ACKNOWLEDGEMENTS

The authors would like to thank Marton Marschall, Jakob Nygård Wincentz and Valentina Zapata Rodriguez for feedback during the development of the simulated environments. Furthermore, we would like to thank the talkers for letting us record their voices.

REFERENCES

Ahrens, A., Marschall, M., and Dau, T. (**2019**). "Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments," Hearing Res., **377**, 307-317. doi: 10.1016/j.heares.2019.02.003.

Ahrens, A., Lund, K.D., Marschall, M., and Dau, T. (**2019**). "Sound source localization with varying amount of visual information in virtual reality," PLOS ONE, **14**(3), e0214603. doi: 10.1371/journal.pone.0214603.

Bradlow, A.R., Torretta G.M., and Pisoni, D.B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun., **20**(3-4), 255-27. doi: 10.1016/S0167-6393(96)00063-5.

Bregman, A.S. (**1994**). "Auditory Scene Analysis: The Perceptual Organization of Sound," MIT Press. doi: 10.1121/1.408434.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust united Ac, **86**(1), 117-128.

Favrot, S. and Buchholz, J.M. (**2010**). "LoRA: A loudspeaker-based room auralization system," Acta Acust united Ac, **96**(2), 364-375. doi: 10.3813/AAA.918285.

Kopčo, N., Best, V., and Carlile, S. (**2010**). "Speech localization in a multitalker mixture," J. Acoust. Soc. Am., **127**(3), 1450-1457. doi: 10.1121/1.3290996.

Weller, T., Best, V., Buchholz, J. M., and Young, T. (**2016**). "A method for assessing auditory spatial analysis in reverberant multitalker environments," J. Am. Acad. Audiol., **27**(7), 601-611. doi: 10.3766/jaaa.15109.

Investigating pupillometry as a reliable measure of individual's listening effort

MIHAELA-BEATRICE NEAGU[1,*], TORSTEN DAU[1], PETTERI HYVÄRINEN[1], PER BÆKGAARD[2], THOMAS LUNNER[3,4], DOROTHEA WENDT[1,3]

[1] *Hearing Systems, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark*

[2] *Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark*

[3] *Eriksholm Research Center Denmark, Snekkersten, Denmark*

[4] *Department of Electrical Engineering, Linköping University, Linköping, Sweden*

Pupillometry as a tool indicating listening effort has been extensively analyzed on a group level, but less is known about how reliable pupil dilation is as an indicator of an individual's listening effort. The aim of this study was to investigate the reliability of the pupil dilation measured during a speech-in-noise task as an indicator of an individual's listening effort. The pupil dilation of 27 normal-hearing (NH) and 24 hearing-impaired (HI) participants was recorded while they performed a speech-in-noise test on two different days. Measures of intraclass correlation coefficient (ICC) absolute agreement were considered in the analysis. The ICC was applied to the peak and mean pupil dilation as well as to the different terms resulting from fitting a third-order orthogonal polynomial within growth curve analysis (intercept, 1st order, 2nd order and 3rd order terms), which are assumed to provide further information about temporal changes of the pupil dilation. High values of test-retest reliability were found on some measures of the pupil response. Furthermore, a Bland-Altman analysis was applied as a graphical representation of the reliability of the pupillometry. The results showed different levels of reliability depending on the different features of the pupil response (slope, rise-fall and mean pupil dilation for the HI listeners; rise-fall, delay and mean pupil dilation for NH).

INTRODUCTION

Pupillometry has been considered as a tool for reflecting listening effort, particularly in HI people who typically have higher listening effort than NH listeners in a given condition (Kramer *et al.*, 2006; Wendt *et al.*, 2015). Changes in listening effort as indicated by changes in the pupil size have been demonstrated on a group level (Zekveld *et al.*, 2010; Wendt *et al.*, 2015). The mentioned studies used speech-in-noise tests in combination with pupillometry to examine the impact of intelligibility, signal-to-noise ratio (SNR) and type of noise on listening effort as indicated by changes in the pupil dilation. However, the reliability of pupillometry as an indicator of individual listening effort has not been systematically studied yet.

*Corresponding author: mnea@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

The current study investigated the reliability of pupillometry as an objective listening effort measure in individuals, while they perform a speech-in-noise test. The most common methods for assessing test-retest reliability are the Intraclass Correlation Coefficient (ICC), proposed by Hays *et al.* (1993), and the Bland and Altman (1986) approach. Alhanbali *et al.* (2019) showed a good reliability (ICC $>$ 0.85) of the mean and the peak pupil dilation (PPD). However, the reliability of other pupil dilation characteristics, such as time-dependent features of the pupil response was not considered in their study. Therefore, the present study focused on the reliability of the pupil dilation as a measure of listening effort by considering features such as the average height of the pupil response function, the slope, the rise and fall around the inflection point and the inflexions at the extremities of the function. These features were extracted when applying the growth curve analysis (GCA) model developed by Mirman *et al.* (2008). Furthermore, this study explored the visual representation of the reliability by using the Bland and Altman (1986) approach describing the individual differences of the two visits against their average. Another element of this study was to perform a cluster analysis on the individual responses of the pupil. The purpose of the cluster analysis was to identify the main features of the pupillary response function that could best characterize listening effort.

METHODS

Data set

Two different data sets were analysed as reported in Wendt *et al.* (2018) and Ohlenforst *et al.* (2018). The first data set was collected by Wendt *et al.* (2018) for a group of 27 NH listeners while the second data set was recorded by Ohlenforst *et al.* (2018) for a group of 24 HI listeners. The pupil dilation was recorded while people performed a speech-in-noise test (HINT, Nielsen and Dau (2011)) at 8 different SNRs. Only two subsets were considered for assessing reliability (two out of eight SNRs for each group, NH and HI: 0dB and 4dB, each tested at a different date) and three subsets for the cluster analysis (8dB, 0dB, -8dB for NH and HI). Four to six weeks were considered in between the two different dates, to avoid learning effects with respect to the sentence material since the sentences were repeatedly used. A list of 25 sentences per condition was presented to the participants in a block-based design. The pupil data were processed using MATLAB and R. To remove any initial effects, the first five sentences (out of 25) of the pupil traces from a list were excluded from the analysis. Data cleaning was performed as reported in Wendt *et al.* (2018). Trials with less than 80% reliable data were removed from the analysis and the other traces were baseline corrected. In total, 40 recordings of each individual were compared between the two dates (2x20x27NH, 2x20x24HI). The mean pupil dilation was calculated as the average pupil dilation over the trials. The PPD was calculated between the 3rd and 8th second of the stimulus presentation as in Zekveld *et al.* (2010).

Growth curve analysis (GCA)

To examine temporal changes of the pupil response function for the two different dates, GCA was applied twice for the 2 different dates. According to Mirman *et al.* (2008), GCA fits orthogonal polynomial terms to time series data with the purpose of showing different variations in the function among individuals. To describe the shape of the function, three orthogonal polynomials (p_1, p_2 and p_3) were used. Pupil size was considered as a dependent variable in the model, predicted by a series of fixed and random effects (Eq. 1). The temporal features of the pupil response for the two dates extracted through GCA were considered when calculating test-retest reliability. According to Kalenine *et al.* (2012), the intercept term represents the averaged height of the pupil response, the linear term reflects the slope, the quadratic term reflects the rise and fall around the central inflection point of the response function, and the cubic term reflects the inflexions at the extremities of the curve referred to as delay in the current study. In other words, an estimate of the 3 coefficients and the intercept were obtained, representing the GCA terms of different orders.

$$pupil \sim (p_1 + p_2 + p_3) * participant + (1 + p_1 + p_2 + p_3 | sentence) \qquad \text{(Eq. 1)}$$

ICC

Intraclass correlation coefficient (ICC) is one of the most used reliability indices in test-retest studies. The ICC can reflect either the degree of consistency or the agreement between measurements. The agreement assumes that the values measured on two different dates are expected to be equal for each respondent. Consistency considers that the values measured on two different occasions are correlated in an additive manner. Thus this measure is less relevant in the current analysis, but is nevertheless still reported. ICC agreement was calculated according to Hays *et al.* (1993), as reflected in Eq. 2, where MS_R is the mean square for rows, MS_C is the mean square for columns, MS_E is the mean square for error and n is the number of subjects.

$$ICC_{agreement} = \frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}} \qquad \text{(Eq. 2)}$$

Bland-Altman (BA) approach

To apply the BA approach, the first step was to calculate the limits of agreement (LoA) as the mean $\pm$ 1.96 standard deviation of the two similarly conditioned tests. The plot is designed to show the difference between the two visits against their mean, according to Bland and Altman (1986). The bias is an important aspect in the interpretation of the BA approach, and it was calculated as a mean applied to the difference between the value determined in the first visit and the value determined in the second visit.

Cluster analysis

The aim of applying a clustering algorithm was to identify whether the data points will group according to the different levels of SNRs, or with respect to the different

characteristics of the pupil traces from the individuals. The *k-means* (k=number of clusters) clustering algorithm applied in this study divides the data into different clusters, based on the distance between points (Euclidean distance). Given the distance between all data points and the centroids (the center of the cluster), the measurement will be assigned to the cluster with the nearest centroid.

RESULTS

Pupillometry data

Fig. 1 shows the pupil response of the most representative 10 (out of 27) individual NH listeners for the two test-retest pupil data sets. Significant effects were obtained on the GCA terms (intercept, linear, quadratic and cubic) with small p-values of polynomials estimates for both visits (between $1.18 \cdot 10^{-08}$- 0.009). Similarly, Fig. 2 shows the pupil response of the 10 most representative (out of 24) individual HI listeners for the two test-retest pupil data sets. Significant effects were obtained on the GCA terms as indicated by small p-values for both visits (between $5.32 \cdot 10^{-15}$- 0.012).

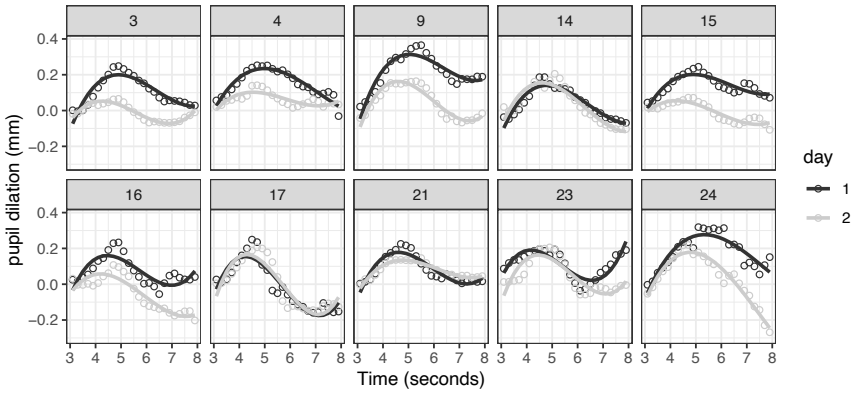


Fig. 1: Growth Curve Analysis for individual NH listeners. Examples of the most relevant pupil responses as a function of time, on the two different visits (black and grey). The open circles represent the actual data, while the lined functions show the fitted GCA model. The numbers in the figure represent the test subjects.

Both figures show that there were individual listeners with comparable pupil responses obtained at the two visits (e.g. NH 14, 17, 21, HI 15). However, there were also individuals showing clearly different responses (e.g. NH 9, HI 17, 19) at the two visits. The dissimilarity could be explained by the difference in the condition tested (0-4 SNR) at the two visits or by other individual factors that need to be identified.

ICC

The classical interpretation of the ICC states that an excellent reliability is reached when ICC values are over 0.75, a good one when ICC is between 0.60 and 0.74 and a fair one for values between 0.4 and 0.59 (Chicchetti, 1994). In the current study,

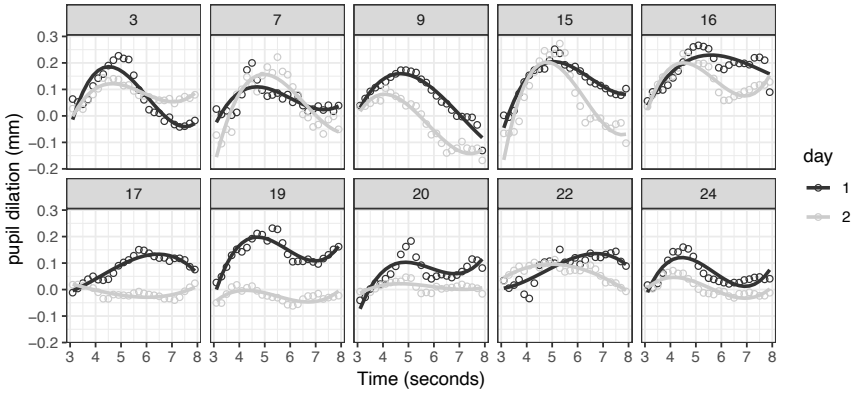


Fig. 2: Growth Curve Analysis for individual HI listeners. Examples of the most relevant pupil responses as a function of time, on the two different visits (black and grey). The open circles represent the actual data, while the lined functions show the fitted GCA model. The numbers in the figure represent the test subjects

the correlation coefficient was calculated for the mean, peak pupil dilation and the time-dependent terms obtained when applying the GCA model. Table 1 shows the ICC values obtained by assessing the reliability of the different features of the pupil response indicating the individual listening effort.

ICC	NH		HI	
	Agreement	Consistency	Agreement	Consistency
GCA Average peak	0.6	**0.62**	0.41	0.54
GCA Slope	0.56	0.58	**0.74**	**0.73**
GCA Rise-fall	**0.60**	**0.69**	**0.64**	**0.66**
GCA Delay	**0.74**	**0.86**	0.27	0.47
Peak pupil dilation	0.48	**0.60**	0.48	**0.64**
Mean pupil dilation	**0.63**	0.59	**0.60**	**0.64**

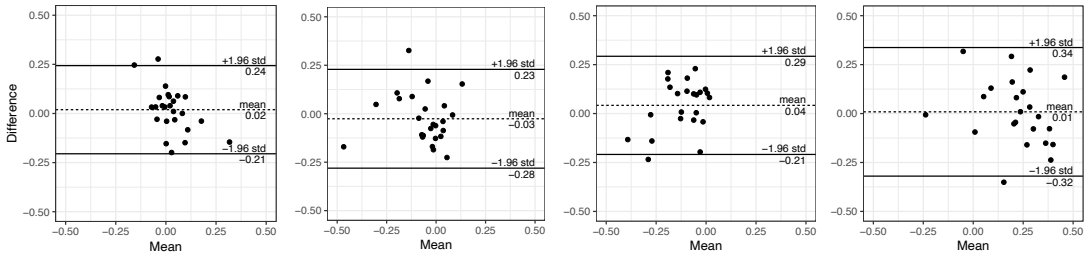
Table 1: ICC agreement and consistency for mean, peak pupil dilation and for different terms of GCA. The ICC values reflect test-retest reliability and bold values are the ones showing good reliability

Different features of the pupil response are reliable for the two listener groups (rise-fall, delay and mean pupil for the NH listener group; slope, rise-fall and mean pupil dilation for the HI listener group).

Bland-Altman visual approach

Fig. 3 shows some examples of the agreement between tests taken on two separate visits as suggested by Bland-Altman. The difference between the two visits is shown against the mean of the two. Sometimes the value obtained on one visit was higher than the other, while sometimes the opposite was found. This contributes to a bias

close to zero. If it is not close to zero, the values of the two visits systematically produce different results, and this represents a low agreement of the method.

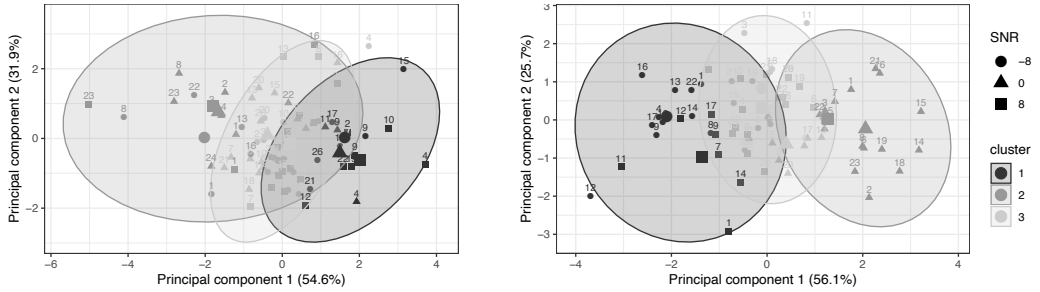


(a) Bland Altman Delay NH **(b)** Bland Altman Rise-fall NH **(c)** Bland Altman Rise-fall HI **(d)** Bland Altman Slope HI

Fig. 3: Example of Bland-Altman plots for NH (a,b) and HI (c,d) groups. The difference between two tests was plotted against their mean. 3a and 3b figures show the BA agreement for delay and rise-fall features (NH group) while the 3c and 3d figures show the BA agreement for the rise-fall and slope features (HI group).

Panels a and b of Fig. 3 show the results for the NH listeners. Most of the data points representing the delay were positioned within the LoA, as in the Fig. 3a. The bias was close to zero showing that there were no significant differences between the two visits. Panels c and d of Fig. 3 show corresponding results for the HI listeners. According to Fig. 3d, the agreement of slope was good, with large LoA values, but the bias was still close to zero. This reflects good agreement, given that the spread of the data points was broader. These results were consistent with the ICC results. Thus, the test-retest reliability was considered as good.

Cluster analysis



(a) Clusters NH 4T GCA 3SNR **(b)** Clusters HI 4T GCA 3SNR

Fig. 4: Clustering of GCA terms for 3 different SNRs (k=3). One point represents one value of the measurement per participant per SNR.

Fig 4. shows the results of clustering the GCA terms for the NH (a) and HI (b) groups at 3 SNR conditions (-8 dB, 0 dB, 8 dB). The choice of the SNR levels to be analysed was made as in Wendt *et al.* (2018). Three different SNRs (out of the eight SNRs

contained by the dataset) with a large range between their PPD were chosen for the cluster analysis. The cluster analysis was applied to both groups, NH and HI, and the results were similar. Listeners with the same SNR were expected to be assigned to the same cluster. According to Fig. 4, the points belonging to the same cluster were data points at different SNRs, suggesting that these clusters could be formed on the base of other factors than those that were considered here.

DISCUSSIONS AND CONCLUSION

This study showed a good reliability for some of the pupil responses features (slope, rise-fall and mean pupil dilation for the HI listeners; rise-fall, delay and mean pupil dilation for the NH listeners). The results obtained with the BA approach were consistent with the ICC results. As Alhanbali *et al.* (2019) also reported, the mean pupil size seems to be a reliable measure for both listeners groups. However, PPD was found to be less reliable than other measures in the current study. Moreover, the time-dependent features of the pupil response seem to be useful for evaluating the reliability of the method. Also, the slope seems to be more reliable for the HI group than for the NH group and it might be an important feature to explore in future studies.

The GCA model reported significant pupil features according to the small p-values of the polynomial estimates. The differences between individual functions obtained with the GCA for the two visits suggest that there could be other factors explaining the variance in the pupil curves (such as listener-dependent factors), apart from the difference in the level conditions (SNR). Zekveld *et al.* (2018) addressed some of these factors and emphasized that further investigations of the individual factors and the effects on the pupil response are required.

The cluster analysis suggested that SNR is not sufficient to classify listening effort, but that there might be some other factors needed for a classification such as listener-dependent factors like age, cognitive abilities and fatigue. Thus, future investigations of the data could consider such individual factors as input features. Furthermore, classification of the listening effort could be modeled with a supervised machine learning algorithm or even a time series analysis.

One of the limitations of the study was the use of different SNR conditions to test the pupil response reliability. It would be valuable to evaluate the reliability of pupillometry in the same acoustic conditions. Eventually, identifying and controlling the factors that can provide insights in cognitive understanding of listening situations will improve the accuracy of pupillometry as an objective measure of listening effort.

Overall, this study showed that rise-fall and mean pupil dilations seem to be important features of the pupil response, demonstrating that the signal is reliable enough in both listener groups. Other time-dependant features seemed to be reliable for one of the groups (Slope for HI and Delay for NH). The reliability results of the method are an important prerequisite for future experimental analysis and for developing pupillometry and the test protocol towards a standardized test for clinical use.

REFERENCES

Alhanbali, S., Dawes, P., Millman, R., and Munro, K. (**2019**). "Measures of listening effort are multidimensional," Ear. Hear., **40**(5), 1084–1097, doi: 10.1097/AUD.0000000000000697.

Bland, J.M. and Altman, D.G. (**1986**). "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet., **327**(8476), 307-310, doi: 10.1016/S0140-6736(86)90837-8.

Cicchetti, D.V (**1994**). "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.," Psychol. Assess., **6**(4), 284-290, doi: 10.1037/1040-3590.6.4.284.

Hays, R.D., Anderson, R., and Revicki, D. (**1993**). "Psychometric considerations in evaluating health-related quality of life measures," Qual. Life Res., **2**(6), 441–449, doi: 10.1007/BF00422218.

Kalenine, S., Mirman, D., Middleton, E.L., and Buxbaum, L.J. (**2012**). "Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts.," J. Exp. Psychol. Learn. Mem. Cogn., **38**(5), pp. 1274-1295, doi: 10.1037/a0027626.

Kramer, S.E., Kapteyn, T.S., and Houtgast, T. (**2006**). "Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work.," Int. J. Audiol., **45**(9), 503–512, doi: 10.1080/14992020600754583.

Mirman, D., Dixon, J.A., and Magnuson, J.S. (**2008**). " Statistical and computational models of the visual world paradigm: growth curves and individual differences," J. Mem. Lang., **59**(4), 475-494, doi: 10.1016/j.jml.2007.11.006.

Nielsen, J.B. and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol., **50**(3), 202-208, doi: 10.3109/14992027.2010.524254.

Ohlenforst, B., Wendt, D., Kramer, S.E., Naylor, G., Zekveld, A.A., and Lunner, T. (**2018**). "Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response," Hear. Res., **365**, 90-99, doi: 10.1016/j.heares.2018.05.003.

Wendt, D., Dau T., and Hjortkjær, J. (**2015**). "Impact of background noise and sentence complexity on processing demands during sentence comprehension.", Front. Psychol., **7**(31), 345, doi: 10.3389/fpsyg.2016.00345.

Wendt, D., Koelewijn, T., Książek, P., Kramer, S.E., and Lunner, T. (**2018**). "Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test," Hear. Res., **369**, 67-78, doi: 10.1016/j.heares.2018.05.006.

Zekveld, A., Kramer, S., and Festen, J. (**2010**). " Pupil response as an indication of effortful listening: the influence of sentence intelligibility.," Ear. Hear., **31**(4), 480-490, doi: 10.1097/AUD.0b013e3181d4f251.

Zekveld, A.A., Koelewijn, T., and Kramer, S.E. (**2018**). " The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge," Trends. Hear., **22**(4412), doi: 10.1177/2331216518777174.

Potential of self-conducted speech audiometry with smart speakers

JASPER OOSTER[1,4,*], KIRSTEN C. WAGENER[2,4], MELANIE KRUEGER[3,4], JÖRG-HENDRIK BACH[2,3,4] AND BERND T. MEYER[1,3,4]

[1]*Medizinische Physik, Carl von Ossietzky Universität, Oldenburg, Germany*
[2]*Hörzentrum GmbH, Oldenburg, Germany*
[3]*HörTech gGmbH, Oldenburg, Germany*
[4]*Cluster of Excellence Hearing4all, Germany*

Speech audiometry in noise based on matrix sentence tests is an important diagnostic tool to assess the speech reception threshold (SRT) of a subject, i.e., the signal-to-noise ratio corresponding to 50% intelligibility. Although the matrix test format allows for self-conducted measurements by applying a visual, closed response format, these tests are mostly performed in open response format with an experimenter entering the correct/incorrect responses (expert-conducted). Using automatic speech recognition (ASR) enables self-conducted measurements without the need of visual presentation of the response alternatives. A combination of these self-conducted measurement procedures with signal presentation via smart speakers could be used to assess individual speech intelligibility in an individual listening environment. Therefore, this paper compares self-conducted SRT measurements using smart speakers with expert-conducted lab measurements. With smart speakers, the experimenter has no control over the absolute presentation level, mode of presentation (headphones vs. loudspeaker), potential errors from the automated response logging, and room acoustics. We present the differences between measurements in the lab and with a smart speaker for normal-hearing, mildly hearing-impaired and moderate hearing-impaired subjects in low, medium, and high reverberation.

INTRODUCTION

Being able to understand speech, especially in noisy conditions, is a crucial factor of social interaction and is often limited for hearing impaired listeners, which can reduce their quality of life. An early diagnosis of the hearing loss can ease this limitation by an early supply of a hearing aid (Arlinger, 2003). A reliable measurement tool with a high accuracy for quantifying the ability of speech understanding in noise is available through matrix sentence tests (Kollmeier *et al.*, 2015). Due to the closed-vocabulary construction of this test, it allows for an unsupervised measurement with a graphical user interface. Nevertheless, such an interface excludes subjects who cannot read (children, illiterate, visually impaired). Hence, we focus on a system that uses

*Corresponding author: jasper.ooster@uni-oldenburg.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

only acoustic communication cues, i.e., speech. This we propose to do with automatic speech recognition (ASR) for the response logging (Ooster *et al.*, 2018). While that system created for clinical (and relatively controlled) environments, an ASR-based conduction has also the potential of increasing the accessibility by performing self-measurements at home.

Smart speakers such as Amazon's *Echo*, Apple's *HomePod* or *Google Home* have the potential of bringing such a test to a broader subject base, since they provide a good audio quality and have a built-in dialogue manager including an ASR component. There have already been approaches to use smart home systems for medical purposes, e.g., to provide acoustic cues to support dementia patients' memory (Boumpa *et al.*, 2019) or to support elderly people in their physical therapy (Vora *et al.*, 2017).

In this work, we present a smart speaker application for measuring and validating the speech reception threshold (SRT), i.e. the signal-to-noise ratio (SNR) corresponding to 50% intelligibility with the matrix sentence test. Smart speaker-based measurements have several differences compared to established clinical setups: (i) We use a high-quality speech synthesis instead of the natural speech files that are protected by copyright, (ii) the sound is presented via the speaker in a reverberant environment, (iii) compressed audio files are presented, and (iv) the listener's response is transcribed via ASR and not logged by an audiometrist. In a first proof-of-concept study the smart speaker-based measurement was already evaluated in a single office room with six normal-hearing (NH) subjects (Ooster *et al.*, 2019). However, the accuracy for hearing-impaired (HI) subjects, which is crucial for speech audiometry, has not been determined. Furthermore, in a real use case, the acoustic conditions in which the test is conducted can exhibit large variability. The user can be asked to avoid any background noise (in order to get an accurate test result), but it is often not possible to easily change the acoustics of the room where the smart speaker is placed. Therefore, in this study, we evaluate the measurement procedure when testing mildly and moderately HI subjects and secondly quantify the influence of acoustic conditions by conducting the experiments with three different kinds of reverberation.

METHODS

Matrix sentence test

The speech audiometric test used in this study is the German matrix sentence test (Wagener *et al.*, 1999). During testing, the subject hears sentences in stationary speech-shaped noise, and the SNR is dynamically adapted to reach the SRT after presenting 20 noisy sentences. The final measurement outcome is estimated by a likelihood fit of a psychometric function to the 20 data points of the whole measurement. The words of the stimulus sentences are randomly selected from a five-by-ten word matrix in order to create sentences with the structure *Name Verb Numeral Adjective Object*. Through this procedure, the individual words of the sentence are independent. This results in a low test-to-retest standard deviation of 1 dB for HI subjects (Wagener and Brand, 2005) and 0.5 dB for NH subjects (Brand and Kollmeier, 2002).

The smart speaker application

The elements of the smart speaker application for the automated SRT measurement are shown in the overview Figure 1 (Ooster *et al.*, 2019). The application is implemented with the Alexa Skill Developer Kit in Python (`github.com/alexa/alexa-skills-kit-sdk-for-python`). When the measurement application is

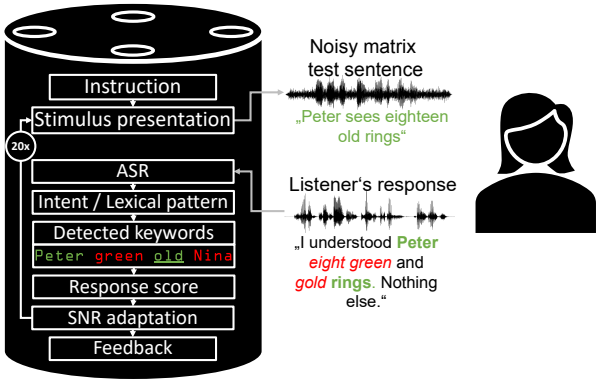


Fig. 1: Overview of the smart speaker measurement application.

started, the subject first hears an instruction about the general measurement procedure and the structure of this hearing test. These instructions are based on the guidelines for the clinical application of the matrix sentence test. During the measurement itself the dialogue manager of the smart speaker uses *intents*, which are derived from lexical patterns on the ASR output to take the next action within the measurement application. The core intent of the measurement application is the response to a matrix stimulus sentence. The lexical pattern to invoke this intent are based on realistic responses obtained in previous work (Ooster *et al.*, 2018). Based on this intent, the keywords in the subjects response are collected; the SNR for the next presentation is adapted based on the resulting score. Since the original speech material of the matrix sentence test is protected by copyright, we used a synthesized version of the sentences from the female German matrix sentence test, which was evaluated in a previous study (Nuesse *et al.*, 2019). All stimulus matrix sentence audio files were premixed with the speech shaped noise at steps of 0.1 dB and converted to the mp3 data format (MPEG version 2, 48 kbps, 16 kHz) in order to be properly played on the smart speaker.

Evaluation measurements

The prototype application was evaluated using an *Amazon Echo Plus 2nd Generation* loudspeaker. The measurements were conducted with subjects with three different hearing-profiles; normal hearing, mildly hearing impaired, and moderately hearing impaired. The subjects were categorized with pure tone average (PTA) criteria from .5 to 4 kHz (Mathers *et al.*, 2001). All subjects were paid for the participation in this study. The smart speaker measurements were conducted in a room that uses

distributed microphones and loudspeakers to simulate different room acoustics. The subjects were sitting in the center of the room with the smart speaker in front on a table at a distance of 2 m. At the beginning of the measurement, the subjects were asked to adjust the volume of the speaker to an easy intelligibility of the speech assistance's voice. To account for different acoustic conditions, rooms with different reverberation times T_{30} were simulated: *Living Room* ($T_{30} = 0.51$s), *Poor Classroom* ($T_{30} = 1.12$s) and *Concert Hall* ($T_{30} = 1.52$s). Every time the room settings were changed, the subjects heard four random sentences at different SNR so they could adapt to the new room and also could adjust the volume of the speaker. The subjects were always allowed to change the volume of the speaker again during the measurement. The subjects were invited for two measurement sessions each with nine SRT measurements in total, as described in Table 1. Overall, each subject conducted 16 measurements with the smart speaker application as well as two clinical reference measurements. The clinical reference measurements were conducted in an isolated sound booth,

Room A				Room B		Room C		Booth
Training1	Training2	Test1	Test2	Test3	Test4	Test5	Test6	Reference

Table 1: Measurement procedure during one of the two sessions for each subject. While the reference measurement with the clinical setup was always in the end, the order of the room settings during the smart speaker measurement was randomly chosen for each subject.

with a calibrated and equalized loudspeaker, the original, female, natural voice for the stimulus sentences (Wagener *et al.*, 2014) and a human supervisor for response scoring. At the end of each measurement session, all recorded audio files in the cloud of the smart speaker were deleted so that the ASR system is not adapted to that speaker for the measurement with the next subject. Parallel to the measurements, a human supervisor scored the subjects responses to have the true value for the scoring for each sentence (assuming that the experienced human supervisor produces no errors when logging the reported words). These true transcripts were later used to quantify the errors of the smart speaker ASR in terms of the score insertion rate (SIR) and the score deletion rate (SDR), i.e., the errors that could actually have an influence on the SRT by inserting or deleting a word. They are defined by

$$SIR = \frac{N_{score\ insertions}}{N_{score}}; SDR = \frac{N_{score\ deletions}}{N_{score}}, \qquad \text{(Eq. 1)}$$

where the number of errors $N_{score\ insertion}$ and $N_{score\ deletion}$ are normalized by the number of correctly repeated matrix sentence test words in the subject's response N_{score}. The order of the words is neglected in this error metric. The full error rates in the classical sense of an ASR system cannot be calculated since the full transcript (including non score relevant words) was not created. Details on the evaluation metrics can be found in Ooster *et al.* (2018).

RESULTS

The evaluation measurements were conducted with 5 subjects from each subject group, resulting in a total of 15 subjects as described in Table 2. Figure 2 describes the

	Normal-hearing	Mild hearing Loss	Moderate hearing loss
N (f/m)	5 (2/3)	5 (1/4)	5 (2/3)
Age	62 +/- 6 years	68 +/- 1 years	60 +/- 11 years
PTA	13 +/- 7 dB	29 +/- 5 dB	47 +/- 8 dB

Table 2: Description of the subjects who participated in the evaluation.

SRT measurement accuracy of the measurement with the smart speaker application compared to clinically acquired estimates. While the black line indicates a potential perfect match between the clinically measured value and the value estimated with the smart speaker application, most of the measured points are above this line. This highly

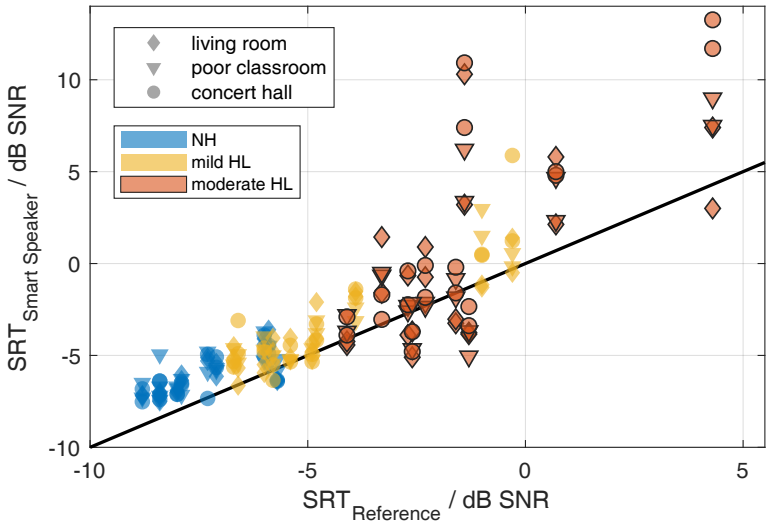


Fig. 2: The measured SRTs with the smart speaker application plotted vs. the SRTs measured with the clinical reference setup in the same session. Color depicts the hearing loss categorization based on the PTA criterion; the shape of data points denotes the room setting during the smart speaker measurement.

significant bias (paired-sample t-test, $p = 3.1 \cdot 10^{-15}$) that amounts to 1.38 dB on average is constant over the acoustic conditions and subjects groups. We did not find any significant difference with the t-test between the different room settings and the subject groups. The intra- and inter-subject standard deviation (SD; 1.37dB/3.79dB) are higher than with the clinical setup (0.76dB/3.11dB) over all subject groups and acoustic conditions. The inter- and the intra-subject SD varied slightly in the three

different acoustic conditions, with the highest increase of the inter-subject SD in the *Concert Hall* condition of about 1 dB. The intra-subject SD is increased by 0.39 dB in this condition. For the mildly HI subjects the intra- and inter-subject SD is slightly increased in comparrsion to the NH subjects by 0.32 dB and 0.30 dB, respectively. For the moderately HI subjects both the intra- and the inter-subject SD is increased by 1.45 dB and 1.76 dB, respectively, in comparison to the mildly HI subjects. This is also due to the fact that two measurement sessions failed completely: In one of these measurement sessions, the ASR performance was with 19.8% SDR (6.9% SIR) very low in comparison to the other subjects, which resulted in an bias of 8.3 dB and a SD of 3.3 dB. In the next measurement session of this subject, the ASR performance was better (at 8.5% SDR and 4.1% SIR), which is consistent with a much higher measurement accuracy with a bias of -0.2 dB and a SD of 1.2 dB. The second inaccurate measurement session is not explainable by the error rates of the ASR system (SDR = 7.4%, SIR = 5.8%). However, we noticed that the subject was speaking very quietly which resulted in several terminations of the measurement application. Although the subject spoke in normal volume towards the end of the session, the terminations could have had a large effects on the ability of the subjects to focus on the listening task. When excluding these two subjects the increase of the inter- and intra-subject SD for the moderately HI subjects goes down to 0.33 dB and 0.70 dB, respectively.

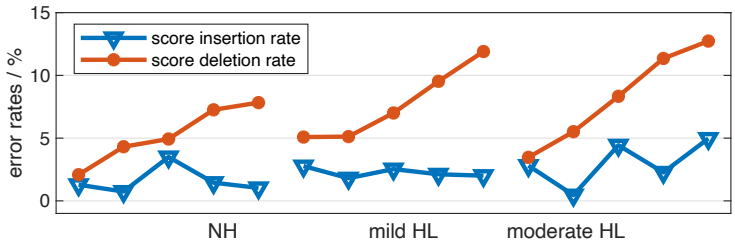


Fig. 3: The ASR performance of the smart speaker for each of the 15 subjects.

The ASR performance of the smart speaker for all subjects is shown in Figure 3. With an average of SDR = $(8.0 \pm 3.2)\%$ are the score deletion errors significantly higher for the HI subjects than the score deletion error of the NH subjects, which had an average of SDR = $(5.3 \pm 2.3)\%$ (two sample t-test, equal variances not assumed, $p = 3.0\%$). Three of the HI subjects showed SDRs above 10%; through post analysis (as discussed above) ASR errors for one of those subjects could be attributed to a strong decrease of the SRT measurement accuracy. The score insertion errors are below 5% for all of the subjects and no significant difference was found between NH and HI subjects.

DISCUSSION

In this study, we investigated the SRT measurement accuracy with a smart speaker-based application in three different acoustic conditions and with three different subject groups. In our previous study regarding the speech-controlled automated matrix

test (SAMT; Ooster *et al.*, 2018), we didn't find any significant decrease of the measurement accuracy conditioned by the errors from the ASR system. In this study, an ASR system, that was not that fine-tuned to the words of the matrix test was used in a more challenging acoustic condition (far-field recognition with reverberation) and therefore the obtained error rates are higher. For one subject this resulted in a very inaccurate measurement, but overall the observed intra-subject SD is very similar to the one of the clinical application. The only subject group with an increase intra-subject SD are the moderately HI subject. This subject group has very similar ASR error rates as the mildly HI subjects, so the decreased replicability of the measurement outcome seems not to be indicated by the smart speaker based measurement itself, but presumably due to an insecurity of subjects in terms of speech-based interaction with a speaker. The obtained ASR error rates during this study represent a lower boundary of the ASR performance, since in a real use case the ASR system should be adapted to the specific user and secondly owners of smart speakers are probably used to speech-based inputs and normal patterns of interaction. The moderately HI subjects showed a decreased measurement accuracy with the smart speaker application, but most of the variance during the measurement with the moderate HI subjects is towards higher (worse) SRTs and therefore would not change the screening result.

CONCLUSIONS

In this paper, we have shown that speech audiometry conducted with a smart speaker for at-home screening of hearing deficits is possible with an intra-subject SD of 1.37 dB. The bias between the clinical and the smart-speaker measurement is significant, but consistent across subject groups and room settings, and no significant difference was found between the groups and conditions, respectively. While normal-hearing and mildly hearing-impaired subjects showed a very similar measurement accuracy as the clinical reference measurement, the inter- and intra-subject SD is increased for moderately hearing-impaired subjects by 1.39 dB and 1.89 dB, respectively. This was attributed to the results for single subjects, whose speech produced high ASR error rates or was too low to properly conduct the measurement. When excluding these subjects from the analyses the increase of inter- and intra-subject SD goes down to 0.33 dB and 0.70 dB, respectively and the overall intra-subjects SD goes down to 0.91 dB which is comparable to the intra-subject SD with the clinical measurement setup of 0.67 dB.

In future work, we will develop an SRT-based criterion to produce a recommendation for the test user (e.g., to seek advice from an audiometrist), based on ratings of his or her performance for the test. This will require a larger number of listeners to be tested to establish a reliable statistical foundation for such a recommendation.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - Project

ID 390895286 and the CRC TRR 31, Transfer Project T01.

REFERENCES

Arlinger, S. (**2003**) "Negative consequences of uncorrected hearing loss - a review." Int. J. Audiol., **42**(2), S17–S20. doi: 10.3109/14992020309074639

Boumpa, E., Gkogkidis, A., Charalampou, I., Ntaliani, A., Kakarountas, A., and Kokkinos, V. (**2019**). "An Acoustic-Based Smart Home System for People Suffering from Dementia". Technol., **7**(1), 29. doi: 10.3390/technologies7010029

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests". J. Acoust. Soc. Am., **111**(6), 2801–2810. doi: 10.1121/1.1479152

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., and Wagener, K. C. (**2015**). "The multilingual matrix test: Principles, applications, and comparison across languages: A review". Int. J. Audiol., **54**(sup2), 3–16. doi: 10.3109/14992027.2015.1020971

Mathers, C., Smith, A., and Concha, M. (**2001**). "Global burden of hearing loss in the year 2000." Global Burden of Disease 2000, **18**(4), 1–30.

Nuesse, T., Wiercinski, B., Brand, T. and Holube, I. (**2019**). "Measuring Speech Recognition With a Matrix Test Using Synthetic Speech." Trends Hear. doi: 10.1177/2331216519862982

Ooster, J., Huber, R., Kollmeier, B., and Meyer, B. T. (**2018**). "Evaluation of an automated speech-controlled listening test with spontaneous and read responses." Speech Commun., **98**, 85–94. doi: 10.1016/j.specom.2018.01.005

Ooster, J., Porysek Moreta, P. N., Bach, J.-H., Holube, I. and Meyer, B.T. (**2019**). "Computer, test my hearing: Accurate speech audiometry with smart speakers" Proceedings of the Interspeech 2019, Graz, Austria. doi 10.21437/Interspeech.2019-2118

Vora, J., Tanwar, S., Tyagi, S., Kumar, N., and Rodrigues, J. J. P. C. (**2017**). "Home-based exercise system for patients using IoT enabled smart speaker". In 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom. doi: 10.1109/HealthCom.2017.8210826

Wagener, K. C., Kühne, V., and Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I-III: Design, Optimierung und Evaluation des Oldenburger Satztests" (Development and evaluation of a German speech intelligibility test. Part I-III: Design, optimization and evaluation of the Oldenburg sentence test). Z Audiol., **38**(1-3), 4-15, 44-56, 86-95.

Wagener, K. C., and Brand, T. (**2005**). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters". Int. J. Audiol., **44**(3), 144–156. doi: 10.1080/14992020500057517

Wagener K., Hochmuth S., Ahrlich M., Zokoll M. and Kollmeier B. (**2014**), "Der weibliche Oldenburger Satztest". 17. DGA Jahrestagung, Oldenburg, Germany.

"Yes, I have experienced that!" – How daily life experiences may be harvested from new hearing aid users

KATJA LUND[1,*], RODRIGO ORDOÑEZ[1], JENS BO NIELSEN[2], AND DORTE HAMMERSHØI[1]

[1] *Department of Electronic Systems, Signals and Information Processing, Aalborg University, Aalborg, Denmark*

[2] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

Both auditory and non-auditory aspects of the rehabilitation process play a role in successful hearing aid uptake. The sound may be experienced differently in the clinic compared to daily life and the skills and knowledge related to HA use vary from patient to patient. The aim of the present study is to assess daily life experiences of new hearing aid users and to explore ways to utilize these assessments in a follow-up situation. The approach is based on online reporting, where the patients over a period of two months "swipe" through 453 possible experiences related to HA use. Seventeen patients volunteered to register experiences for a period of two months, and participated in a follow-up interview, where the registered data were presented. Results suggest that data can shed light on the development within various categories of hearing aid experience and promote reflection on the hearing rehabilitation process.

INTRODUCTION

New users often experience challenges when first fitted with hearing aids (HA). The challenges may be related to various factors such as benefit, comfort, maintenance, attitude, dexterity etc. (Bertoli *et al.*, 2010; Hickson *et al.*, 2014; McCormack *et al.*, 2013). Due to the diversity of factors related to use, it is difficult to get reliable data on patient experiences during the first months of use. The aim of the present study is to collect data on what new HA users experience over a two-month period to gain insight into the experiences they have and how these experiences develop over time. The research question is thus: *How may daily life experiences be harvested from new adult hearing aid users and do these data represent added value for patients or hearing care professionals?*

By presenting a tool for logging a wide range of everyday life experiences related to HA use, we expect to enable registrations of positive and negative experiences among new HA users. A hypothesis is that the tool will also provide users with a vocabulary to give words to these experiences in a follow-up interview. Fig. 1, shows the interface developed for this purpose.

*Corresponding author: klu@es.aau.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

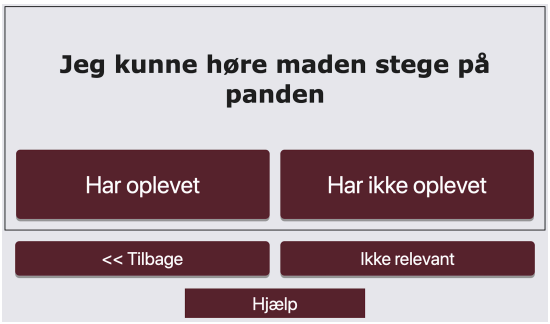


Fig. 1: Example of possible experience: 'I could hear the food sizzling on the frying pan', and the response options presented in the interface: 'Have experienced', 'Have not experienced' or 'Not relevant'.

METHODS

Twenty nine participants were included in the study. Seven female and ten male participants, in all fifteen new and two experienced HA users, completed the pilot study by logging experiences over a period of two months. Twelve participants withdrew from the study. The mean age was 67 (youngest 54, oldest 83) years. Fifteen participants agreed to face-to-face follow-up interviews and two gave written feedback instead of an interview.

Data were collected online and consisted of 453 pre-fabricated sentences representing experiences related to HA use. The sentences were developed through a participatory design process involving various stakeholders such as patients, practitioners and hearing care experts (Lund *et al.*, 2020).

The sentences representing possible experiences were presented using an online tool that allowed individual logging. The participants could log as many sentences as they wanted and spend as much time as they liked. The online tool allowed participants to 'swipe' through the experiences, which were presented in random order (with random repetition after completing all 453 experiences), see example in Fig. 1. Subjects would then be able to indicate, whether they had had a given experience or not, and were also allowed to skip decision by answering "irrelevant". The users were encouraged to log experiences every day, or at least once a week. Participants were hoped to find their own preferred rhythm for logging that would suit their specific daily routines.

After 30 responses, a break would be suggested. Users could also take breaks whenever they preferred, and would continue where they left off by signing in with their unique credentials. The possible experiences were subdivided into the following 13 categories containing between 9 and 85 possible experiences: 1) speech intelligibility, 2) spatial, 3) sound quality, 4) adaptation to new sounds, 5) noise, 6) loudness, 7) fatigue, 8) use, 9) tinnitus, 10) handling and maintaining the HA, 11) fit, 12) support, and 13) quality of life.

The participants did not receive feedback on data and did not have access to the data during the period of logging. If there were no log-activity for more than two weeks, a reminder was sent by email to the participants, who had shared email contact information in advance. After two months of logging experiences, an individual interview was made by one of the researchers (Lund). An interview-protocol was followed and log-data for each patient were visualized (as shown in Fig. 2) and reflected on. The interview-protocol included the following questions:

- Have you experienced improvement over the two months / challenges that have been overcome?
- Do you experience continuing challenges? If yes, please describe them.
- Did you have contact with the clinic within the two months? If yes, what was it about?
- What has it been like to register your experiences? Has it been easy / difficult etc.?
- Did you involve anyone else in the log-activity / talk to someone about it?
- Do you feel that the activity of logging experiences has made you act differently? Have you, for example, become more aware of something because it was presented to you in the sentences?
- Have there been sentences that were completely irrelevant to you or something that annoyed you when logging the experiences?
- Have you had experiences, which you would have liked to log, but which were not represented in the sentences (missing sentences)?
- Do you think you will use your HAs in the future?
- Would you have liked to continue logging your experiences if possible?

After answering the questions in the protocol, the log-data were presented to the participants. Some of the positive and negative experiences that dominated in the beginning of the log-period and after the two-month period were read aloud to help the participant remember the experience referred to and elaborate on it.

RESULTS

Log-data

The participants who completed the study logged between 170 and 3171 experiences in the period from March until August 2019. The approaches to logging differed in both number of logs and the distribution of log-activities as shown in Fig. 3.

The average number of answers per patient is 753. This is, however, based on some of the patients being very active (example Figs. 4 and 5) and others being very inactive (example Figs. 6 and 7).

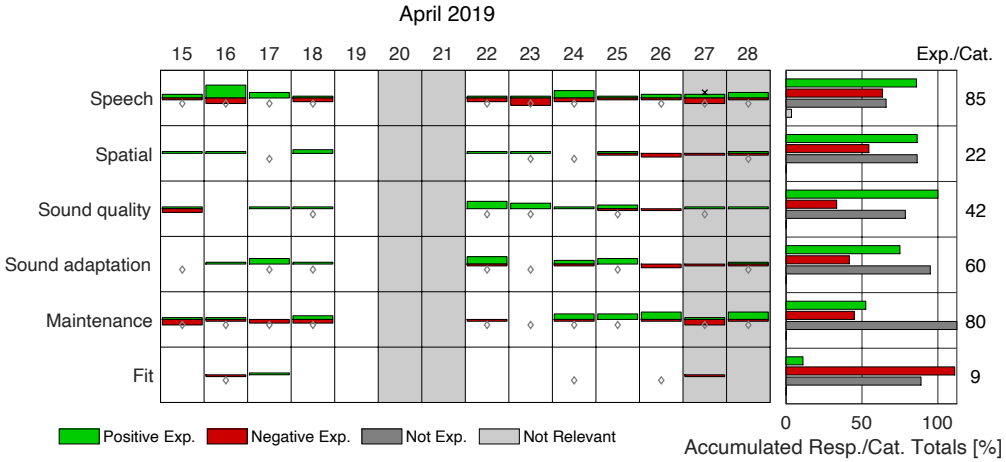


Fig. 2: Data for each participant displayed in a calendar-like format showing the logs of each day and month (only two weeks displayed here). Categories are listed to the left (only six of the 13 categories are shown here) and to the right the accumulated scores of positive, negative, not experienced and not relevant experiences are listed. The number to the far right represents the number of sentences in each category.

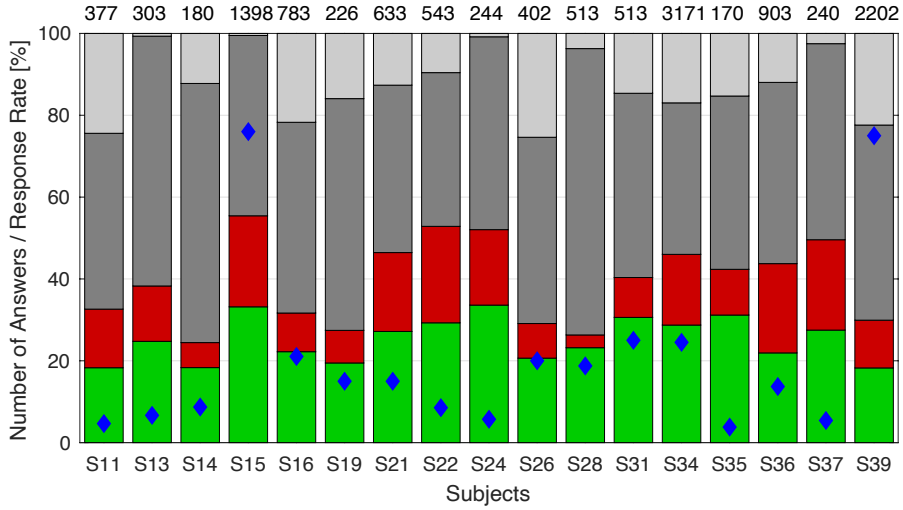


Fig. 3: The number of positive (green), negative (red), not experienced (dark grey) and not relevant (light grey) answers are shown in percentage for each participant. The blue diamond shows the response rate where 100% represents log-activity every day of the log period (10%: log-activity two days every three weeks; 20% log-activity three days every two weeks). The number of answers for each participant is shown with the numbers at the top of each bar.

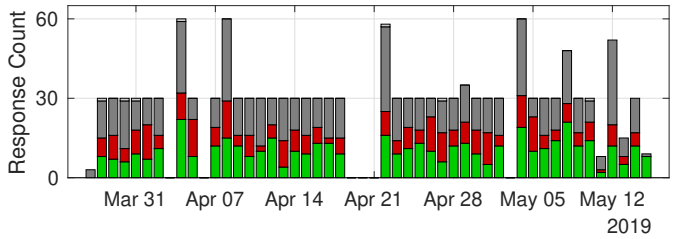


Fig. 4: Response rate and number of responses per system entry for one of the most active participants (S15, new HA-user). The period from the first to the last entry was seven weeks. The number of responses are displayed on the y-axis while the log-period is shown along the x-axis.

The distribution of responses in Fig. 4 indicates that participants stopped logging, when the system suggests a break after 30 or 60 sentences. This is a general observation across subjects, regardless the number of answers each patient made.

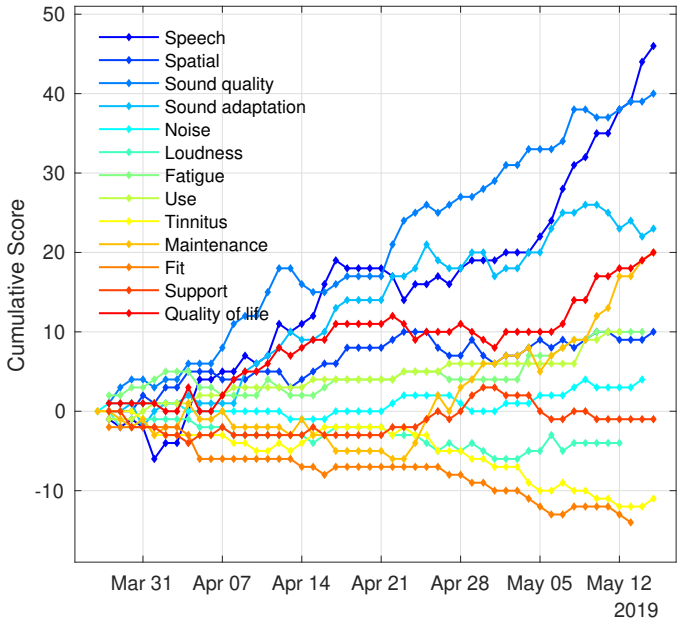


Fig. 5: Cumulative score for the same participant as in Fig. 4 (S15). Zero on the y-axis represents an equal amount of positive and negative responses in a category while the numbers above zero represent a positive cumulative score and the numbers below represent a negative score.

Fig. 5 shows that the patient experienced challenges with fit and tinnitus. This was a general observation, which applied to several of the patients. In particular, fit was a challenge to almost all participants. Negative experiences with tinnitus were not necessarily related to the new HA and the challenges could even have decreased after fitting despite a negative score in the accumulated data. No objective before-after

measure was available for critical assessment of the suggested progression. Most patients registered positive accumulated scores in the categories 'Speech', 'Sound adaptation' and 'Sound quality' (as S15), which were not verified objectively either.

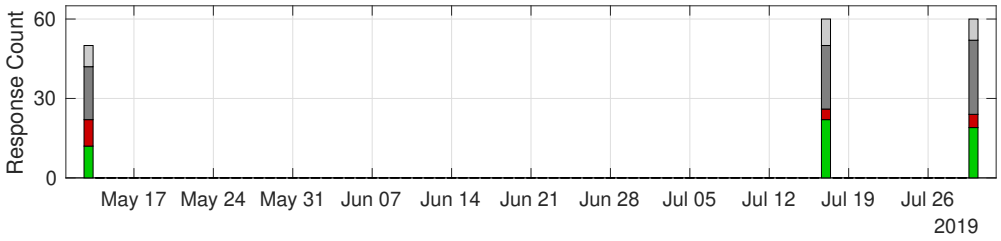


Fig. 6: Response rate and number of responses per system entry for one of the least active participants (S35, new HA-user). The period from the first to the last entry was 12 weeks.

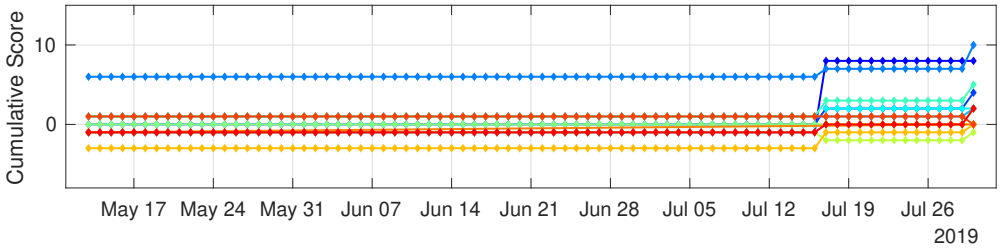


Fig. 7: Cumulative score for the same participant as in Fig. 6 (S35). The x-axis shows the period from the first to the last patient log. The dots along the lines indicate the days of the log-period. Each line represents a category. Changes along the lines indicate that the patient has registered experiences in the system.

In the data analysis in Fig. 7, the logs have been accumulated to indicate whether there is progression over time. Some of the lines are above zero (the accumulated score suggests that the patient has had more positive than negative experiences) on the days of the three registrations, and some are below zero (the accumulated score shows more negative than positive experiences). Three registrations seem, however, not to be sufficient data to draw a proper picture of progression over time.

A daily rhythm of logging experiences seems favorable to reach an amount of log-data that shows either a positive or a negative progression over time as in participant S15 (Figs. 4 and 5). S35 (Figs. 6 and 7) had only three system entries over the period of 12 weeks reaching a total of 170 answers. The cumulative score shows to some extend an overweight of positive experiences, but gives only a vague indication on progression due to the low response rate.

Follow-up interviews

In the follow-up interview, the participants were asked to elaborate on the activity of logging: Sixteen of the 17 participants said that the system was easy to operate. Nine participants expressed that sentences presented were frequently not relevant to them, and it was therefore difficult to relate to them. Four had become aware of functionalities in the HA, which they were not familiar with when first fitted. Three participants experienced technical challenges such as difficulties with logging on to the system, which were overcome during the log-period.

The participants were asked if they were missing sentences representing situations they had been in during the log-period. Five had missed sentences for describing experiences related to the acoustical environment, the HA falling off, HA comfort, the sound of paper, television volume and the sound quality of different voices, whereas 12 could not think of any situations they would have liked to register, which did not occur in the sentences.

When looking at the log-data, examples of responses were read aloud, which lead to reflections among participants with both high and low response rates. In general, all participants who agreed to the face-to-face interview reflected on the sentences read aloud. A reaction could, for example, be related to the sentence 'I was overwhelmed by the sound as I went to a restaurant'. One of the participants said: *"That's about getting used to sounds, which I haven't been able to hear. I'm looking forward to going hunting in the end of August. For many years, I haven't been able to hear if something was rustling in the scrub."* The experienced users also reflected on sentences such as 'I heard noise coming from the coffee machine' e.g. by saying: *"I could hear it and I'm not used to that. Even when I'm sitting in the living room I can hear it in the kitchen. That's an alright experience because it means that I hear more than I did before."* The reflections were generally positive, also when reflecting on a negative response and focus was to a large extend on progression.

CONCLUSIONS

As expected, the number and frequency of patient logs differed. The suggested breaks after 30 and 60 responses were decisive for the number of logs per time a patient would access the system. The majority of the participants had negative accumulated scores in the categories 'Fit' and 'Tinnitus' and positive accumulated scores in 'Speech', 'Sound adaptation' and 'Sound qualities'. It is unknown whether this reflects an improvement after HA fitting compared to unaided hearing, as there are no objective measures before and after the period of observations. Data may however show progression (positive / negative) in various categories of the hearing experience during a two-month period after fitting.

Despite challenges with relating to some of the sentences, a part of the participants experienced that the sentences had caused them to think more about different aspects of use. The participants were able to express their experiences and to reflect on data as

well as describe aspects of the process from fitting until follow-up. A daily rhythm of logging experiences seems favorable to reach an amount of log-data that shows either a positive or a negative progression over time. The tool was used to tap the memory of and promote reflections in participants with both many and few responses as well as new and experienced HA-users. The results apply for the method and the options for data retrieval.

ACKNOWLEDGEMENTS

The authors would like to thank the participating patients, and the staff of the Audiological Department of Aalborg University Hospital. The study is part of the BEAR project funded by Innovation Fund Denmark and partners (incl. Force Technology, Oticon, GN Hearing and Widex Sivantos Audiology). Funding and collaboration is sincerely appreciated.

REFERENCES

Bertoli, S., Bodmer, D., and Probst, R. (**2010**). "Survey on hearing aid outcome in Switzerland: Associations with type of fitting (bilateral/unilateral), level of hearing aid signal processing, and hearing loss," Int. J. Audiol., **29**(5), 333–346. doi: 10.3109/14992020903473431.

Hickson, L., Meyer, C., Lovelock, K., Lampert, M., and Khan, A. (**2014**). "Factors associated with success with hearing aids in older adults," Int. J. Audiol., **53** (sup1), S18–S27. doi: 10.3109/14992027.2013.860488.

Lund, K., Ordoñez, R. P., Nielsen, J. B., and Hammershøi, D. (**2020**). "Sentence-based experience-logging in new hearing aid users," Am. J. Audiol., Special Issue: Internet and Audiology (in press).

McCormack, A. and Fortnum, H. (**2013**). "Why do people fitted with hearing aids not wear them?" Int. J. Audiol., **52**(5), 360–368. doi: 10.3109/14992027.2013.769066.

Speech related hearing aid benefit index derived from standardized self-reported questionnaire data

SREERAM KAITHALI NARAYANAN[1,*], TOBIAS PIECHOWIAK[2], ANNE WOLFF[3], SABINA S HOUMØLLER[4], VIJAYA KUMAR NARNE[5], GÉRARD LOQUET[6], DAN DUPONT HOUGAARD[3,6], MICHAEL GAIHEDE[3,6], JESPER HVASS SCHMIDT[4], & DORTE HAMMERSHØI[1]

[1] *Department of Electronic Systems, Signals and Information Processing, Aalborg University, 9220 Aalborg, Denmark*

[2] *GN Hearing A/S, Ballerup, Denmark*

[3]*Department of Otolaryngology, Head and Neck Surgery and Audiology, Aalborg University Hospital, Aalborg, Denmark*

[4] *Department of Oto-rhino-laryngology, Odense University Hospital, Odense, Denmark*

[5] *Institute of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark*

[6] *Department of Clinical Medicine, Aalborg University, Aalborg, Denmark*

Speech understanding in noisy environments has been the most desired hearing-aid (HA) benefit sought by HA users. This paper examines the possibility of developing a speech-related HA benefit index from the speech-related questions in the self-reported questionnaire data. One question from Health-Related Quality of Life (HRQoL) instrument 15D and nine questions from the Speech, Spatial and Qualities of Hearing Scale (SSQ) having a direct implication to speech were selected for the analysis. After applying weights relevant to 15D, a delta of base-line (prior to HA fitting) and follow-up (two months after the initial fitting) responses to the selected questions were determined. A principal component analysis (PCA) was performed on the scaled and centered delta values. The resultant principal component scores were used to derive the composite index indicative of speech-related HA benefit.

INTRODUCTION

The speech intelligibility in challenging environments is, according to Kochkin (2002), the most desired improvement in hearing-aid (HA) rehabilitation. Studies also suggest that conversation in the presence of noise was the listening situation rated as the lowest in satisfaction by HA users (Abrams and Kihm, 2015). The subjective improvement in hearing ability in challenging listening situations involving speech

*Corresponding author: skn@es.aau.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

understanding reflected in the self-reported questionnaire responses can be used as an effective HA outcome measure, this is established in previous studies, including Cox (2009) and Lopez-Poveda (2017).

This study is part of an effort to identify potential sub-populations with low HA benefit in the population of 1,961 patients provided HA rehabilitation in two audiological departments at two university hospitals in Denmark, Wolff *et al.* (2017). The data analysed was collected as part of the Better hEAring Rehabilitation project (BEAR), a Danish national project envisioned to improve hearing rehabilitation through an evidence-based renewal of clinical practice. The purpose of the present study is to develop a speech-related HA benefit index to facilitate a single-dimension scale from the differently scaled self-reported questionnaires collected in the BEAR project.

METHOD

Only the questions with direct implication to speech understanding were considered for the analysis. Principal Component Analysis (PCA) was used to understand the underlying relationship between the selected questions. This approach is inspired by similar approaches in the field of socio-economic studies (Antony and Rao, 2007; Howe *et al.*, 2008; Chao and Wu, 2017). The resultant principal component (PC) scores and their contribution to explaining overall variance in the responses are used to calculate a composite index that can be an indicator of speech-related HA benefit.

Data

Data from the 1,961 patients registered in the centralized clinical database of the BEAR project are analyzed. The mean age of the patients was 67 years (ranging from 19 years to 100 years), with 72% first-time HA users and 28% experienced HA users. The records consisted of audiometric data (including air- and bone-conduction hearing thresholds, acoustic reflex, tympanometry, speech reception thresholds, and speech recognition scores), self-reported quality of life evaluation questionnaires (15D Sintonen and Pekurinen (1993) and a non-standardized health-related questionnaire, and Tinnitus Handicap Inventory (THI) Newman *et al.* (1996)), specific standard questionnaires to understand the hearing disabilities and HA outcome (Speech, Spatial and Qualities of Hearing Scale (SSQ) Gatehouse and Noble (2004), and International Outcome Inventory for HA (IOI-HA) Cox and Alexander (2002)), and HA data (HA type, fitting rationale, HA log time, and Real-Ear Measurements, REM).

Seventeen of the original SSQ questions were included, 12 of which are from the standard short form of SSQ49; the SSQ12, and five extra questions including one question from the speech domain, and four questions from the quality of hearing domain of the SSQ49. The SSQ questions were provided with a scale from 0 to 100 in the RedCap (Harris *et al.* (2019)) implementation, assuming that the responses divided by 10 would give the same scaling as in the standard SSQ from 0 to 10 (Lorentzen *et al.*, 2019). The responses were recorded online before the planned visit of the patients for the HA fitting, and before the scheduled follow-up visit (approximately

two months after the initial fitting). For the present analysis, question 3 of the 15D, and the nine questions from the customized implementation of the SSQ that all had direct implication to speech understanding were included (See Tab. 1).

Question ref.	Pragmatic sub-scale	Questions
15D3	Not applicable	Question 3 Hearing: 1. I can hear normally 2. I hear normal speech with little difficulty 3. I hear normal speech with considerable difficulty 4. I hear even loud voices poorly; I am almost deaf 5. I am completely deaf.
SSQ49-1.1	Speech in noise	You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the person you're talking to says?
SSQ49-1.4	Speech in noise	You are in a group of about five people in a busy restaurant. You can see everyone else in the group. Can you follow the conversation?
SSQ49-1.10	Multiple speech streams	You are listening to someone talking to you, while at the same time trying to follow the news on TV. Can you follow what both people are saying?
SSQ49-1.11	Multiple speech streams	You are in conversation with one person in a room where there are many other people talking. Can you follow what the person you are talking to is saying?
SSQ49-1.12	Multiple speech streams	You are with a group and the conversation switches from one person to another. Can you easily follow the conversation without missing the start of what each new speaker is saying?
SSQ49-1.14	Multiple speech streams	You are listening to someone on the telephone and someone next to you starts talking. Can you follow what's being said by both speakers?
SSQ49-3.14	Listening effort	Do you have to concentrate very much when listening to someone or something?
SSQ49-3.15	Listening effort	Do you have to put in a lot of effort to hear what is being said in conversation with others?
SSQ49-3.16	Not applicable	When you are the driver in a car can you easily hear what someone is saying who is sitting alongside you?
SSQ49-3.17	Not applicable	When you are a passenger can you easily hear what the driver is saying sitting alongside you?

Table 1: Questions included in the defined set of questions for analysis, including pragmatic subscales according to Gatehouse and Akeroyd (2006).

Analysis

A total of $n = 1{,}148$ out of 1,961 patients had given valid responses to all ten questions of interest for the present analysis consisting of both first time and experienced HA users. The standardized weights for the Danish version of 15D according to Wittrup-Jensen and Pedersen (2008) were applied, and the delta values (difference between baseline before fitting and before follow-up visit) calculated for the 15D and the subset of the SSQ questions. The delta values were scaled and centred before PCA.

The PCs with eigenvalues higher than 1 were considered for determining the composite index. The derived principal component score of the respective PC was weighted with a ratio of the percentage contribution of the PC to the overall percentage of variance explained by all the considered PCs. The summation of the weighted component scores of all the PCs determined the composite index, as shown in Eq. 1. This approach is an adaptation of method described in Antony and Rao (2007).

$$Index_i = \sum_{k=1}^{n} [PC_k / PC_{total}] * ComponentScore_{ik} \quad \text{(Eq. 1)}$$

where,

$$ComponentScore_{ik} = \sum_{j=1}^{n} Observation_{ji} * Loading_{kj} \quad \text{(Eq. 2)}$$

- i - number of patients
- k - number of principal components considered
- j - number of questions considered
- PC_k - percentage of variance explained by the k^{th} principal component
- PC_{total} - overall percentage of variance explained by all k principal components
- $Observation_{ji}$ - Scaled and centred recorded response values for j^{th} question for i^{th} individual.
- $Loadings_{kj}$ - Rotated principal axis values for k^{th} component for j^{th} question.

RESULTS

The scree plot in Fig. 1 shows the percentage of variance in the data explained by each PC. The first two components are considered for further analysis with respect to the eigenvalues of these components (overall: 60.8%, PC1: 49.3%, and PC2: 11.5%). Even though only two components were considered for deriving the index, this choice was statistically verified by performing an Analysis of Variance (ANOVA) on the index derived by including one to six PCs. Including more than two components did

not show any statistically significant change in the index. Tab. 2 shows the percentage contribution of each question towards the respective PC considered in the further analysis. It can be seen that PC1 is built by contributions from all SSQ questions included, whereas PC2 is clearly dominated by the two questions (SSQ-Q14 and SSQ-Q15) that relate to listening effort.

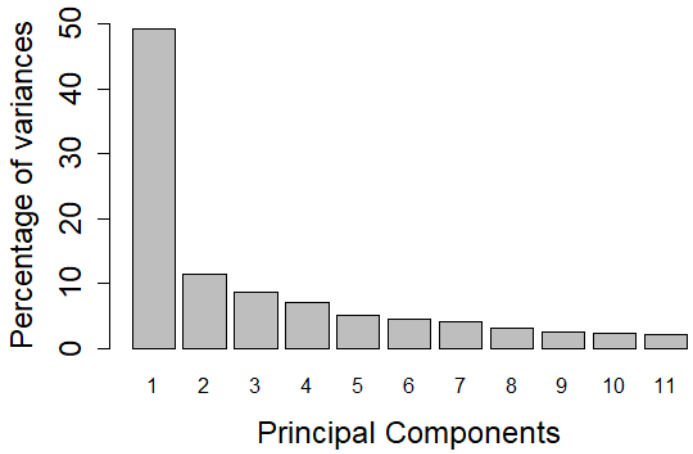


Fig. 1: Scree plot indicating percentage of variance explained by each principal component.

From the component score, the loading, and the percentage contribution of the respective PC, the speech-related benefit index is derived using Eq. 1, see Fig. 2. The negative index refers to the particular individual having negative correlation to the dominant PCs.

DISCUSSION

A composite index related to speech benefit using a HA is derived from selected self-reported questionnaire data. The PCA reveals a first component dominated by the questions in SSQ, with slightly higher loads on questions 4, 10, and 11. These represent three different pragmatic sub-scales; speech in noise, multiple speech streams, and speech in speech, respectively. The second component is loaded significantly by the listening effort dimension.

It is hypothesized that the composite index proposed is indicative for the speech-related HA benefit of a given individual. Negative values in Fig. 2 would then indicate a low benefit of the HA in the functional domain of speech understanding. This is valid as the negative index represents a negative correlation of the responses to the PCs considered. However, a margin could be applied to account for the false positives (misclassified low benefit users). Although the index is not normally distributed, as a rough estimator we can consider one standard deviation length as the error margin. Thus, negative indices below -1.9, which is one standard deviation length away from

Question	PC 1	PC 2
15D3	1.47	1.07
SSQ49-1.1	**10.09**	2.83
SSQ49-1.4	**11.64**	4.51
SSQ49-1.10	**11.43**	**9.81**
SSQ49-1.11	**10.77**	5.47
SSQ49-1.12	**9.48**	3.40
SSQ49-1.14	**9.57**	**10.71**
SSQ49-3.14	7.62	**22.05**
SSQ49-3.15	8.53	**21.96**
SSQ49-3.16	**9.91**	8.71
SSQ49-3.17	**9.46**	9.47

Table 2: Percentage of variance accounted for each question by each principal component.

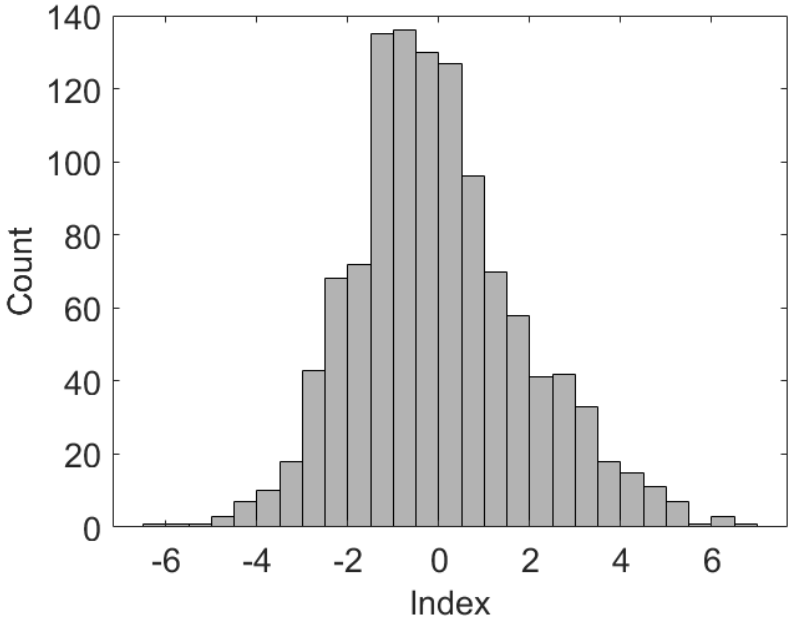


Fig. 2: Histogram showing the distribution of the derived speech-related benefit index, using Eq. 1.

the mean (which is -7e^{-12} and thus close to zero) could be an indicator of low benefit, and indices higher than $+1.9$ could represent the desired benefit, then 164 out of 1,148 patients in the present population have negative benefit (indices between -1.9, and -6). This accounts for a total of 14% of the patients included. Many studies have associated non-regular usage of the HA to lower benefit (Kochkin, 2007; Dillon *et al.*, 1999). The trend in the percentage of patients with non-regular HA usage found in

these studies is similar to the percentage of patients with low benefit in the current study. This suggests further investigation of the HA usage of the patients identified as low-benefit users. The absolute criteria of one standard deviation length as error margin also have to be statistically validated.

CONCLUSIONS

A composite index has been derived based on the available self-reported questionnaire data relating to speech understanding. Considering the population studied, an individual having a negative speech-related benefit index less than -1.9 could be a potential low-benefit HA user.

ACKNOWLEDGEMENT

This work was supported by Innovation Fund Denmark Grand Solutions 5164-00011B (BEAR project), Oticon, GN Hearing, Widex Sivantos Audiology, and other partners (Aalborg University, University of Southern Denmark, the Technical University of Denmark, Force Technology (Technical Audiological Laboratory, Odense), Aalborg, Odense and Copenhagen University Hospitals). The funding and collaboration of all partners is sincerely acknowledged. The authors sincerely thank Palle Rye and Michal Fereczkowski for their valuable inputs and critical comments.

REFERENCES

Abrams, H. and Kihm, J. (**2015**). "An Introduction to MarkeTrak IX: A New Baseline for the Hearing Aid Market," in Hearing Review, **22**(6), 16.

Chao, Y.-S. and Wu, C.-J. (**2017**). "Principal component-based weighted indices and a framework to evaluate indices: Results from the Medical Expenditure Panel Survey 1996 to 2011," in PLoS ONE **12**(9): e0183997.

Cox, R. M. (**2009**). "The International Outcome Inventory for Hearing Aids (IOI-HA): Psychometric properties of the english version," in Int. J. Audiol., **42**(sup1), 90–96.

Cox, R. M. and Alexander, G. C. (**2009**). "Assessment of subjective outcome of hearing aid fitting: getting the client's point of view," in Int. J. Audiol., **41**(1), 30–35.

Dillon, H., Birtles, G., and Lovegrove, R. (**1999**). "Measuring the Outcomes of a National Rehabilitation Program: Normative Data for the Client Oriented Scale of Improvement (COSI) and the Hearing Aid User's Questionnaire (HAUQ)," in J. Am. Acad. Audiol., **10**, 67-79.

Gatehouse, S. and Akeroyd, M. (**2006**). "Two-eared listening in dynamic situations," in Int. J. Audiol., **45** sup1, 120-124.

Gatehouse, S. and Noble, I. (**2004**). "The Speech, Spatial and Qualities of Hearing Scale (SSQ)," in Int. J. Audiol., **43**, 85–99.

Antony, G. M. and Rao, K. V. (**2007**). "To calculate the Human Development Index (HDI) and Human Poverty Index (HPI) of Indian states; to trace the indicators

useful for finding variations in poverty; and to develop a composite index that may explain variations in poverty, health, nutritional status and standard of living.," in Public Health, **121**(8), 578-87.

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., and Duda, S. N. (**2019**). "The redcap consortium: Building an international community of software platform partners," in J. Biomed. Inform., **95**, 103208.

Howe, L. D., Hargreaves, J. R., and Huttly, S. R. A. (**2008**). "Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries," in Emerg. Themes in Epidem., **5**:3.

Kochkin, S. (**2002**). "Marketrak VI: Consumers rate improvements sought in hearing instruments," in Hearing Review, **9**(11), 18–22.

Kochkin, S. (**2007**). "MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids," in The Hearing Journal, **60**(4), 24–51.

Lopez-Poveda, E. A., Johannesen, P. T., Pérez-González, P., Blanco, J. L., Kalluri, S., and Edwards, B. (**2017**)."Predictors of Hearing-Aid Outcomes," in Trends Hear, **21**. doi: 10.1177/2331216517730526

Lorentzen, L. N., Narne, V. K., Wolff, A., and Schmidt, J. H. (**2019**). "Validation of the Danish version of the SSQ12 and the preliminary results on correlation between Quality of Hearing and Quality of Life of people with hearing impairment," Poster presented at the 7^{th} International Symposium on Auditory and Audiological Research, August 21 - 23 2019, Nyborg, Denmark.

Newman, C. W., Jacobson, G. P., and Spitzer, J. B(**1996**). "Development of the tinnitus handicap inventory," in Arch. Otolaryngol. Head Neck Surg., **122**(2), 143-148.

Sintonen, H. and Pekurinen, M. (**1993**). "A fifteen-dimensional measure of health-related quality of life (15D) and its applications," in *Quality of Life Assessment: Key Issues in the 1990s*, 185-195. Springer, Dordrecht.

Wittrup-Jensen, K. and Pedersen, K. (**2008**). "Modelling Danish Weights for the 15D Quality of Life Questionnaire by Applying Multi-Attribute Utility Theory (MAUT)," in Health Economics Papers, 7.

Wolff, A., Houmøller, S. S., Gaihede, M., Hougaard, D. D., and Hammershøi, D. (**2017**). "National Better hEAring Rehabilitation (BEAR) project: A status on the database with special focus on patients' motivation on hearing aid treatment," Poster presented at the 6^{th} International Symposium on Auditory and Audiological Research, August 23 - 25 2019, Nyborg, Denmark.

Applicability and outcomes of a test for binaural phase sensitivity in elderly listeners

INGA HOLUBE[1,3], THERESA NUESSE[1,3], OLAF STRELCYK[2], ANNAEUS WILTFANG[1], PETRA VON GABLENZ[1,3] AND ANNE SCHLUETER[1,3]

[1] *Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, D-26121 Oldenburg, Germany*

[2] *Sonova U.S. Corporate Services, Cincinnati, OH, USA*

[3] *Cluster of Excellence "Hearing4All", Oldenburg, Germany*

Interaural phase difference (IPD) discrimination in the binaural auditory system has been shown to be related to localization abilities and speech intelligibility in background noise. One of the tests for binaural phase sensitivity determines the highest frequency of the test tone for which an interaural phase difference of 180° is detectable (IPD-FR). This test was included in a test battery together with examination of visual and hearing abilities, balance, tactile- and motor-skills, and cognitive abilities. The IPD-FR test was conducted with an adaptive 3-AFC experiment starting with a test-tone frequency of 250 Hz. Sixty-five of 220 participants could not perform the IPD-FR task. The main predictors for inability to perform the IPD-FR task are hearing loss at 250 Hz, fluid intelligence, measurement number, and gender. A linear regression analysis revealed that the test result IPD-FR threshold is related to pure-tone thresholds at low frequencies, composite score of cognition, and composite score of tactile sensitivity, fine motor skills, and vision, as well as gender. A correlation analysis shows that the IPD-FR threshold is not related to speech recognition in non-dynamic listening conditions with speech from the front if low-frequency hearing loss is taken into account.

INTRODUCTION

Speech recognition performance appears to be influenced by multiple factors such as hearing loss, age, cognitive abilities, and supra-threshold auditory processing. One of the factors for supra-threshold auditory processing is the sensitivity to interaural phase differences (IPD) in the binaural auditory system. IPD has been shown to be related to localization abilities and speech intelligibility in background noise (see e.g., Strelcyk and Dau, 2009). Several test paradigms are used in audiological research to measure IPD. One of the tests determines the highest frequency for which an IPD of 180° is detectable (IPD-FR threshold, test names "IPD-FR" in Neher *et al.*, 2011, and "TFS-AF" in Füllgrabe *et al.*, 2017). The performance of participants in these kinds of tests can be affected by personal characteristics such as age and hearing loss (see Füllgrabe and Moore, 2018 for a meta-analysis). Furthermore, Füllgrabe and Moore

*Corresponding author: inga.holube@jade-hs.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

(2018) assumed that cognitive abilities might have an influence on test performance as a substantial amount of variance could not be explained by other factors.

The IPD-FR test was regarded as a test for supra-threshold auditory processing ability. It was included in an extensive test battery containing audiological standard diagnostics, questionnaires, motor and vision skills as well as cognitive tasks to analyze their relation to speech recognition. As a relatively large number of elderly listeners was unable to perform the IPD-FR test, we investigated: (i) Prediction of the individual inability to perform the IPD-FR task, (ii) Explanation of variance in the measured IPD thresholds.

METHODS

Participants

Two hundred and twenty-three volunteers (77 from Hörzentrum Oldenburg GmbH database and 146 from public announcement) aged from 55 to 81 years participated in an extensive test battery (see Table 1). Three participants were excluded due to single sided deafness or technical issues. All participants were numbered based on the order of their participation date (measurement number).

	Age cohorts (years)					
	55-60	**61-65**	**66-70**	**71-75**	**76-81**	**all**
Number of participants	51	43	46	42	38	220
Number female/male	33/18	21/22	25/21	22/20	20/18	121/99
Median PTA-4 (dB HL)	12.5	20.6	27.5	35.6	34.4	23.8
Median PTA-low (dB HL)	9.4	15.6	15.3	20.6	20.6	14.4
IPD-FR ability/inability	48/3	31/12	31/15	23/19	22/16	155/65
Mean IPD-FR threshold (Hz)	757.4	678.0	657.5	574.0	549.8	664.9

Table 1: Characteristics of the participants. PTA-4 denotes the average hearing loss at 0.5, 1, 2, and 4 kHz. PTA-low denotes the average hearing loss at 0.25, 0.5, 0.75, and 1 kHz averaged between both ears.

IPD-FR

An adaptive 3-AFC experiment (1-up-2-down rule) was used to determine the highest frequency for which an IPD of 180° was detectable (Neher *et al.*, 2011). The outcome of the experiment was the IPD-FR threshold in Hz. Within each interval, a sinusoid with a duration of 2 s and an amplitude modulation of 1 Hz was presented over headphones. The two reference intervals included diotic stimuli whereas the interaural

phase changed between 0 and 180° every 0.5 s in the target interval. The presentation level of all stimuli was 30 dB SL. The experiment started with a frequency of 250 Hz and a step size of 250 Hz. The step size was halved after each upper of a total of eight reversals. A minimum bound of 125 Hz and a maximum of 20 kHz was chosen for the IPD-FR threshold. Oral instructions were given to the participants and a training with 10 trials at minimum was carried out. All participants except the first 29 administered the task using a touch screen.

Test battery

Pure-tone hearing thresholds were measured via air conduction (0.125 to 8 kHz) and via bone conduction (0.5 to 6 kHz).

Several tests were used to determine cognitive skills, visual abilities, fine motor skills, and tactile sensitivity. To ensure audibility of instructions during those tests, participants wore hearing aids if German indication criteria (G-BA, 2017) were fulfilled. Hearing aids (Phonak Bolero V90-P or V90-SP depending on the severity of hearing loss) were fitted to those participants who did not own hearing aids. Hearing aid owners could decide, based on their subjective preference, whether they wanted to use their own or the newly fitted hearing aids.

1) Composite scores of cognitive tests:

The test outcomes of the cognitive tests were z-transformed and averaged to obtain three composite scores related to a typical categorization of cognitive abilities:

- Fluid intelligence: Ruff2 & 7 (Ruff and Allen, 1996), Trail-Making Test A and B (Reitan, 1992), STROOP (Puhr and Wagner, 2012), TAP divided attention test (Zimmermann and Fimm, 2013), digit span forward and backward (Petermann, 2012)
- Crystallized intelligence: Regensburg word fluency test (Aschenbrenner *et al.*, 2000), multiple choice vocabulary test (Lehrl, 2005)
- Verbal memory: Verbal learning and memory test (Helmstaedter *et al.*, 2001)

2) Vision, fine motor skills and tactile sensitivity

Three outcomes for visual abilities, fine motor skills, and tactile sensitivity were used as single values. These skills might be necessary to handle the touch screen in the IPD-FR task although the keys on the screen were large and easy to catch.

- Tactile sensitivity was measured as the 75%-threshold of spatial resolution for fingertips using JVP domes (Johnson *et al.*, 1997).
- Fine motor skills: The MLS test battery of *Schuhfried* (Neuwirth and Benesch, 2012) was applied. The outcomes were standardized and combined to four factors (factors 1-3, and 5 in Neuwirth and Benesch, 2012). For further analysis, only factor 5 characterizing the movement speed of arm, hand, and fingers was used.

- Visual acuity was measured bespectacled if glasses were prescribed and available using *Optovist* (VISTEC Vision Technologies, 2010).

3) Speech recognition

Speech levels for recognition scores of 50% (SRT) for speech in quiet were determined with the German Freiburg digit test and the Göttingen sentence test (GÖSA; Kollmeier and Wesselkamp, 1997). The SRT for speech in background noise was measured for GÖSA in the standardized stationary Gönoise using headphones and via loudspeaker in the TASCAR system (Grimm *et al.*, 2015). In two other TASCAR conditions, IFFM (Holube *et al.*, 2010) and a cafeteria recording were used as noise sources. The target sentences were always presented from the front. Gönoise and IFFM were also presented from the front and the cafeteria recording from all eight loudspeakers of the system. In all speech tests, the participants wore the same hearing aids (Phonak Bolero V90-P/SP, see above) fitted according to NAL-NL2 fitting procedure (Keidser *et al.*, 2011) if they fulfilled the German indication criteria. In the speech tests, in contradiction to the other tests described in (1) and (2), the participants were not allowed to use their own hearing aids if available to ensure similar amplification schemes in the speech tests.

RESULTS

Predicting individual inability to perform the IPD-FR task

Sixty-five participants (approx. 30%) could not detect the IPD of 180° at any frequency down to 125 Hz. The age distribution of the inability groups relative to the ability group is additionally given in Table 1. Logistic regressions were used to find predictors of inability to perform the IPD-FR task. In the following model calculations, 5 participants were excluded due to outliers in the cognitive variables.

In the first step of the model development process, air conduction pure-tone threshold at 250 Hz was identified as having the highest significance of all audiogram parameters. This variable was kept in the model and further single variables were added (but not kept). Significant contributions were found for fluid intelligence ($p < 0.001$), measurement number ($p = 0.004$), age ($p = 0.010$), touch screen usage ($p < 0.012$), and crystallized intelligence ($p = 0.037$). There was no gender effect ($p = 0.057$). Verbal memory and none of the vision acuity, tactile, or motor variables were significant.

In a next step, the significant variables and female gender were added to the logistic model starting with fluid intelligence and kept if they had a significant contribution to the model. Otherwise, they were omitted. In addition, a combination (mean) of fluid and crystallized intelligence was tested, but resulted in a lower R^2 compared to fluid intelligence alone. The procedure resulted in a model with four independent variables shown in Table 2 ($R^2_{(\text{Nagelkerkes})} = 0.576$, sensitivity = 90.7%, specificity = 65.6%).

The odds (likelihood) for inability increases by a factor of 1.78 per 5 dB increase in hearing loss in the worse ear at 250 Hz. Besides this, fluid intelligence, but not

crystallized intelligence or verbal memory, has a significant contribution to the model. The order of the participants, coded in the measurement number, contributes significantly to inability. The odds for inability increases e.g., by a factor of 1.1 per decreasing of measurement number by 10. The measurement number includes: A) Non-usage of a touch screen for the first 29 participants. B) Recruitment of the first 77 participants from the database of Hörzentrum Oldenburg GmbH. On average, those participants had a higher hearing loss and included more hearing aid users compared to those from public announcement. Therefore, this variable might include hearing loss differences not covered by the threshold at 250 Hz. C) Possible training effects including optimization of the oral instructions of the examiner. D) Other unknown temporal or recruiting effects. In addition, gender has a small but significant effect: The odds for inability is 2.8 times higher in female than in male.

Independent variables	Wald	*p*	Odds ratio (95% C.I.)
AC WEHL 250 Hz	38.7	< 0.001	1.12 (1.08-1.16)
Fluid intelligence	12.0	0.001	4.07 (1.84-8.99)
Measurement number	7.4	0.007	1.01 (1.00-1.02)
Female gender	5.4	0.020	2.79 (1.18-6.60)

Table 2: Logistic regression model for inability of the IPD-FR task. AC WEHL 250 Hz: air conduction thresholds in the worse ear at 250 Hz.

Predicting IPD-FR thresholds

A linear regression analysis was calculated for the 145 participants who could perform the IPD-FR task and for whom complete composite scores of cognitive skills and tactile/motor/vision were available. Dropouts were equally distributed over age groups.

In a first step, PTA-low was identified as most significant variable of all audiogram variables in the linear regression model. Thereafter, each variable was included in the linear regression model (but not kept). Measurement number, touch screen usage, and verbal memory did not contribute significantly to the model. The other variables were added stepwise to the linear regression model and kept if they had a significant contribution. To reduce the number of variables, several of them were combined: A significant correlation between the cognitive composite z-scores for fluid and crystallized intelligence of 0.499 ($p < 0.001$) was observed. Therefore, the mean of both cognitive z-scores was used as a combined variable. In addition, fine motor skills, tactile sensibility, and vision acuity were significantly correlated (motor-tactile: 0.210, $p = 0.011$; motor-vision: 0.181, $p = 0.029$; tactile-vision: 0.244, $p = 0.003$).

Hence, the z-scores of all three variables were averaged forming a new variable tactile/motor/vision. The resulting linear regression model with four independent variables is shown in Table 3.

Independent variables	R	R^2 corr.	R^2 change	Coeff B
PTA-low	0.371	0.132	0.132	-7.0
Mean of fluid and crystalline intelligence z-score	0.468	0.208	0.076	101.9
Tactile/Motor/Vision z-score	0.532	0.268	0.060	95.0
Male gender	0.584	0.322	0.054	113.4

Table 3: Results of the linear regression model with four independent variables.

Table 3 reveals that the IPD-FR threshold decreased by about 70 Hz for an increase in mean low-frequency pure-tone threshold by 10 dB. The IPD-FR threshold increased by about 100 Hz for an increase in cognitive and tactile/motor/vision skills by one standard deviation. Although, it has to be pointed out that the two independent variables are significantly correlated ($r = 0.359$, $p < 0.001$). In addition, males had on average about 110 Hz higher IPD-FR thresholds than females. It should be noted that similar to the logistic regression, age had no additional significant contribution to the model when the cognitive abilities were taken into account.

RELATION OF IPD-FR THRESHOLD TO SPEECH RECOGNITION

IPD-FR threshold was significantly correlated with the SRT for all speech test conditions measured with hearing aids. Spearman's correlation coefficients are 0.240 to 0.362 ($p \leq 0.004$). The highest correlation coefficient was observed for the fluctuating masker IFFM. When controlling for low-frequency pure-tone thresholds, the absolute values of partial correlations between IPD-FR threshold and SRTs dropped to the range from 0.015 to 0.146 and were no longer significant ($0.082 \leq p \leq 0.851$).

DISCUSSION

The first objective of this study was to predict the individual inability to perform the IPD-FR task using logistic regression. Personal characteristics that predicted the inability were hearing loss in the worse ear at 250 Hz, fluid intelligence, measurement number, and gender. Especially the effect of the factor measurement number left some open questions. Generally, it was surprising that the IPD-FR task was an insuperable obstacle for many participants. In contradiction to this observation, Füllgrabe *et al.*

(2018) stated that "reliable threshold estimates can be obtained relatively quickly […] and without practice" for (nearly) all listeners. 14 out of the 65 participants who could not perform the task met the inclusion criteria of Füllgrabe *et al.* (2018), i.e. air-conduction thresholds up to 1.5 kHz ≤ 25 dB HL. For future studies, more training (e.g., with ILD cues) might improve the percentage of participants who can perform the test and increase specificity. Furthermore, the IPD-FR minimum of 125 Hz might be too high (compared to 30 Hz in TFS-AF) and the start frequency should be lowered if participants were not able to use the cue at 250 Hz.

In a second step, the IPD-FR thresholds of those participants who were able to perform the task were analyzed with the aim to explain the variance. As presumed by Füllgrabe and Moore (2018) in their meta-analysis and consistent with Strelcyk *et al.* (2019), cognitive abilities were significantly predictive. An additional influence of the parameter "age" on the thresholds was not found in the present data. However, other sensory and motor skills also had predictive power.

With regard to the lack of correlation between IPD-FR threshold and speech recognition when controlling for low-frequency hearing loss, the applied speech tasks should be reconsidered. The relationship might be observable only for more spatial and dynamic speech conditions (e.g., Neher *et al.*, 2011).

ACKNOWLEDGEMENT

Thanks to Olga Schwarz and Anika Morgenstern for conducting part of the measurements, as well as Ulrike Lemke, Sigrid Scherpiet, and Emily Urry for project support. Research funded by the governmental funding initiative "Niedersächsisches Vorab" of the Lower Saxony Minstry for Science and Culture in cooperation with Sonova AG (Stäfa, Switzerland).

REFERENCES

Aschenbrenner, S., Tucha, O., and Lange, K.W. (**2000**). "RWT: Regensburger Wortflüssigkeits-Test," Göttingen Hogrefe.

Füllgrabe, C., Harland, A.J., Sek, A.P., and Moore, B.C.J. (**2017**). "Development of a method for determining sensitivity to temporal fine structure," Int. J. Audiol., **56**, 926-935. doi: 10.1080/14992027.2017.1366078.

Füllgrabe, C., and Moore, B.C.J. (**2018**). "The association between the processing of binaural temporal-fine-structure information and audiometric threshold and age: A meta-analysis," Trends Hear., **22**. doi: 10.1177/2331216518797259.

Füllgrabe, C., Sek, A.P., and Moore, B.C.J. (**2018**). "Senescent changes in sensitivity to binaural temporal fine structure," Trends Hear., **22**, 1-16. doi: 10.1177/2331216518788224.

Grimm, G., Luberadzka, J., Herzke, T., and Hohmann, V. (**2015**). "Toolbox for acoustic scene creation and rendering (TASCAR): Render methods and research applications," In: Proceedings of the Linux Audio Conference, Johannes Gutenberg University (JGU). Mainz, Germany.

Helmstaedter, C., Lendt, M., and Lux, S. (**2001**). "VLMT: Verbaler Lern- und Merkfähigkeitstest," Göttingen, Hogrefe.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). „Development and analysis of an International Speech Test Signal (ISTS)," Int. J. Audiol., **49**(12), 891–903. doi: 10.3109/14992027.2010.506889.

Johnson, K.O., van Boven, R.W., and Phillips, J.R. (**1997**). "J.V.P. Domes for cutaneous spatial resolution measurement. Operation Manual," Stoelting Co., Wood Dale, Illinois (USA).

Keidser, G., Dillon, H., Flax, M., Ching, T.Y.C., and Brewer, S. (**2011**). "The NAL-NL2 prescription procedure," Audiol. Res., **1**(e24), 88-90. doi: 10.4081/audiores.2011.e24.

Kollmeier, B., and Wesselkamp, M. (**1997**). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," J. Acoust. Soc. Am., **102** (4), 2412–2421. doi: 10.1121/1.419624.

Lehrl, S. (**2005**). "Mehrfachwahl-Wortschatz-Intelligenztest: Manual mit Block MWT-B," 5th edition, Balingen Spitta Verlag.

Neher, T., Laugesen, S., Søgaard Jensen, N., and Kragelund, L. (**2011**). "Can basic auditory and cognitive measures predict hearing-impaired listeners' localization and spatial speech recognition abilities?" J. Acoust. Soc. Am., **130**, 1542-1558. doi: 10.1121/1.3608122.

Neuwirth, W., and Benesch, M. (**2012**). "Manual Motorische Leistungsserie. Kurzbezeichnung MLS," Mödling: Schuhfried.

VISTEC Vision Technologies (**2010**). "Optovist. Gebrauchsanweisung ab Softwareversion 1.0.007," VISTEC AG - Olching (Germany).

Petermann, F. (Hg.) (**2012**). "Wechsler-Adult-Intelligence-Scale-Fourth Edition (WAIS-IV). Manual," Frankfurt/Main Pearson Assessment and Information GmbH.

Puhr, U., and Wagner, M. (**2012**). "Handanweisung Interferenztest nach Stroop. Kurzbezeichnung STROOP," 25th edition, Mödling SCHUHFRIED GmbH.

Reitan, R. M. (**1992**). "Trail Making Test: Manual for administration and scoring," Tucson Reitan Neuropsychology Laboratory.

Ruff, RM., and Allen, C.C. (**1996**). "Ruff 2 & 7 Selective Attention Test: Professional Manual," Lutz PAR. Retrieved from www.parinc.com.

Strelcyk, O., and Dau, T. (**2009**). "Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing," J. Acoust. Soc. Am., **125**, 3328-3345. doi: 10.1121/1.3097469.

Strelcyk, O., Zahorik, P., Shehorn, J., Patro, C., and Derleth, R.P. (**2019**). "Sensitivity to interaural phase in older hearing-impaired listeners correlated with nonauditory trail making scores and with a spatial auditory task of unrelated peripheral origin," Trends Hear., **23**, 1-21. doi: 10.1177/2331216519864499.

Zimmermann, P., and Fimm, B. (**2013**). " Testbatterie zur Aufmerksamkeitsprüfung," Version 2.3. Teil 1. 2nd edition, Herzogenrath Psytest.

Task repetition influence on pupil response during encoding of auditory information in normal-hearing adults

MISEUNG KOO[1], MYUNG-WHAN SUH[1,2], JUN HO LEE[1,2], SEUNG-HA OH[1,2], AND MOO KYUN PARK[1,2*]

[1] *Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, Seoul, Korea*

[2] *Sensory Organ Research Institute, Seoul National University Medical Research Center, Seoul, Korea*

Although numerous behavioural measures to estimate listening effort have been developed in recent years using free recall or dual-task paradigms, relatively little is known about physiological measures, such as pupil dilation, in response to cognitively demanding tasks. This study used a repeated-measure experimental design and aimed to investigate the cognitive resource allocation process of spoken words in an immediate free recall paradigm. Here, ten adults with normal hearing (NH) attended 2 days of trials with 14 trials per day. The listeners heard four-speaker babble noise along with seven sentences and then tried to remember the first words of all seven sentences. Recall performance on the first day only showed a significant serial position effect ($p < 0.05$). With increasing memory load imposed by the subsequent recall task, baseline pupil size significantly enlarged ($p < 0.01$), and the PPDs significantly decreased ($p < 0.01$) during the encoding process, implying that a gradual increase in resources allocated to memory capacity corresponded to a decline in resources allocated to listening. Real-time allocation of cognitive resources during the encoding of spoken words can be monitored independently by the analysis of pupil dilation averaged over multiple trials.

INTRODUCTION

Hearing-impaired (HI) listeners may have to devote more effort to perceiving speech under adverse listening conditions (Kahneman, 1973). Interestingly, effortful listening in everyday conversation where the speech is fully audible and intelligible has been reported (Lunner *et al.*, 2016; Pichora-Fuller *et al.*, 2016), implying that individuals devote different amounts of effort to facilitate understanding even though behavioural performances may not differ. Listening effort, defined in the framework for understanding effortful listening (FUEL) as "deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task" (Pichora-Fuller *et al.*, 2016), has become a subject of increasing interest in cognitive hearing science (Rudner *et al.*, 2014). Numerous assessment tools have been developed to assess individual cognitive spare capacity in relation to varying effort in listening, including the sentence-final word identification and recall test (SWIR) (Ng *et al.*, 2013; Ng *et al.*, 2015) that inspired our research. However, behavioural measures

*Corresponding author: aseptic@snu.ac.kr

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

alone might not fully describe the quantitative change in listening effort during encoding of auditory stimuli.

Pupillometry has recently gained considerable attention as the most promising physiological measure underlying cognitive processing in response to hearing-related tasks (Pichora-Fuller *et al.*, 2016; Zekveld *et al.*, 2018). Studies on speech recognition in noise frequently utilise peak pupil dilation (PPD) to quantify changes in listening effort (Ohlenforst *et al.*, 2017). In addition, a greater pre-stimulus baseline, also known as tonic or baseline pupil size (BS), was observed in participants with higher working memory capacity (WMC) (Heitz *et al.*, 2008; Tsukahara *et al.*, 2016). We hypothesised that BS would vary not only from subject to subject, reflecting variations in individual WMC, but also between sentences in a trial, with changes in memory demand imposed by the recall task. This study had the following aims:

1. To investigate whether a repeated-measure experimental design would have a favourable impact on either behavioural assessment or pupillometric data in response to a cognitive task;
2. To monitor task-evoked changes in pupil size in response to the recall task to determine whether real-time cognitive resource allocation can be detected based on pupil size during the encoding of auditory stimuli.

MATERIALS AND METHODS

Twelve fluent Korean speakers (mean age 24.6 years, range 22–29 years, eight males) with NH were initially recruited. All participants reported normal or corrected-to-normal vision as well as bilateral NH and attended two visits (days 1 and 2) with at least a 3-week interval to avoid any learning effect (Ohlenforst *et al.*, 2018; Simonsen *et al.*, 2016). They were told to refrain from caffeine consumption for at least 6 h before each visit. After their first visit, two of the participants were excluded from data collection due to large amounts of missing data.

Stimuli

Fourteen seven-sentence lists from the Korean version of the hearing in noise test (HINT) (Moon *et al.*, 2005) were selected for the SWIR in accordance with the published study protocol (Lunner *et al.*, 2016; Ng *et al.*, 2013; Ng *et al.*, 2015). Target speech, spoken by a male speaker, was presented at 65 dB SPL along with four-talker babble noise (two males and two females) starting 2 s before sentence onset and ending 2 s after sentence offset (Fig. 1). To evaluate serial-position effects, sentences in each set of seven sentences were allocated as follows: the first and second sentences to the primacy, third to fifth sentences to the asymptote, and sixth to seventh sentences to the recency position.

Procedure

During the first visit, the HINT speech reception thresholds (SRTs), speech and noise from 0° and at 80% correct performance, were obtained from individuals using a published HINT procedure (Hallgren *et al.*, 2006). Subsequently, during SWIR

training with four practices, the $SRT_{80\%}$ was tuned to reach the signal-to-noise ratio (SNR) 95% correct performance depending on repetition performance of the first words of each sentence (identification task), as described in Ng et al. (2015), although listeners were instructed to complete both tasks: identification and free recall. The recall phase began with the presentation of a 0.5 s beep sound, and participants were prompted to report the first words in any order as many as possible (recall task). In the following blocks comprising 10 trials of a recall task in the SWIR and pupil diameter recording, participants were not required to give any verbal response before the beep in order to prevent rehearsal of to-be-remembered items that might potentially influence the subsequent recall performance. At the second visit, participants repeated the SWIR training and SWIR with pupil data recording while maintaining the SNRs obtained from the first visit (Fig. 1).

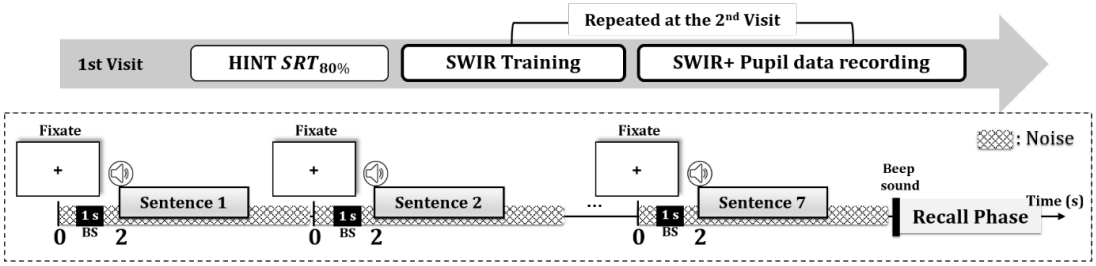


Fig. 1: Upper panel: Three experimental sessions used in this study. Lower panel: Encoding and recall phases of a trial in SWIR. The inter-stimulus interval between sentence offset and onset of the following sentence was longer than 4 s. "0" represents the start of four-talker babble noise used to calculate the peak pupil latency.

Pupil diameter was recorded using a wearable eye-tracking headset (Pupil Labs, Germany) with 200 Hz binocular eye cameras positioned in front of the eyes and using Matlab software (Release 2018a) provided by Oticon Medical A/S, Smørum, Denmark. The raw pupil diameter data were pre-processed to remove samples with blink artifacts or dilation speed outliers using median absolute deviation method, as described in Kret *et al.* (2019). Pupil diameter values greater or lower than the median ± 2.5 times the standard deviation of the remaining data were defined as blink. We also applied divisive baseline correction (proportion change: corrected pupil size = pupil size/baseline) to the pre-processed data at the end of normalization. Room illumination was provided by two LED lights positioned on the ceiling of the testing booth and varied depending on the dynamic range of the participant's pupil size to prevent floor and ceiling effects, and was individually adjusted to the pupil-size midpoint, from dim (~30 lux) to bright (~230 lux) prior to data collection, with an average illuminance of 110 lux.

Statistical analyses

A nonparametric repeated-measures analysis of variance (RM ANOVA) was used to analyse recall performance with two within-subjects factors: word position (primacy, asymptote, recency) and visit (day 1 and 2) because the recall score was a discrete variable rather than a continuous one that follows a normal distribution.

Nonparametric analysis of repeated data was performed with R (nparLD package) (Brunner *et al.*, 2002; Noguchi *et al.*, 2012).

Statistical analyses of the pupil data collected from the encoding phase were performed using SPSS software, version 25 (Chicago, IL, USA). A linear mixed model (LMM) was employed to examine the data because of its ability to handle missing values due to the large number of blinks and to statistically compare the fixed effects of stimulus presentation order and visit on BS, PPD, and peak pupil latency (time interval between sentence onset and PPD). The average pupil size during the 1 s pre-stimulus period served as BS. Post hoc analyses with Bonferroni correction were used to adjust for multiple comparisons. Eye was not included as a fixed effect because, in our preliminary experiments, no significant influence of eye on pupil response was observed. A p-value <0.05 was considered significant. After data selection, we identified 298 invalid pupil traces out of 2,800 pupil traces, recorded per sentence, due to missing PPD values for containing either more than 25% blinks or erroneous recording. We measured 35.7 pupil recordings per participant on average, regardless of eye position.

RESULTS

Recall performance

As depicted in Fig. 2, the nonparametric RM ANOVA results revealed a significant interaction between word position and visit day, indicating that a significant serial position effect was found in recall performance on the first day only ($p = 0.0274$). Post hoc analyses with Bonferroni correction showed significantly better performance for early (primacy) than for late (recency) items in the list ($p < 0.0167$). This pattern, however, was not seen in recall performance on the second day. No other significant main effects or interactions were observed.

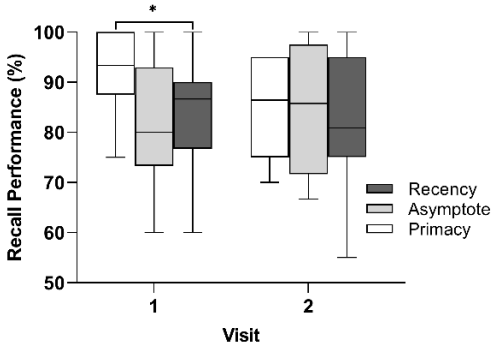


Fig. 2: Results of participants' recall scores as a function of word position (primacy, asymptote, and recency) and visit (day 1 and 2).

BS, PPD, and peak pupil latency during memory encoding of spoken words

The LMM results for BS revealed significant main effects of visit ($F = 67.62$, $p < 0.001$) and stimulus presentation order ($F = 41.63$, $p < 0.001$), in addition to a

significant interaction between these variables ($F = 3.48$, $p = 0.002$). Compared with BS for the first item, listeners' pupil size increased progressively with increasing memory load in preparation for the subsequent recall task (Fig. 3a). The incremental increase in recall performance was greater on the second day than that on the first day (Fig. 3b).

The LMM results for the PPD revealed significant main effects of stimulus presentation order ($F = 13.10$, $p < 0.001$) and visit day ($F = 8.67$, $p = 0.003$) in addition to a significant interaction between these variables ($F = 2.52$, $p = 0.02$). There was a progressive decline in PPD as the number of words to be remembered increased (Fig. 3c), and the PPD on the second day was significantly greater than that on the first day (Fig. 3d).

The LMM for peak pupil latency revealed significant main effects of visit day ($F = 46.08$, $p < 0.01$) and stimulus presentation order ($F = 4.58$, $p < 0.01$); however, no significant interaction was observed. Post hoc analyses using Bonferroni correction showed that the latency was significantly shorter for the first items than for the middle items, i.e., the third ($p = 0.027$), fourth ($p < 0.001$), and sixth sentences ($p = 0.004$) (Fig. 3e). The latency on day 1 was significantly shorter than that on day 2 (Fig. 3f).

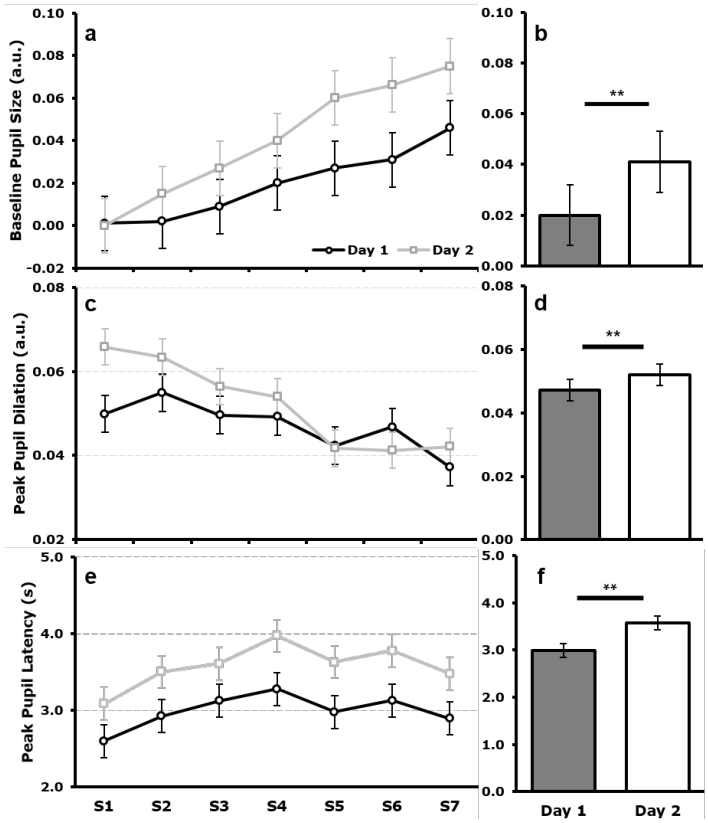


Fig. 3: Mean ± 1 SE of BS (a, b), PPD (c, d), and peak pupil latency (e, f) during memory encoding as a function of stimulus presentation order (S1 to S7) and visit.

DISCUSSION

Our findings agree with earlier literature showing that single-trial analysis is insufficient to provide a reliable interpretation of pupillometry data (Winn *et al.*, 2018). This experiment, which attempted to quantify real-time allocation of cognitive resources during encoding of auditory inputs, also expanded the findings of previous studies (Lunner *et al.*, 2016; Ng *et al.*, 2013) by adding time-locked eye tracking analysis of the memory-encoding period followed by the free recall task.

Our most important finding was that the BS tended to increase; while the PPD tended to decrease during the phase when participants were encoding words heard against a competing speech background (see Fig. 3). These incremental changes are closely associated with increasing memory demands as listeners strive to memorise the items for a subsequent recall task. BS, recorded sentence-by-sentence during the pre-stimulus period, seemed to be an effective indicator of the changes in memory load and the ability to store a sequence of auditory items, in line with previous findings (Tsukahara *et al.*, 2016). There are a number of different perspectives on the role of BS. So far most studies have used a trial-based analysis, including Gilzenrat *et al.* (2010) who linked increases in BS to reduced task utility and disengagement from a given task with the purpose to demonstrate an inversed relationship between BS and task-evoked pupil dilations, predicted by the adaptive-gain theory (Aston-Jones *et al.*, 2005). Since the listening condition used in this study was relatively easy for NH listeners, it is predetermined that the recall task might be more and more rewarding for them to explore, relative to the identification (control) task, thereby producing BS that progressively enlarged over the course of a trial.

The longest peak latency for the middle items seemed to associate item difficulty or high processing load on items in the asymptote position (primacy effect). Similarly, Koelewijn *et al.* (2014) found longer latency to PPD in the dual-sentence condition compared to the single-sentence condition. However, recall performance in asymptote varied between the visits, and indeed it was the highest in the second visit. This consistent pattern of peak latency needs to be incorporated with recall performance, which was not examined in our study, in further study. In addition, researchers when designing multiple conditions and experiments repeated on different days may need to consider the effect of test-day on pupil response because a statistically significant increase in BS, PPD, and peak latency was found in our study. We would recommend a minimum of 10 repetitions to be carried out for each experimental condition excluding the first few trials (i.e. familiarization training).

One weakness of our study is that the use of contact lenses was not restricted due to difficulties in recruiting suitable subjects who had both NH and normal, uncorrected vision. Rather some participants were allowed to use them to avoid any unnecessary fatigue in maintaining eye fixation (not the focus of our study). To increase data accuracy and reduce variability, our findings should be confirmed in further studies using a large numbers of subjects and examining between-subject factors such as age, hearing status (e.g. HI listeners), and cognitive ability because listening effort studies may help future hearing rehabilitation practices and approaches (Richmond et al.,

2011). Since task-evoked pupil dilation is apparently much smaller than other pupillary reflexes, we focused on NH young adults who reportedly exhibit greater changes in pupil response than older age and HI groups (Winn *et al.*, 2018). Moreover, an adequate normalisation or analysis method (e.g. growth curve analysis) should be employed to minimise possible individual variance.

In conclusion, pupillometry can be an independent indicator for monitoring online resource allocation in a free recall paradigm where a repeated-measure design is adopted. Although we could not explore inter-individual variance in cognitive processing using the analysis of pupil dilation in this study, pupillometry was able to detect the ongoing changes during the memory-encoding phase while behavioural assessments, measured offline, could not provide such information.

REFERENCES

Aston-Jones, G., and Cohen, J. D. (**2005**). "An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance," Annu. Rev. Neurosci., **28**, 403-450. doi:10.1146/annurev.neuro.28.061604.135709.

Brunner, E., Domhof, S., and Langer, F. (**2002**). "Nonparametric analysis of longitudinal data in factorial experiments," New York, NY: J. Wiley.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., and Cohen, J. D. (**2010**). "Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function," Cogn. Affect. Behav. Neurosci., **10**(2), 252-269. doi:10.3758/CABN.10.2.252.

Hallgren, M., Larsby, B., and Arlinger, S. (**2006**). "A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition," Int. J. Audiol., **45**(4), 227-237. doi:10.1080/14992020500429583.

Heitz, R. P., Schrock, J. C., Payne, T. W., and Engle, R. W. (**2008**). "Effects of incentive on working memory capacity: behavioral and pupillometric data," Psychophysiology, **45**(1), 119-129. doi:10.1111/j.1469-8986.2007.00605.x.

Kahneman, D. (**1973**). "Attention and effort," Englewood Cliffs, N.J., Prentice-Hall.

Kret, M. E., and Sjak-Shie, E. E. (**2019**). "Preprocessing pupil size data: Guidelines and code," Behav. Res. Methods, **51**(3), 1336-1342. doi:10.3758/s13428-018-1075-y.

Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., and Kramer, S. E. (**2014**). "The pupil response is sensitive to divided attention during speech processing," Hear. Res., **312**, 114-120. doi:10.1016/j.heares.2014.03.010.

Lunner, T., Rudner, M., Rosenbom, T., Agren, J., and Ng, E. H. (**2016**). "Using Speech Recall in Hearing Aid Fitting and Outcome Evaluation Under Ecological Test Conditions," Ear Hear., **37**(Suppl 1), 145S-154S. doi: 110.1097/AUD.0000000000000294.

Moon, S. K., Mun, H. A., Jung, H. K., Soli, S. D., Lee, J. H., and Park, K. (**2005**). "Development of Sentences for Korean Hearing in Noise Test (KHINT)," Korean J. Otolaryngol., **48**, 724-728. doi:10.1097/AUD.0b013e31803154d0.

Ng, E. H., Rudner, M., Lunner, T., Pedersen, M. S., and Ronnberg, J. (**2013**). "Effects of noise and working memory capacity on memory processing of

speech for hearing-aid users," Int. J. Audiol., **52**(7), 433-441. doi:10.3109/14992027.2013.776181.

Ng, E. H., Rudner, M., Lunner, T., and Ronnberg, J. (**2015**). "Noise reduction improves memory for target language speech in competing native but not foreign language speech," Ear Hear., **36**(1), 82-91. doi:10.1097/AUD.0000000000000080.

Noguchi, K., Gel, Y. R., Brunner, E., and Konietschke, F. (**2012**). "nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments," Journal of Statistical Software, **50**(12), 1-23.

Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., and Lunner, T. (**2018**). "Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response," Hear. Res., **365**, 90-99. doi:10.1016/j.heares.2018.05.003.

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., . . . Kramer, S. E. (**2017**). "Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation," Hear. Res., **351**:68-79. doi: 10.1016/j.heares.2017.05.012. Epub 2017 May 1025.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., . . . Wingfield, A. (**2016**). "Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL)," Ear Hear., **37**(Suppl 1), 5S-27S. doi: 10.1097/AUD.0000000000000312.

Richmond, L. L., Morrison, A. B., Chein, J. M., and Olson, I. R. (**2011**). "Working memory training and transfer in older adults," Psychol. Aging, **26**(4), 813-822. doi:10.1037/a0023631.

Rudner, M., and Lunner, T. (**2014**). "Cognitive spare capacity and speech communication: a narrative overview," BioMed Res. Int., 2014, 869726. doi:10.1155/2014/869726.

Simonsen, L. B., Hietkamp, R. K., and Bramsløw, L. (**2016**). "Learning effects of repeated exposure to Hearing In Noise Test," Paper presented at the Annual Conference of the British Society of Audiology, Coventry, UK.

Tsukahara, J. S., Harrison, T. L., and Engle, R. W. (**2016**). "The relationship between baseline pupil size and intelligence," Cogn. Psychol., **91**, 109-123. doi:10.1016/j.cogpsych.2016.10.001.

Winn, M. B., Wendt, D., Koelewijn, T., and Kuchinsky, S. E. (**2018**). "Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started," Trends Hear., **22**, 1-32. doi:10.1177/2331216518800869.

Zekveld, A. A., Koelewijn, T., and Kramer, S. E. (**2018**). "The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge," Trends Hear., **22**, 2331216518777174. doi:10.1177/2331216518777174.

Development of a Danish test material for assessing speech-in-noise reception in school-age children

SHNO KOIEK[1], JENS BO NIELSEN[2], LAILA KJÆRBÆK[3], MARIA BALTZER GORMSEN[4] AND TOBIAS NEHER[1,*]

[1] *Institute of Clinical Research, University of Southern Denmark, Odense, Denmark*

[2] *Hearing Systems, Technical University of Denmark, Lyngby, Denmark*

[3] *Department of Language and Communication, University of Southern Denmark, Odense, Denmark*

[4] *Department of Audiology, Odense University Hospital, Odense, Denmark*

For the audiological assessment of the speech-in-noise abilities of children with normal or impaired hearing, an appropriate test material is required. However, there is no standardized speech material for children in Denmark. The purpose of the current study was to develop a Danish sentence material suitable for school-age children. Based on the 600 test sentences from the Danish DAT corpus (Nielsen *et al*., 2014), 11 test lists comprising 20 sentences each were carefully compiled. These lists were evaluated in terms of their perceptual similarity and reliability with a group of 20 typically-developing normal-hearing children aged 6-12 yrs. Using stationary speech-shaped noise and diotic stimulus presentation, speech reception thresholds (SRTs) were measured twice per list and participant at two separate visits. The analyses showed that six test lists were perceptually equivalent. These lists are characterized by a grand average SRT of −2.5 dB SNR, a test-retest improvement of 0.4 dB, and a within-subject standard deviation of 1.1 dB SNR. The remaining test lists produced slightly higher SRTs but were generally also usable. Altogether, it is concluded that the developed test material is suited for assessing speech-in-noise reception in Danish school-age children.

INTRODUCTION

Children are often exposed to noise (e.g., in classrooms), which causes difficulties with speech understanding (e.g., Shield and Dockrell, 2003). Reliable methods for assessing speech reception in noise in school-age children are essential, specifically when difficulties in noise are suspected. In Germany, for instance, the "Oldenburger Kinder Satztest" (OlKiSa) was developed for that purpose (Neumann *et al*., 2012). OlKiSa consists of three-word pseudo-sentences, each with a numeral, an adjective, and an object noun (e.g., 'four red flowers') which is applicable for children from age four.

*Corresponding author: tneher@health.sdu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

In Denmark, a number of speech materials are available for clinical and research purposes, e.g., DANTALE-I (Elberling *et al.*, 1989) or DANTALE-II (Wagener *et al*., 2003). DANTALE-I includes lists of monosyllabic words for the measurement of the discrimination score (DS) for children and younger children. DANTALE-II contains semantically unpredictable, nonsensical sentences that are difficult to memorize. However, significant learning effects have been observed (Wagener *et al*., 2003).

Nielsen and Dau (2009) developed a Danish speech intelligibility test named conversational language understanding evaluation (CLUE). This test was based on the principles and test procedure of the original Hearing in Noise Test (HINT; Nilsson *et al*., 1994). However, the speech materials of CLUE are not well-suited for children (Nielsen and Dau, 2011). In 2010, Nielsen and Dau developed a Danish version of HINT that was based on the same speech materials as CLUE with some modifications (Nielsen and Dau, 2011). Furthermore, Nielsen *et al.* have developed a Danish open-set speech corpus (DAT) containing 600 unique sentences that were systematically distributed in 30 test lists with three talkers (Nielsen *et al*., 2014). The DAT material was validated using free-field speech-on-speech measurements in normal-hearing and hearing-impaired Danish adult listeners. However, there is currently no standardized Danish speech test that is suited for testing speech reception in noise in children. An open-set speech material that simulates a real-life communication situation more than a close-set speech material is required to assess speech reception in noise abilities among children.

To summarize, there is no standardized Danish test material that is suited for children. The purpose of the current study was to address this shortcoming. In particular, the aim was to develop a set of test lists that is characterized by small training effects, high test list equivalence and low measurement uncertainty, which are suited for assessing speech reception in noise in Danish school-age children. To ensure reliable SRT measurements that are independent of the applied test list, these should result in very similar SRT measurements. In the current study, the reliability of the results was examined by a retest 5-15 days after the initial test.

MATERIALS AND METHODS

Generation of test lists

For the compilation of the test lists, the 600 test sentences from the Danish DAT corpus (Nielsen *et al*., 2014) were used. The DAT corpus is an open-set, low-context, multi-talker speech corpus. All sentences in this corpus have a fixed, simple structure. That is, they start with a name [Dagmar (D), Asta (A) or Tine (T)] and contain two short keywords (nouns), e.g., “Dagmar tænkte på en teske og en næse i går” (“Dagmar thought of a teaspoon and a nose yesterday”). In terms of their semantic properties, the noun pairs are not related, which makes them difficult to predict. For each name, there are 200 test sentences uttered by one of three professional female talkers with similar voice characteristics. For the current study, 220 sentences suitable for children with keywords judged to belong to the vocabulary of a typical 6-year-old were selected. For this selection, two audiologists and one psycholinguist (three of the

authors of the current study) individually went through all 600 sentences of DAT corpus. They each individually decided whether which sentences are suitable for a 6 years old child. Finally, they selected those 220 sentences that they all agreed being suitable for children and belong to the vocabulary of a typical 6 years old child. These sentences were combined into 11 lists containing 20 sentences each. All sentences in a given list are uttered by the same talker and start therefore with the same name. Specifically, four D-lists, three A-lists and four T-lists were created. The intelligibility was defined as the average SNR at which both keywords could be correctly identified by the 16 participating adults (Nielsen *et al*., 2014). In the present project, these intelligibilities were assumed to be valid for the 220 child-friendly sentences. In the children's lists, the sentences with relatively high and low intelligibilities were counterbalanced at the beginning of each test list, while the sentences with approximately equal intelligibility were put in towards the end of each list.

Participants

Twenty typically-developing, normal-hearing children (13 female) participated in the study. They were aged 6-12 yrs (mean: 8.7 yrs). All participants fulfilled the following inclusion criteria: (i) normal middle-ear function, (ii) pure-tone hearing thresholds ≤ 25 dB HL at all standard audiometric frequencies from 250 to 8000 Hz, (iii) native Danish speakers, (iv) normal language development, and (v) normal cognitive function. Middle-ear function and hearing thresholds were assessed using standard tympanometry and audiometry. Language development of the children was assessed using the Peabody Picture Vocabulary Test. Cognitive development was assessed based on parental reports.

Apparatus and procedures

All measurements were conducted in a soundproof booth. To evaluate the 11 created test lists in terms of their perceptual similarity and reliability, SRT measurements were made. The speech stimuli were presented diotically in stationary speech-shaped noise via supra-aural headphones (Sennheiser HDA200). The order of the test lists was balanced across the participants. The starting level of the speech signal was 67 dB SPL. The level of the noise was fixed at 60 dB SPL. The SRTs were measured using the adaptive procedure from the standard HINT (Nielsen *et al*., 2014). Before the start of the actual measurements, the participants were verbally instructed to repeat the two keys words in each sentence. In case of any doubts, they were encouraged to guess. Reponses were scored as correct if both keywords were repeated accurately. In this case, the level of the target speech was decreased by 2 dB. Otherwise, the level of target speech was increased by 1 dB. To familiarize them with the procedure and the speech material, all participants performed one SRT measurement in quiet and two SRT measurements in noise. The lists used for these purposes were training lists from the original DAT material (Nielsen *et al*., 2014). A short break was included after the first five SRT measurements and whenever a participant felt tired. A set of retest measurements was made on average 10 days (range: 5-19 days) after the first set of measurements.

Statistical analysis

The collected data were analyzed using SPSS version 25. To begin with, test-retest reliability was assessed, resulting in the data of one child being excluded from all subsequent analyses because of large inconsistencies. To assess the influence of the talker, visit and test list repeated-measures analyses of variance (ANOVAs) were performed on the SRTs. In all cases, a significance level of 5% was used.

RESULTS

The grand average SRT across all test lists and participants was −2.0 dB SNR with a standard deviation (SD) of 1.3 dB SNR. For the test measurements only, the mean SRT was −1.7 dB SNR with an SD of 1.5 dB SNR; for the retest measurements, the corresponding values were −2.4 and 1.4 dB SNR. The within-subject SD for all 11 test lists was 1.2 dB SNR. Figure 1 shows the mean list SRTs for the two visits.

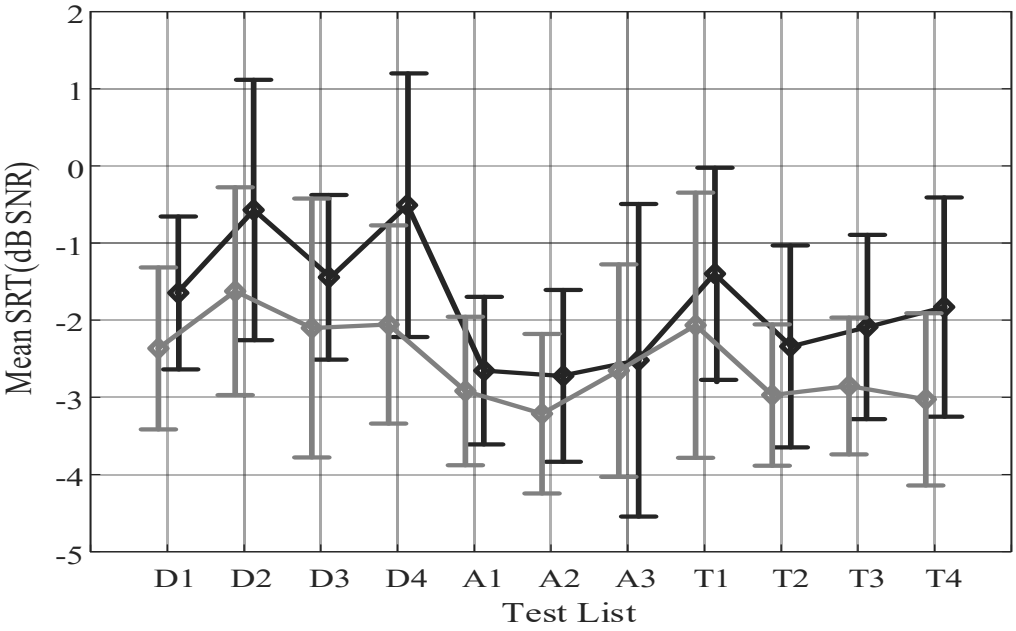


Fig. 1: Mean list SRTs for the first (solid line) and second (dashed line) visit.

Recall that the sentences of the D-, A- and T-lists were uttered by three different talkers. The overall mean SRTs of the three talkers were −1.4 dB SNR (D), −2.6 dB SNR (A), and −2.2 dB SNR (T), respectively. A one-way ANOVA comparing the mean SRTs of the three talkers showed a significant effect [$F_{(2, 206)} = 19.2, p < 0.001$]. Post hoc comparisons using Tukey's test revealed that the mean SRT of talker D was significantly higher than those of talkers A and T, whereas the mean SRTs of talkers A and T did not differ from each other (see Table 1).

Talker 1	Talker 2	Mean difference (dB)	*p*-value
T	A	0.4	0.126
D	T	0.8	< 0.001
	A	1.2	< 0.001

Table 1: Results of post hoc tests comparing the mean SRTs of talkers D, A and T.

To investigate the perceptual similarity of the seven test lists of talkers A and T, a two-way repeated-measures ANOVA with the within-subject factors visit and test list was carried out. This showed statistically significant effects of test list [$F_{(6, 108)} = 3.6$, $p = 0.002$] and visit [$F_{(1,18)} = 7.9$, $p = 0.012$]. Post-hoc comparisons using Tukey's test showed that the T1-list differed significantly from T2, T3 and A lists (all $p < 0.05$). T2, T3, T4, A1, A2, and A3 lists, on the other hand, did not differ from one another (all $p > 0.05$).

Figure 2 shows the mean SRTs of the test and retest of the 11 test lists. For eight of these test lists (T1, T2, T3, T4, A1, A2, A3, and D1), the mean SRTs were within 1 dB of each other. For the six lists that were found to be perceptually equivalent, the grand average SRT was – 2.5 dB SNR, the average test-retest improvement was 0.4 dB, and the within-subject SD was 1.1 dB SNR.

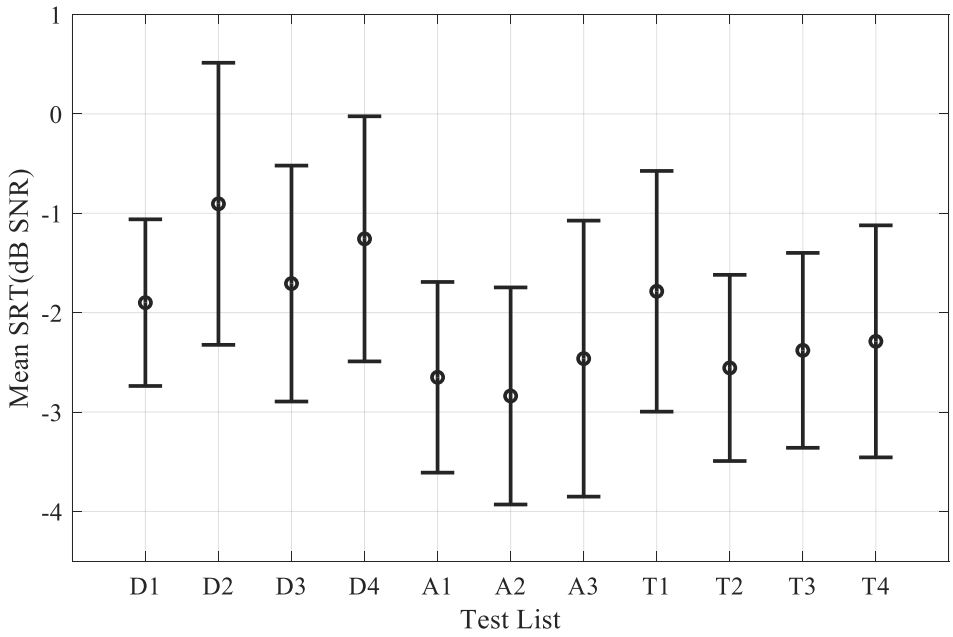


Fig. 2: A graph of mean SRTs and ±1 standard deviation for the 11 test lists.

Given that the children who participated in the current study covered a rather wide age span (6-12 yrs), we also investigated the effect of age on the SRT results. Figure 4 shows a scatter plot of age against the mean SRT. As expected, older children achieved lower (better) SRTs compared to younger children. The relationship between age and mean SRT was statistically significant ($r(19) = -0.53$, $p < 0.05$).

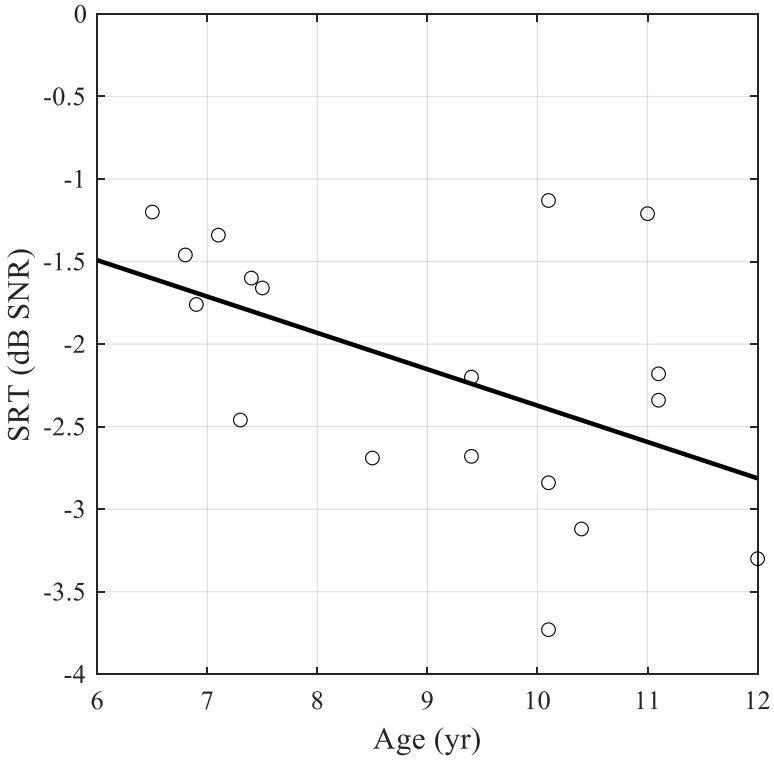


Fig. 3: Scatter plot of mean SRT versus age. The solid line shows a linear regression line that shows the average trend.

DISCUSSION

The aim of the current study was to develop a Danish test material, which is suitable for assessing speech reception in noise in school-age children. More specifically, the objective was to develop a set of test lists with small within-subject and between-list variation for performing SRT measurements with 6-12 yrs olds. Eleven test lists comprising 20 sentences each were created, and their equivalence was examined with the help of 20 typically-developing, normal-hearing native Danish children.

We verified our results by comparing them to the study by Nielsen and Dau with Danish HINT (Nielsen and Dau, 2011). However, Danish HINT has been studied among adult listeners. But in both studies, SRTs have been assessed using sentence reception in stationary speech-shaped noise. The overall mean SRT of the 11 lists in the current was −2.0 dB SNR, which is slightly higher than the mean SRT of the Danish HINT obtained with 16 normal-hearing adults (−2.5 dB SNR; Nielsen and

Dau, 2011). The within-subject SD for these lists was 1.2 dB SNR, which is somewhat larger than that of the Danish HINT (0.9 dB SNR). The relatively higher mean SRT and within-subject SD might be due to differences in the speech material used in the two studies. DAT sentences are without context whereas the HINT sentences are everyday sentences with contextual information. Another explanation could be the large age difference between the participants of the two studies (children vs. adults).

Since the sentences of the D-, A- and T-lists were uttered by three different talkers, we considered the influence of talker on our results. We found that the D-lists resulted in significantly higher mean SRTs than the A- and T-lists. Ideally, the test lists of a given speech material should result in very similar SRT measurements, so they can be used interchangeably. Based on our results, lists T2, T3, T4, A1, A2 and A3 are equivalent in terms of mean SRT. These test lists can be used in actual measurements in future speech-based research in 6-12 years old children in Denmark. The rest of the test lists including D1, D3, D4, and T1 were found with equivalent mean SRTs. They can be used as training lists in measurements. They can also be used in tests with less conditions of measurements. The mean test-retest improvement for these lists was 0.4 dB, corresponding to that observed for the Danish HINT (Nielsen and Dau, 2011). The within-subject SD across six test lists was 1.1 dB, which is slightly larger than the within-subject SD found by Nielsen and Dau (2011) with normal-hearing Danish adults (0.9 dB SNR). This can also be explained by speech material differences (without context vs with context) as well as age difference (children vs adults).

In addition to the within-subject SD, we also calculated the list-SRT SDs for lists T2, T3, T4, A1, A2 and A3. The result was 0.2 dB SNR. This result is very similar to the list-SRT SD of the Danish HINT (0.3 dB SNR). Furthermore, the maximum deviation from the overall mean SRT of 0.3 dB SNR was observed for list A2. It is smaller than the maximum deviation from the overall mean SRT that Nielsen and Dau (2011) found (0.6 dB SNR). This indicates a high equivalence of these six test lists with respect to the measured SRT.

CONCLUSION

Eleven test lists compiled from the Danish DAT corpus (Nielsen *et al*., 2014) were evaluated in terms of their perceptual similarity and reliability with 20 native Danish 6-12-year-old children. Six of these test lists (T2, T3, T4, A1, A2 and A3) were found to be suitable for speech-based studies among Danish 6-12-year-olds. These lists produced a grand average SRT of −2.5 dB SNR. The observed test-retest improvement of 0.4 dB, which suggests that reusing the lists after about 10 days is possible. The A- and T-lists produced mean SRTs that were within 1 dB of each other. The D-lists resulted in mean SRTs that were on average 1 dB higher than the other lists. They may be used for training purposes.

REFERENCES

Elberling, C., Ludvigsen, C., and Lyregaard, P. E. (**1989**). "DANTALE: A new Danish speech material," Scand. Audiol., **18**(3), 169-175. doi: 10.3109/01050398909070742

Nielsen, J., Dau, T., and Neher, T. (**2014**). "A Danish open-set speech corpus for competing-speech studies," J. Acoust. Soc. Am., **135** (1), 407-420. doi: 10.1121/1.4835935

Nielsen, J. B., and Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol., **48**(10), 729-741. doi: 10.1080/14992020903019312

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol, **50**(3), 202-208. doi: 10.3109/14992027.2010.524254

Nilsson, M., Soli, S.D., and Sullivan, J.A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am., **95** (2), 1085-1099. doi: 10.1121/1.408469

Neumann, K., Baumeister, N., Baumann, U., Sick, U., Euler, H. A., and Weißgerber, T. (**2012**). "Speech audiometry in quiet with the Oldenburg Sentence Test for Children," Int. J. Audiol., **51** (3), 157-163. doi: 10.3109/14992027.2011.633935

Shield, B. M., and Dockrell, J. E. (**2003**). "The effects of noise on children at school: a review," Build. Acoust., **10**(2), 97-116. doi: 10.1260/135101003768965960

Wagener, K., Josvassen, J.L., and Ardenkjær, R. (**2003**) "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol., **42** (1),10-17. doi: 10.3109/14992020309056080

Looking for objective correlates between tinnitus and cochlear synaptopathy

CHIARA CASOLANI[1,*], JAMES MICHAEL HARTE[2] AND BASTIAN EPP[1]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Interacoustics Research Unit, part of Oticon A/S, DK-2800 Lyngby, Denmark*

Tinnitus is the perception of a sound in the absence of acoustic stimulation. While usually connected to a hearing loss, there exists a subset of tinnitus sufferers with audiologically normal hearing, whose tinnitus was often initiated by a noise trauma. Noise-induced tinnitus might be connected to the noise exposure that leads to a permanent impairment of the hearing system without affecting sensitivity to sound. This is commonly referred to as hidden hearing loss (HHL) and might be connected to cochlear synaptopathy. The hypothesis that HHL is one of the causes underlying tinnitus is based on suppositions that both phenomena are related to deafferentation of auditory nerve fibres and related central gain adjustments. To investigate this connection, a screening procedure consisting of high frequency audiometry (HFA), tinnitus likeness spectrum and loudness, psychophysical tuning curves (PTC) and tinnitus masking curves (TMC), adaptive categorical loudness scaling, and middle-ear muscle reflex test was developed. Pilot results show that all measurements can be completed within a short time frame, due to a Bayesian procedure being adopted to measure HFA, PTC and TMC. These procedures may contribute to investigating the connection between tinnitus and HHL with a large number of outcome measures. This connection will provide important insights toward the development of better diagnoses and treatment methods.

INTRODUCTION

Tinnitus is a complex phenomenon whose causes and mechanisms are not yet completely understood. Research in this topic is still inconclusive, and treatments are still not effective in many cases. In addition, the results of different studies are inconsistent and often suggest opposing theories to explain the observed phenomena. Because tinnitus has many underlying causes, a categorization of tinnitus is required to obtain an understanding of the underlying mechanisms. Many studies include listeners with tinnitus and a hearing loss (Baracca *et al.*, 2011), which complicates the interpretation of the results. An interesting population to study is therefore tinnitus sufferers with a normal audiogram. It has been suggested that a deafferentation of auditory nerve fibres in the inner ear leads to a reduced excitatory input into the brainstem, which, in turn, leads to a compensatory increase in neural gain and tinnitus.

*Corresponding author: chiarac@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

This will be in agreement with numerical simulations (Schaette & McAlpine, 2011). This hypothesis is also in line with recent findings in animal models finding that noise overexposure followed by a temporal threshold shift leads to deafferentation in the inner ear (Kujawa & Liberman, 2015). The aim of the present study was to develop a screening procedure composed of psychophysical and acoustic measures suggested to assess tinnitus and the potential presence of synaptopathy, and thereby to categorize a subclass of tinnitus sufferers with normal hearing thresholds. One of the main underlying assumptions is that a noise trauma is the triggering event for the generation of tinnitus by initiating a progressive degeneration of spiral ganglion neurons (SGN) caused by noise-induced synaptopathy (Fig. 1). A shift in hearing thresholds will first be evident after a critical time and after a critical amount of SGN degeneration has been reached (Lobarinas *et al.*, 2016).

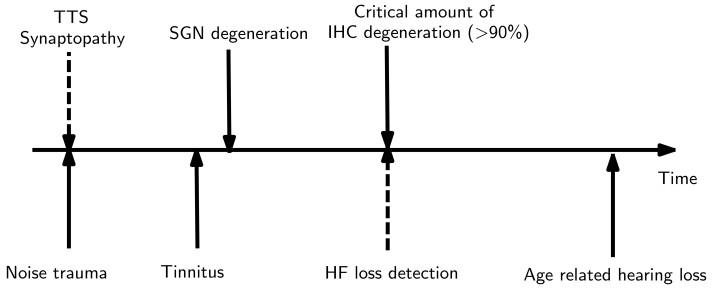


Fig. 1: Assumption behind the connection between tinnitus, cochlear synaptopathy and high-frequency hearing loss. It is assumed that a noise overexposure initiates a temporal threshold shift with subsequent cochlear synaptopathy. Reduced input into the brainstem then leads to a regulation of gain and an overexcitation of specific frequencies, leading to tinnitus. A progressive degeneration of spiral ganglion cells as a consequence of synaptopathy will then, after a critical amount of damage, lead to an increase in audiometric threshold.

Because animal models showed a higher prevalence for synaptopathy at tonotopic places connected to high frequencies, HFA was selected as a proxy of the presence of synaptopathy. It has also recently been suggested that the presence of tinnitus might be related to the functioning of the efferent system (Wojtczak *et al.*, 2017) which, in turn, might be connected to cochlear synaptopathy (Furman *et al.*, 2013). Middle ear muscle reflex (MEMR) measurements will be used as an acoustic characterization measure, based on results by Wojtczak *et al.* (2017) where tinnitus sufferers had lower MEMRs compared to a control group. Psychophysical tuning curves (PTC) and tinnitus masking curves (TMC) have been shown to differ within a group of listeners suffering from tinnitus (Fournier *et al.*, In press). This variability provides information about the tuning properties of the system and the potential impact of the mechanism underlying tinnitus on tuning. Combining clinical measures of tinnitus assessment

with more experimental procedures in the same listeners will provide a more detailed profile of the tinnitus.

METHODS

The inclusion criteria for the subjects were: a) the tinnitus had to be constant but not intrusive or bothersome, i.e. the ideal participant can ignore the tinnitus the majority of the time. b) tinnitus had to be chronic, not related to acute temporal threshold shift (TTS) or upper respiratory tract infections (URTI). c) audiometric thresholds from 125 Hz to 8 kHz measured with standard audiogram had to be lower than 20 dB HL. The data of the tinnitus group will be compared to the data of a control group, matched in hearing thresholds and age. To have a further characterization of the tinnitus, the Tinnitus Handicap Inventory (THI) questionnaire will be included.

The HFA, PTC (both at 1 kHz and at tinnitus frequency) and the TMC were implemented based on a Bayesian algorithm (Schlittenlacher *et al.*, 2018) for standard audiometry. In short, a continuous function is fitted to estimated values based on a prior grid and a Bayesian statistical approach maximizing the information in each trial to cover the desired spectral range. Initially, specific frequencies with a distance of $\frac{1}{4}$th octave (See Fig. 2) were tested to get at least one positive ("I heard the tone") and one negative answer ("I didn't hear the tone"). The algorithm then iterated through the frequencies and finally applied a continuous fit to the data points. Using an initial grid allowed a reduction in measurement time.

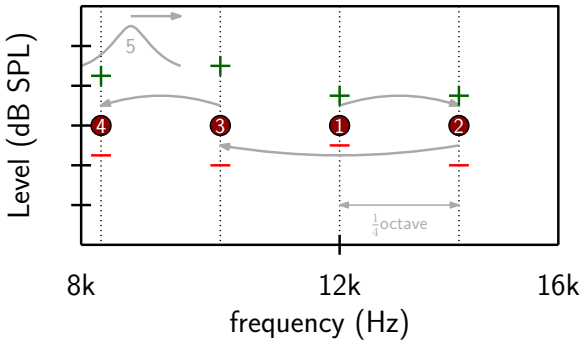


Fig. 2: Bayesian procedure used in the high frequency audiometry. Starting at 12 kHz, an initial estimate of the threshold was made by obtaining a positive and a negative response from the listener. Additional estimates were obtained in the frequency range between 8 kHz and 16 kHz with a spectral distance of 1/4th octaves. After estimation of the lowest frequency threshold, the Bayesian procedure was applied, selecting the frequency and level by maximizing the information. Finally, all estimated points were connected by a continuous differentiable fitting function.

All three experiments consisted of 50 trials. For the HFA and PTC, each interval

consisted of 3 pulses of 250 ms duration with a rise/fall time of 20 ms spaced by 100 ms silence intervals. For the TMC, each interval consisted of one pulse of 2s with a rise/fall time of 20 ms. The sampling rate was set to 48 kHz and the step size between the frequencies was set to 0.1 octaves. In the HFA the participant had to indicate if the tone was detected (YES) or not (NO). The tested frequency range was 8 kHz to 16 kHz. To reduce bias effects, 10% of silent trials (in addition to the 50 trials) were randomly presented to test the reliability of the participant. The PTC was measured for a probe tone at 1 kHz and at the listeners' tinnitus frequency in the presence of a narrowband noise masker. The bandwidth of the noise masker was 20% of the probe frequency and at maximum 320 Hz (Kluk & Moore, 2005). The masker centre frequency varied in the range of 0.5 kHz to 2 kHz (for the 1 kHz probe) or within a 1 octave width centred at the tinnitus frequency. For the TMC, the same approach as the PTC was used but in absence of the probe.

To measure the middle-ear muscle reflex (MEMR), wideband tympanometry (WBT) was measured (226 Hz to 8 kHz) to determine the tympanometric peak pressure. Then, the MEMR was measured in correspondence to the maximum of the absorbance found with WBT. The stimulation pattern (Fig. 3) is based on Liberman *et al.* (personal communication). The elicitor was a broad band white noise noise presented ipsilaterally (Interacoustics Titan Suite). The levels were increasing from 60 to 95 dB SPL, or up to 110 dB SPL if the participant did not indicate annoyance.

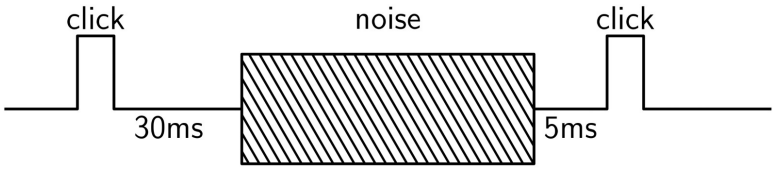


Fig. 3: The stimulus used to measure the middle ear muscle reflex (MEMR). A click preceded a burst of white noise, followed by another click . All stimuli were presented on the ipsilateral side. The paradigm was implemented using the Interacoustics TITAN research platform with custom code.

RESULTS - EFFICACY OF THE PIPELINE

High frequency audiometry

The modified HFA procedure was compared to a 3-alternative forced choice procedure (AFC) (Fig. 4). The figure shows example results for the same subject measured with both the procedures. AFC results are shown by filled symbols, the Bayesian algorithm by circles, crosses and solid line. The AFC procedure took between 8 and 10 minutes to complete, while the Bayesian procedure between 4 and 6 minutes. Moreover, the AFC procedure provided estimates of five discrete frequencies, while the Bayesian

procedure provided a continuous estimate of the threshold. For this participant (Fig. 4, control group) the deviations between the AFC results and the estimate based on the Bayesian procedure were below 5 dB. Other participants showed differences between the two methods around 10 dB and even higher, especially at 16 kHz. The similarity of the two methods is promising, but due to the intrinsic high variability found in high-frequency audiometry especially from 14 to 16 kHz (Schmuziger *et al.*, 2004), the comparability of the two measures needs to be further investigated.

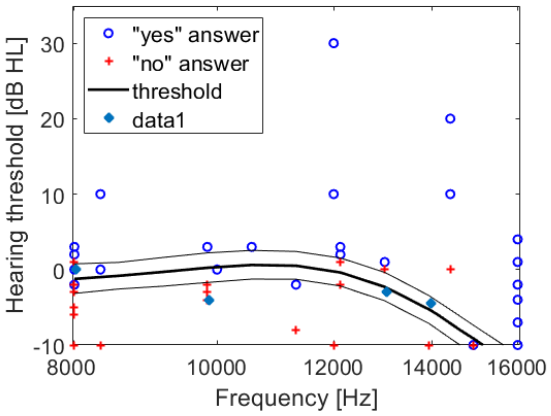


Fig. 4: Comparison of the results obtained using an alternative-forced choice procedure (AFC, filled symbols) and the Bayesian algorithm. Open symbols indicate positive responses of the listeners (tone heard), crosses negative responses (tone not heard). The black line indicates the final estimate with an estimate of the uncertainty.

Psychophysical Tuning curve (PTC) and Middle Ear Muscle Reflex (MEMR)

Figure 5 shows an example of a PTC using a probe tone at 1 kHz for a normal hearing subject (control group). Symbols indicate listeners' responses ("heard", circles; "not heard", red crosses). The solid line indicates the interpolation of the measured points. The PTC showed a clear minimum around the frequency of the probe tone and shallow slopes at the minimum and maximum frequencies. Compared to classical methods where single frequencies are measured, the peak was broader (Q10 = 4). The lack of a sharp peak is a consequence of the fitting procedure which requires the function to be differentiable. Despite the lack of a sharp peak, a clear tuning was found which might be sufficient for a rough classification, but not for indications of, for example, potential outer hair cell loss. The obtained continuous estimate might, however, allow different metrics of PTC tuning using this method. Fig. 6 shows an example result of the measured MEMR (control group). The difference in absorbance was higher for higher levels of the elicitor, in agreement with data from the literature. The difference across elicitor levels was small for the lowest (250 Hz) and the highest (800 Hz) frequencies,

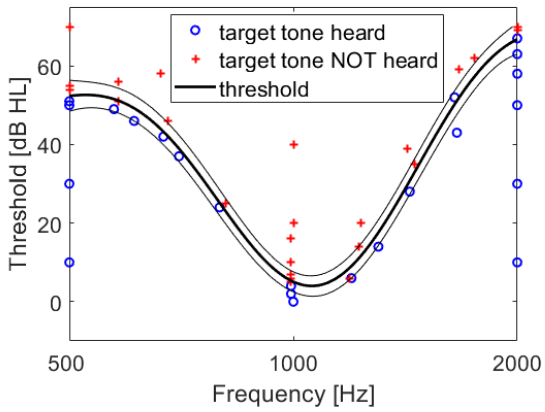


Fig. 5: An example of PTC using the Bayesian algorithm. Compared to literature, the peak is broader using this algorithm. The algorithm used for PTC was modified to use a finer grid and a weighting function in order to collect more data around the frequency of the presented tone (1 kHz) compared to the HFA. The Gaussian process interpolation will provide a differentiable function which makes the result deviate from the classical shape with a sharp tip.

showed a crossing point around 1100 Hz and the highest excursion towards negative absorbance changes between 500 and 700 Hz. The response was clearly visible even after only short measurement time. It is interesting to note that the stimulation with a noise rather than a pure tone as in standard protocols has been reported as less bothering. While of minor importance for normal hearing listeners, this plays an important role when handling participants that suffer from tinnitus. Moreover, the use of the wideband acoustic reflex is particularly recommended for its sensitivity (Hein *et al.*, 2017).

DISCUSSION

The preliminary results for the HFA for a listener in the control group indicate that the Bayesian procedure developed for standard audiometric frequencies (Schlittenlacher *et al.*, 2018) can also be applied to high-frequency audiometry. Despite the challenges connected to high-frequency audiometry in terms of ear-canal acoustics and challenges of the listeners to decide about the presence or absence of the stimulus, HFA obtained with AFC and the Bayesian procedure were very similar in the pilot data. In comparison with the AFC procedure, the Bayesian procedure not only provided an advantage in terms of measurement duration, but also allows an estimation of the threshold directly at any frequency within the measured interval. This is an advantage for measures that rely on the tinnitus frequency identified by the listeners. The comparison between PTC at tinnitus frequency and TMC has been proposed by

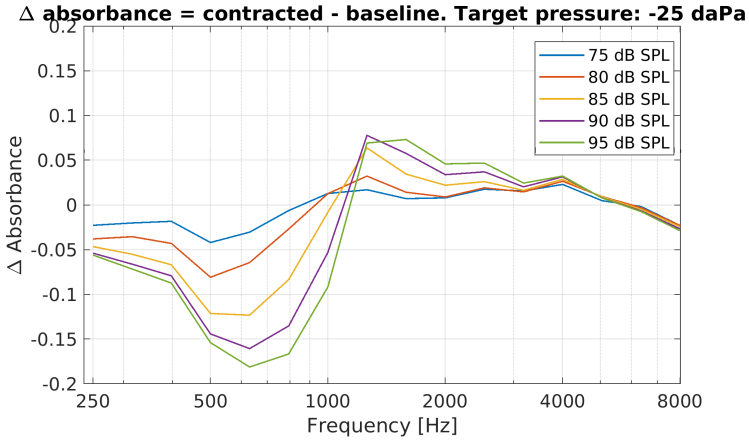


Fig. 6: The MEMR of a healthy subject. The figure shows the changes in the ear-canal sound pressure followed by the presentation of the elicitor for different levels.

(Fournier *et al.*, In press). In that study, the participants were clustered into three groups: 1) V/U shape of the PTC and TMC, 2) V/U shape of the PTC and flat TMC and 3) flat PTC and TMC. The PTC test is a measure of the frequency tuning of the auditory system. In connection to tinnitus and synaptopathy, differences in the PTC for tinnitus sufferers with a normal audiogram can reveal differences about the underlying neural mechanisms. These might differ between listeners suffering from synaptopathy compared to listeners not affected by synaptopathy. The TMC is a way to characterize and quantify the masking effect of an external stimulus on the tinnitus. Depending on the neural origin of the tinnitus, this effect might be different for peripheral mechanisms (including synaptopathy) and central mechanisms. In the study by (Fournier *et al.*, In press) a relationship between the maskability and the type of tinnitus (tonal or noise) was found. Hence, PTC and TMC might provide a more fine grained characterization of the tinnitus sufferers' profile.

The total duration of the suggested screening method is about one hour. This shortened measurement time was mainly achieved by using the Bayesian procedure rather than classical AFC procedures and by using optimized clinical equipment for the measurement of the MEMR. The decreased duration might play an important role when dealing with tinnitus sufferers which often are subjected to stress and report that their tinnitus can be modulated in intensity by external factors.

In conclusion, the combination of the suggested measures and the application of the Bayesian procedure has the potential to provide a detailed, yet fast characterization of the mechanisms underlying tinnitus in listeners with a normal audiogram.

ACKNOWLEDGEMENTS

This work was funded by the TIN-ACT EU initial training network (grant agreement 764604). We would like to thank Josef Schlittenlacher for help with the Bayesian procedure.

REFERENCES

Baracca, G., Del Bo, L., and Ambrosetti, U. (**2011**). "Tinnitus and hearing loss," Textbook of Tinnitus, Springer New York, 285–291, doi: 10.1007/978-1-60761-145-535.

Fournier, P., Wrzosek, M., Paolino M., Paolino F., Quemar A., and Norena A. (**In press**). "Comparing the patterns of tuning for tinnitus and tinnitus-like sound," Trends Hear.

Furman, A.C., Furman, A.C., Kujawa, S.G., Kujawa, S.G., and Liberman, M.C. (**2013**). "Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates," J. Neurophysiol., **110**(3), 577–586, doi: 10.1152/jn.00164.2013.

Hein, T., Hatzopoulos, S., Skarzynski, P., and Colella-Santos, M. (**2017**). "Wideband Tympanometry," Advances in Clinical Audiology, 29-42, doi: 10.5772/67155.

Kluk, K., and Moore, B.C.J. (**2005**). "Factors affecting psychophysical tuning curves for hearing-impaired subjects with high-frequency dead regions," Hearing Res., **200**(1-2), 115–131, doi: 10.1016/j.heares.2004.09.003.

Kujawa, S.G., and Liberman, M.C. (**2015**). "Synaptopathy in the noise-exposed and aging cochlea: Primary neural degeneration in acquired sensorineural hearing loss," Hearing Res., **330**(Part B, Sp. Iss. SI), 191–199, doi: 10.1016/j.heares.2015.02.009.

Lobarinas, E., Salvi, R., and Ding, D. (**2016**). "Selective Inner Hair Cell Dysfunction in Chinchillas Impairs Hearing-in-Noise in the Absence of Outer Hair Cell Loss," J. Assoc. Res. Oto., **17**(2), 89–101, doi: 10.1007/s10162-015-0550-8.

Schaette, R., and McAlpine, D. (**2011**). "Tinnitus with a normal audiogram: Physiological evidence for hidden hearing loss and computational model," J. Neurosci., **31**(38), 13452–13457, doi:10.1523/jneurosci.2156-11.2011.

Schlittenlacher, J., Turner, R.E., and Moore, B.C.J. (**2018**). "Audiogram estimation using Bayesian active learning," J. Acoust. Soc. Am., **144**(1), 421–430, doi: 10.1121/1.5047436.

Schmuziger, N., Probst, R., and Smurzynski, J. (**2004**). "Test-Retest Reliability of Pure-Tone Thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-2 Earphones," Ear Hear., **25**(2), 127–132, doi: 10.1097/01.aud.0000120361.87401.c8.

Wojtczak, M., Beim, J.A., and Oxenham, A.J. (**2017**). "Weak middle-ear-muscle reflex in humans with noise-induced tinnitus and normal hearing may reflect cochlear synaptopathy," eNeuro, **4**(6), e0363–17.2017, e0363–17.2017, doi: 10.1523/eneuro.0363-17.2017.

Comparison of clinical feasibility of behavioural and physiological estimates of peripheral compression

MICHAL FERECZKOWSKI, TORSTEN DAU, AND EWEN N. MACDONALD

Hearing Systems Section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark

Previous research has shown that the rate of peripheral compression (estimated from the slope of the basilar-membrane input-output function) is not correlated with the pure-tone sensitivity (the audiogram). However, efficient estimation of peripheral compression has proven challenging and the methods are based on several assumptions. The aim of this study was to investigate and compare results from three methods of estimating peripheral compression in terms of their accuracy and clinical feasibility. Two psychoacoustic behavioural measures, based on forward (temporal masking curves, TMC) and simultaneous masking with notched-noise (NN), were investigated together with a physiological, distortion-product otoacoustic-emissions (DPOAE), based measure. Forty-five hearing-impaired (HI) listeners with mild-to-moderate hearing loss were tested. Correlation analysis of the data was performed, including partial-correlations, in order to factor out the potential influence of the pure tone-thresholds on the compression estimates. The results demonstrated limitations of each of the considered methods; however, the experiment involving estimates of auditory filters showed good stability and small training requirements across the listeners.

INTRODUCTION

Currently, the main method used to categorize hearing loss is the audiogram, reflecting sensitivity to pure tones. While useful, it is not sufficient to predict supra-threshold perception and performance of individual hearing-impaired (HI) listeners. In other words, two individuals with similar audiograms can differ widely on perceptual and hearing-aid outcome measures (Fereczkowski *et al*., 2017). Therefore, there is a need for additional supra-threshold measures that may improve categorization of HI listeners.

Plack *et al*. (2004) used the temporal masking curve (TMC) method of Nelson *et al.*, (2001) to measure peripheral compression behaviourally and found the compression rate to be uncorrelated with audiometric thresholds for listeners with mild-to-moderate hearing loss. However, a major challenge in estimating peripheral compression with psychoacoustical methods is the relatively low time-efficiency of the measurements. In particular, the TMC method is very time consuming, requires extensive training before listeners reach stable performance, and the estimated thresholds are often characterized by large within-subject variability (Rosengard *et al.*, 2005). Physiological methods, based on otoacoustic

*Corresponding author: mfer@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

emissions (OAE), have been developed (e.g., Kummer *et al.*, 1998) but are also limited by variability, such as notches and plateaus in the estimated input/output (I/O) characteristics (e.g., Johannesen and Lopez-Poveda, 2010).

Estimates of auditory filter characteristics, and thus frequency selectivity, can be derived behaviourally by examining how thresholds for a tone in notched noise varies as a function of notch bandwidth (Patterson *et al.*, 1976). Since the outer hair cells (OHC) influence both the nonlinear compression and frequency selectivity in the cochlea, some characteristics of the auditory filter and estimates of cochlear compression should be related.

The aim of the present study was to investigate the relation between different measures believed to estimate cochlear compression (and reflect the OHC status), i.e., TMC- and OAE-based estimates and auditory filter bandwidths in a large group of listeners, to explore whether these measures are consistent and/or could complement each other.

METHODS

Participants

Forty-five HI listeners (twenty-three female), with a mean age of 72.5 years participated in the study. The listeners had sensorineural hearing loss (with an air-bone gap not greater than 10 dB at any audiological frequency) and were recruited and tested at the audiological departments of the Bispebjerg Hospital and the State Hospital (Rigshospitalet), located in Copenhagen, Denmark. In order to participate in the study, the listener's threshold at 1 or 2 kHz could not exceed 45 dB HL in at least one ear. All participants provided informed consent and the experiment was approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

Setup and Procedure

The experiments were performed over three sessions, on three separate days, in a double-walled listening booth. The OAE task was performed using the ER10-X probe system and the behavioural tasks were performed on a PC equipped with Matlab with stimuli presented via calibrated Sennheiser HDA200 headphones connected directly to an RME Fireface UCX soundcard. All tests focused on investigation of the cochlear nonlinearity at two frequencies: 1 and 2 kHz.

Distortion product otoacoustic emissions (DPOAEs) were measured on the first day, to make sure that a good seal to the better ear could be obtained with the probe tip. If a stable seal could not be obtained, the other ear was used for all further tests. Two DPOAE measurement paradigms were employed – one based on swept tones and the other on pure-tones. The calibration (including estimation of the forward-pressure level), swept-tone DPOAE measurement, and analysis (including source unmixing) procedures were very similar to the scissors-rule-based paradigm used in Anyfantakis *et al*. (2017), but with two major changes. First, the presentation levels of the second primary (L2) were set to 6 values that uniformly span the range from 40 to 80 dB SPL.

Second, in an attempt to reduce the biological noise, and thus improve the signal-to-noise ratio (SNR), participants were seated in the test booth with the lights turned on and could observe a digital timer indicating the time remaining until the next break. In the pure-tone paradigm, the same presentation levels, level rule, and primary frequency ratio (1.22) were used, as in the swept-tone paradigm. The pure-tones presented had a duration of 1.1 s, including 50 ms-long ramps. Nine recordings of each combination of test-frequency and presentation level were performed. While measurements within the swept-tone paradigm took around 90 minutes per participant, the pure-tone paradigm required only 2 minutes per frequency (with 6 levels measured).

The classic notched-noise paradigm (NN; Patterson *et al.*, 1976) was used to investigate the sharpness of the auditory filters and it was performed on day two. In the NN task, the broadband, 300 ms masker had a constant spectral density of 40 dB/Hz and the participants' task was to detect in which of two intervals a 200 ms target tone was present. A 1-up-3-down version of the Grid method (Fereczkowski, 2015) was used to adaptively obtain thresholds. The minimum notch width was set to 0 (which corresponds to the tone-in-noise threshold) and the maximum was set to 0.85 (as a proportion of the test frequency). The step sizes were 3 dB in the target-tone level and 0.05 in the notch width dimensions. Prior to the test, each participant was trained in the 2-alternative-forced-choice (AFC) task and the tone-detection thresholds were measured using the standard 2-AFC, 1-up-3-down paradigm. Next, guided-stepwise training was used to familiarize the listener with the NN task and a warm-up run with the Grid method was administered for the 1 kHz target tone. This training procedure typically took 10-15 minutes. Subsequently 3 test runs were executed at 1 kHz. If the standard error of the estimated thresholds (averaged across the tested notch-widths) exceeded 3 dB, a fourth run was administered and thresholds from the three final runs were averaged. Finally, a warm-up run was performed for the 2 kHz target and 3 or 4 test runs followed. On average, the NN task took a total of 66 minutes, which included all the described training procedures and breaks administered after every two runs.

The TMC experiments were performed in the last session. In this task, a 200 ms pure tone masker was followed by a 16 ms pure tone target (presented at 12 dB sensation level). 8 ms ramps were applied to both the masker and target tones. As in the NN task, a 2-AFC 1-up 3-down version of the Grid method was used. The minimum and maximum masker-target temporal gaps were 10 and 200 ms and the step sizes were set to 3 dB for the tone-level and 5 ms for the temporal gap. Three conditions were tested: 2 kHz off- and on-frequency and 1 kHz on frequency, in that order. In the on-frequency condition, the target and the masker tones were the same frequency. In the off-frequency condition, the masker frequency was set to 55% of the target frequency. Guided-stepwise learning was employed before each test condition and in each condition one warm-up and two test runs were performed. If the average standard error exceeded 3 dB, up to three extra runs could be administered and thresholds from the three final runs were averaged. The TMC experiment took, on average, 94

minutes, which included all the training, test sessions, and breaks administered after every two runs.

Data analysis

The DPOAE data from the swept-tone paradigm was processed using the same procedure as used by Anyfantakis *et al*. (2017), which included source unmixing. The distortion product (DP) response at a given input level was considered valid, if the estimated SNR exceeded 5 dB. For the case of the pure-tone paradigm, the 8 recordings with the lowest RMS were selected and high-pass filtered at 500 Hz in an attempt to reduce artefacts and excessive noise. The 8 recordings were then averaged and FFT analysis was performed on the 1 s long fragment that did not contain the ramps. The strength of the $2F_1$-F_2 component was recorded as the final DP response and the SNR criterion was 10 dB.

The DP I/O curves, resulting from both presentation paradigms, were fitted independently with a broken-stick (1, 2 or 3 sections) function using a similar constrained fitting procedure as Fereczkowski *et al.* (2017). However, no constraints were set on the location of the knee-points. Here, the slope of the most compressive portion of the I/O function, henceforth referred to as compression exponent (CE), was recorded.

To reduce the variance of the OAE-based CE estimates, and facilitate the comparison with the psychophysical methods, the CE values obtained from the pure-tone and swept-tone DPOAE paradigms were averaged to obtain the final OAE CE estimate. This was done since the CE values from both methods were found not to be significantly different, according to a permutation test ($p > .17$).

The TMC I/O curves were obtained by combining the on- and off-frequency data, averaged across all test-runs. The CE estimate was obtained in the same manner as for the DP I/O paradigm described above. For the auditory filter estimates, fitting rounded exponential functions occasionally led to unstable (i.e., very high) estimates of the rounding parameter, p. Thus, the sharpness of tuning in the NN task was estimated in a simpler way. Through interpolation, the notch width that resulted in a threshold 10 dB lower than the tone-in-noise threshold was found and termed the NN10 threshold.

To avoid potential distribution-related issues, Spearman correlation coefficients are reported.

RESULTS

The left panel of Fig. 1 shows the distribution of the test-ear thresholds for frequencies between 500 and 4000 Hz. The right panel of Fig. 1 shows the NN10 data as a function of the audiometric threshold at the tested frequency. Blue crosses represent data collected at 1 kHz and red circles represent the 2-kHz data. A reference NH value of the NN10 threshold was estimated to be 0.148 (Rosen and Baker, 1994), which corresponds well with the results here for hearing thresholds not exceeding 20 dB HL. The estimated correlation, measured across both frequencies, is moderate and

significant [$r_S(n = 81) = .51$, $p < 1e\text{-}4$], indicating that the sharpness of tuning generally decreases with increasing hearing thresholds. The NN10 value could be estimated in 81 out of a total of 90 cases (2 frequencies for each of 45 listeners). For the remaining cases, the threshold curves decreased by less than 10 dB over the tested notch-widths.

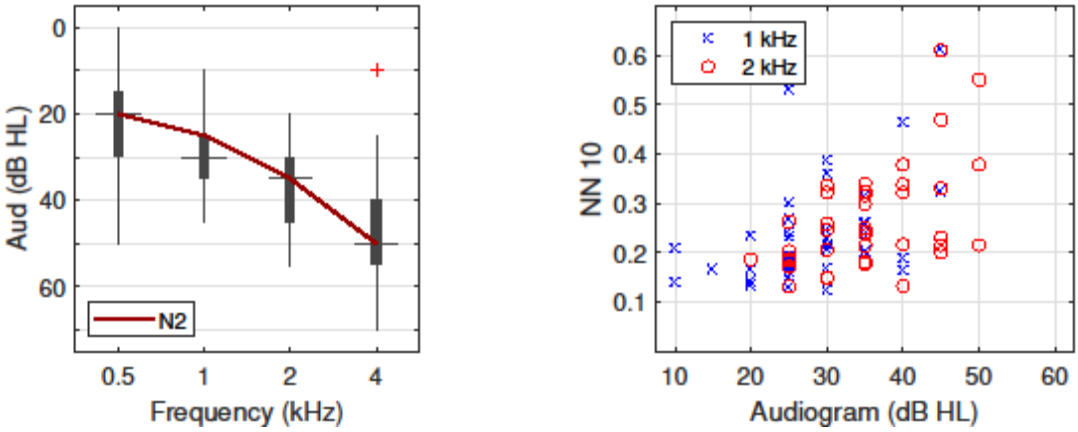


Fig. 1: Left: The boxplots show the better-ear hearing thresholds from the 45 listeners. The median thresholds were similar to the N2 audiogram (Bisgaard *et al.*, 2010). **Right:** Estimated tuning-sharpness (NN10) as a function of the audiometric threshold.

The left panel of Fig. 2 shows the corresponding data for CE estimates from the TMC task (TMC CE). Here, the estimated correlation is similar to the previous case, and significant [$r_S(89) = .50$, $p < 1e\text{-}4$].

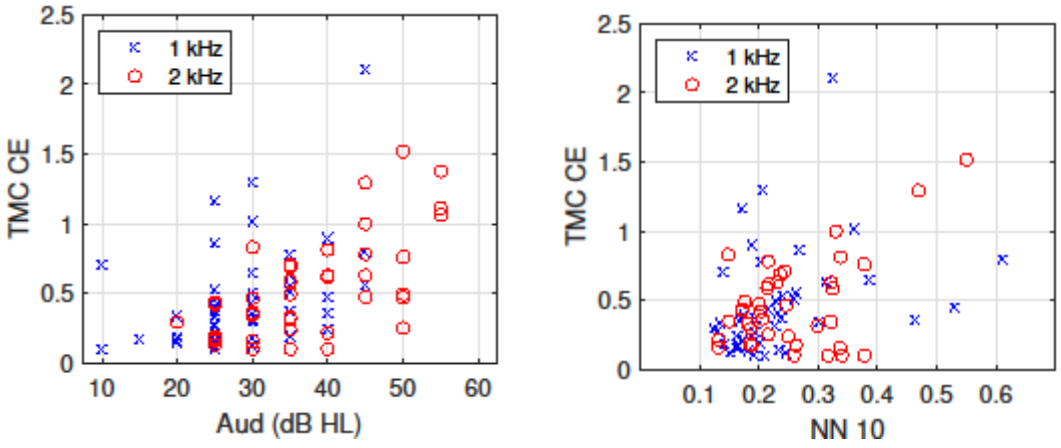


Fig. 2: Left: CE estimates from the I/O functions fitted to the TMC data as a function of the audiometric threshold. In both panels, blue crosses and red circles represent data from 1 and 2 kHz, respectively. **Right:** Compression exponent of the I/O function estimated from the TMC task as a function of tuning-sharpness (NN10).

The right panel of Fig. 2 presents the dependence between the NN10 and the TMC CE estimates from individual listeners. The correlation is weak, but

significant [$r_S(80) = .31$, $p = .0052$], which suggests that higher CE indicates broader auditory filters. However, the estimated partial correlation between TMC CE and NN10, controlling for the audiometric threshold, turned out to be weak and insignificant ($r_{SP}(80) = .12$, $p > .29$).

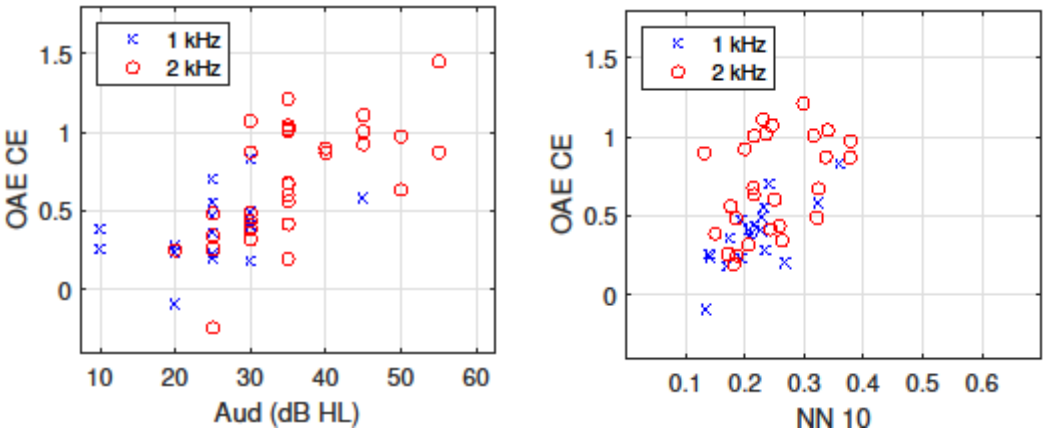


Fig. 3: Left: CE estimates from the DP I/O functions averaged across the OAE paradigms as a function of the hearing threshold. **Right:** Same CE estimates shown as a function of tuning-sharpness estimate. In both panels the OAE CE estimates tend to increase with the increasing abscissa value.

The left panel of Fig. 3 presents the dependence between the audiometric threshold and the OAE CE estimate from individual listeners. Also in this case the correlation is strong and significant [$r_S(48) = .72$, $p <$ 1e-4]. Note, however, that data was obtained in only 48 out of 90 possible cases (a valid OAE CE estimate required DPOAE responses measured at minimum least two input levels). The right panel of Fig. 3 shows a scatterplot of the same OAE CE estimates against the NN10 estimates. The correlation between the two estimates is moderate and significant [$r_S(45) = .55$, $p =$ 2e-4], and the partial correlation (controlling for the audiogram) is, again, moderate and significant [$r_{SP}(45) = .44$, $p = 0.0028$]. Moreover, a comparison of linear models of OAE CE reveals that a model using both hearing threshold and NN10 values as predictors is explains significantly more variance (adjusted $R^2 = 0.48$) than the model employing hearing thresholds alone (adjusted $R^2 = 0.39$) with $F(2, 42) = 21.48$ and $p <$ 1e-4. However, no significant correlation was found between the TMC CE and the OAE CE estimates ($r_{SP}(48) = 0.23$, $p = 0.12$)

DISCUSSION

The data analysis revealed that while there are significant correlations between hearing thresholds and auditory filter sharpness estimates, compression exponents from the TMC task, and compression exponents from the DP I/O

function, as well as a significant correlation between auditory filter sharpness and the OAE CE, the correlation between the TMC CEs and the auditory filter sharpness is mediated by the audiogram and there is no correlation between TMC and OAE CEs.

The findings concerning the TMC CE estimates are somewhat inconsistent with the previous literature. Plack *et al*. (2004) reported no correlation between the hearing thresholds and the TMC CE and Anyfantakis *et al*. (2017) reported a correlation between TMC and OAE CEs. However, those two studies had a smaller number of participants than the current study. A power analysis reveals that a correlation coefficient of 0.5 (the value reported above for TMC CE and the hearing thresholds) requires at least 29 data points, while Plack *et al*. (2004) had 26 points in total and only 12 from HI listeners. Here, the corresponding sample size was 89. This suggests that the Plack *et al*. (2004) study may have not had enough statistical power to detect the existing relations. Similarly, the correlation analyses in Anyfantakis *et al*. (2017) were based on just 8 data points and the lack of correlation between TMC and OAE CEs reported here is consistent with Johannesen and Lopez-Poveda (2010), who found no relation at frequencies below 4 kHz. However, the main limitation of the TMC CE estimates in the current study is their high variability, with values ranging between 0.1 (where a constraint was set) and 2.2. While this variability is consistent with Rosengard *et al*. (2005), it limits the strength of the current conclusion.

On the other hand, the current analysis suggests that NN10 estimates complement the audiogram, as NN10 provides information on OAE CEs that is not present in the hearing threshold data. While, the DPOAE method may return a CE estimate within a few minutes, many HI listeners did not produce strong enough DP responses to estimate an I/O function. However, all listeners could perform the NN task. Therefore, the NN task may be useful to assess the BM nonlinearity when DP responses cannot be measured. Moreover, while the NN task may be too time-consuming to be performed in the clinic, its low training requirement and good stability make it an interesting candidate for at-home, mobile-platform-based testing (e.g., Hyvärinen *et al*., 2019).

SUMMARY

Three methods, two behavioural (TMC and NN) and one physiological (DPOAE), of estimating peripheral compression were tested. The results were, in all cases, significantly correlated with audiometric thresholds. When the audiometric thresholds were controlled for, only the correlation between NN and DPOAE estimates remained significant. While the DPOAE task is more time-efficient than the NN task, it did not return an estimate in several individual cases. These results suggest that the NN task could serve as a complementary test of estimating peripheral compression for listeners where DPOAEs are difficult to measure in the individual listener.

ACKNOWLEDGMENTS

Parts of this study were supported by the Oticon Foundation.

REFERENCES

Anyfantakis, K., MacDonald, E. N., Epp, B., and Fereczkowski, M. (**2017**). "Comparison of objective and subjective measures of cochlear compression in normal-hearing and hearing-impaired listeners". Proc. ISAAR, **6**, 167-174.

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (**2010**). "Standard audiograms for the IEC 60118-15 measurement procedure". Trends Amplif., **14**(2), 113-120. doi: 10.1177/1084713810379609.

Fereczkowski, M. (**2015**). "Time-efficient behavioural estimates of cochlear compression". PhD thesis, Contributions to Hearing Research, **20**.

Fereczkowski, M., Jepsen M. L., Dau, T., and MacDonald E. N. (**2017**). "Investigating time-efficiency of forward masking paradigms for estimating basilar membrane input-output characteristics". PloS ONE, **12**(3), e0174776. doi: 10.1371/journal.pone.0174776.

Hyvärinen, P., Fereczkowski, M, and MacDonald, E., (**2019**). "Evaluation of a notched-noise test on a mobile phone". Proc. ISAAR, **7**, 125-132.

Johannesen, P. T., and Lopez-Poveda, E. A. (**2010**). "Correspondence between behavioral and individually optimized otoacoustic emission estimates of human cochlear input/output curves". J. Acoust. Soc. Am., **127**(6), 3602-3613. doi: 10.1121/1.3377087.

Nelson, D. A., Schroder, A. C., and Wojtczak, M. (**2001**). "A new procedure for measuring peripheral compression in normal-hearing and hearing-impaired listeners". J. Acoust. Soc. Am., **110**(4), 2045-2064. doi: 10.1121/1.1404439

Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli". J. Acoust. Soc. Am., **59**(3), 640-654 doi: 10.1121/1.380914.

Plack, C. J., Drga, V., and Lopez-Poveda, E. A. (**2004**). "Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss". J. Acoust. Soc. Am., **115**(4), 1684-1695. doi: 10.1121/1.1675812.

Rosen, S., and Baker, R. J. (**1994**). "Characterising auditory filter nonlinearity". Hearing Res., **73**(7), 231–243.

Rosengard, P. S., Oxenham, A. J., and Braida, L. D. (**2005**). "Comparing different estimates of cochlear compression in listeners with normal and impaired hearing". J. Acoust. Soc. Am., **117**(5), 3028-3041. doi: 10.1121/1.1883367.

Characterizing the speech-in-noise abilities of school-age children with a history of middle-ear diseases

SHNO KOIEK[1], JENS BO NIELSEN[2], CHRISTIAN BRANDT[1], JESPER HVASS SCHMIDT[3], AND TOBIAS NEHER[1,*]

[1] *Institute of Clinical Research, University of Southern Denmark, Odense, Denmark*

[2] *Hearing Systems, Technical University of Denmark, Lyngby, Denmark*

[3] *Department of Audiology, Odense University Hospital, Odense, Denmark*

Recently, a number of studies have indicated that recurrent or chronic middle-ear disease during early childhood may lead to long-term supra-threshold hearing deficits. The current study followed up on this by investigating differences in monaural and binaural hearing abilities in noise among school-age children with or without a history of middle-ear diseases. Groups of children aged 6-12 years with either a history of recurrent otitis media (OM) with infection or effusion or without any previous ear diseases participated. All participants had normal middle-ear function and normal audiometric hearing thresholds at the time of testing. Measurements included monaural and binaural speech reception thresholds in the presence of stationary noise or competing speech. Sensitivity to binaural phase information was also assessed. Preliminary analyses based on the data from the first 31 participants suggest that, on average, OM children have poorer thresholds in conditions with binaural or spatial differences compared to children without any previous middle-ear problems. Follow-up analyses based on a larger dataset will substantiate these initial findings and relate them to information obtainable from the OM children's medical records (e.g., age of onset or duration of conductive hearing loss).

INTRODUCTION

Otitis media (OM) is the most common reason for temporary hearing loss in children before the age of four (Bennett & Haggard, 1999). OM can produce mild-to-moderate intermittent hearing loss of up to 40 dB HL, typically in the lower frequency range (Moore *et al*., 2003). Auditory abilities are known to develop greatly during early childhood (Cameron *et al*., 2009). Consequently, fluctuating and/or asymmetrical hearing loss caused by OM during this critical period has generated much speculation about potential longer-term effects on auditory system development, language acquisition and perception.

A number of studies have suggested adverse effects of conductive hearing loss due to OM during early childhood on higher-order auditory abilities including binaural hearing (Moore *et al*. 2003; Keogh *et al*., 2005; Tomlin & Rance, 2014; Graydon *et al*., 2017). To illustrate, Tomlin and Rance (2014) used the 'Listening in Spatialized

*Corresponding author: tneher@health.sdu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Noise-Sentences Test' (LISN-S), which is an established tool for testing spatial processing abilities. They observed poorer abilities in OM children years after their hearing thresholds had returned to normal. In their case, OM history was determined based on parental reports. Graydon *et al*. (2017) also demonstrated longer-term effects of early-childhood conductive hearing loss on spatial processing abilities. In their study, OM history was determined based on the children's medical records. In contrast to Tomlin and Rance (2014) and Graydon *et al*. (2017), however, other researchers did not find evidence for long-term consequences of early conductive hearing loss on auditory development (e.g., Hartley *et al*., 2001). These inconsistent findings could be due to different aspects related to the history of middle-ear diseases (e.g., age of onset and duration of conductive hearing loss) and auditory system recovery (Keogh *et al*., 2005; Lawless *et al*., 1981; Moore *et al*., 2003).

Given the abovementioned inconsistences, the purpose of the current study was to look more closely at potential longer-term effects of early-childhood conductive hearing loss on binaural hearing abilities. We assessed binaural hearing using tone-in-noise, speech-in-noise and speech-on-speech stimuli with or without binaural differences between the competing signals. In this way, we examined the influence of early-childhood OM on different levels of auditory processing. Moreover, we will explore if these effects are associated with information obtainable from the OM children's medical records (e.g. age of onset and duration or of conductive hearing loss) in follow-up analyses based on a larger dataset. Our hypothesis was that early-childhood conductive hearing loss results in longer-term binaural hearing deficits and that this is associated with information about middle-ear status during early childhood.

MATERIALS AND METHODS

Ethical approval for the current study was obtained from the Regional Committees on Health Research Ethics for Southern Denmark.

Participants

Twenty children aged 6-12 years (mean: 10.8 years; standard deviation, SD: 1.7 years; 15 male) with a documented history of middle-ear infection or effusion ('OM group') and 11 children within the same age range (mean: 10.2 years; SD: 1.9 years; 4 male) without any previous ear diseases ('control group') participated. The inclusion criteria for all participants were normal middle-ear function (type-A tympanogram), pure-tone average (PTA) hearing thresholds calculated across 500, 1000, 2000 and 4000 Hz (PTA4) $\leq$ 20 dB HL, and normal speech, language and cognitive development at the time of testing. Fulfilment of these criteria was assessed using standard audiological measurements and parental questionnaires. Socioeconomic status is known to influence academic development. Therefore, comparability of the two groups in terms of socioeconomic status was also verified using a questionnaire. We used a custom-made questionnaire that included five questions related to the child's mother tongue, if the child was monolingual, the level of education of the child's parents and income of the child's parents. All the children who participated in the current study were monolingual, native Danish speakers with similar socioeconomic

status. In addition, children belonging to the OM group were required to have had at least three episodes of middle-ear infection or effusion (type-B tympanogram and conductive hearing loss with PTA4 >25 dB HL in each affected ear) for several months in at least one ear before the age of five. If a given child had also experienced middle-ear issues afterwards but otherwise fulfilled the inclusion criteria at the time of testing, it was still included in the study. History of middle-ear diseases was verified using the medical records from the children's otologists.

Design and procedure

The participants attended three appointments lasting 45-60 minutes each at the audiological laboratory of the University of Southern Denmark. The first visit included (1) completion of the parental questionnaires, (2) otoscopy and tympanometry, (3) standard pure-tone audiometry, (4) monaural measurements of speech reception in quiet, and (5) monaural and binaural measurements of speech reception in noise. At the second and third visit, speech reception in the presence of competing speech and sensitivity to binaural phase information was assessed. For each participant and type of measurement, a set of test and retest measurements was performed. If a given retest measurement deviated by more than 3 dB from the corresponding test measurement, another repetition was carried out. For the data analyses, the median of each set of measurements was used. The experiments were controlled via customized MATLAB scripts. The stimuli were presented via an external sound card and free field-equalized Sennheiser HDA200 headphones. All measurements were conducted in a large sound-attenuating booth.

Sensitivity to binaural phase information

Sensitivity to binaural phase information in the presence of noise was assessed using binaural masking level difference (BMLD) measurements. A 3-interval, 3-alternative, forced-choice design with a 1-up 2-down procedure was used. On each trial, all three intervals contained bandpass-filtered, 65 dB SPL Gaussian noise that was interaurally in-phase and centred at either 500 or 1000 Hz. One randomly chosen interval contained a 500 or 1000 Hz pure tone (corresponding to the centre frequency of the noise) that was either interaurally in phase ('N0S0') or π radians out of phase ('N0Sπ'). Each interval was 500 ms long and included 25 ms raised-cosine on- and offset ramps. Intervals were separated by 333 ms of silence. The starting signal-to-noise ratio (SNR) was +1 dB with additive step sizes of 8, 4 and 2 dB. After 10 reversals, a measurement was terminated, and the threshold estimated as the geometric mean of the adaptive variable at the last six reversals. The BMLD was calculated by taking the difference between corresponding N0Sπ and N0S0 thresholds. Before the actual measurements, all participants completed a training run in the N0Sπ condition with a starting SNR of +7 dB. The participants responded by pressing one of three buttons displayed on a touch screen. In case of doubt, they were encouraged to guess. A break was given after half of the measurements and whenever a child felt tired.

Speech-in-noise reception

To assess speech-in-noise abilities, 50%-correct speech reception thresholds (SRTs) were measured in two conditions. Using anechoic head-related impulse responses (Gardner & Martin, 1994), the target speech was presented from in front (0º azimuth) and stationary speech-shaped noise from the side (90º or 270º azimuth) of the listener. The stimuli were presented either binaurally ('binaural SRT') or monaurally ('monaural SRT') to the ear opposite the noise. The speech level was initially set to 68 dB SPL and then varied according to the adaptive procedure of the Danish hearing in noise test (HINT; Nielsen & Dau, 2011). The noise was presented at 65 dB SPL. The target speech consisted of the sentence material from the pediatric DAT material (Koiek *et al*., 2020). All of these sentences have a fixed, simple structure, i.e. they start with a name (Dagmar, Asta or Tine) and contain two short keywords, for example "Dagmar tænkte på en teske og en bjørn i går". The participants were instructed to repeat the two keywords per sentence. To quantify the binaural contribution to speech-in-noise reception, the binaural intelligibility level difference (BILD; Kollmeier, 1996) was calculated by subtracting the binaural SRTs from the monaural SRTs. For training purposes, the participants performed one binaural SRT measurement in quiet and one binaural SRT measurement in noise.

Speech reception with competing speech

In addition to the speech-in-noise measurements, we assessed speech reception in the presence of competing speech (speech-on-speech measurements). Using anechoic head-related impulse responses, sentences from the pediatric DAT material (see above) were presented from in front (0º azimuth) of the listeners. As interferers, two female talkers with different voice characteristics were used. These were either collocated with (0º azimuth; '$S_fV_fV_f$') or spatially separated from (±90º azimuth; '$S_fV_rV_l$') the target speech. The spatial advantage was determined by subtracting the SRTs of the $S_fV_rV_l$ condition from the SRTs of the $S_fV_fV_f$ condition. The target speech level was initially set to 62 dB SPL and then varied according to the adaptive procedure of the Danish HINT. The level of the two competing talkers was fixed at 55 dB SPL. Before the start of the measurements, the participants were instructed to pay attention to the sentence starting with a specific name (Dagmar, Asta or Tine) and to repeat the two keywords in that sentence. For training purposes, they performed an SRT measurement in the $S_fV_rV_l$ condition. Following the training, two measurements (test and retest) per condition were performed.

Statistical analyses

To examine the distributions of the collected datasets, Shapiro-Wilk's test, normal Q-Q plots and box plots were used. To verify equality of variances, Levene's test was used, showing equality for all datasets (all $p > 0.05$). To explore differences between the two groups of children, two-sample *t*-tests were applied for the normally distributed datasets. For the non-normally distributed datasets (i.e. the N0Sπ data at 500 Hz, the binaural SRTs in noise, and the spatial advantage scores), non-parametric Mann-Whitney *U* tests were applied.

RESULTS

Sensitivity to binaural phase information

The mean and SD of the N0S0, N0Sπ and BMLD scores for the 500 and 1000 Hz conditions and two groups of participants are shown in Table 1.

Condition	Control group	OM group
N0S0, 500 Hz	−6.7 ± 2.7 dB SNR	−7.4 ± 3.4 dB SNR
N0Sπ, 500 Hz	−21.0 ± 2.6 dB SNR	−19.6 ± 5.4 dB SNR
BMLD, 500 Hz	14.2 ± 2.5 dB	12.3 ± 4.2 dB
N0S0, 1000 Hz	−8.0 ± 1.6 dB SNR	−7.9 ± 2.8 dB SNR
N0Sπ, 1000 Hz	−18.3 ± 3.2 dB SNR	−16.0 ± 3.7 dB SNR
BMLD, 1000 Hz	10.3 ± 3.1 dB	8.1 ± 3.0 dB

Table 1. Mean and SD for the N0S0, N0Sπ and BMLD scores for the 500 and 1000 Hz conditions and two groups of participants.

For the OM group, mean thresholds were higher and mean BMLDs were smaller than for the controls. However, the statistical tests revealed no significant differences in terms of these outcomes (all $p > 0.05$).

Speech-in-noise reception

Figure 1 shows the monaural and binaural SRT measurements as well as the BILD scores.

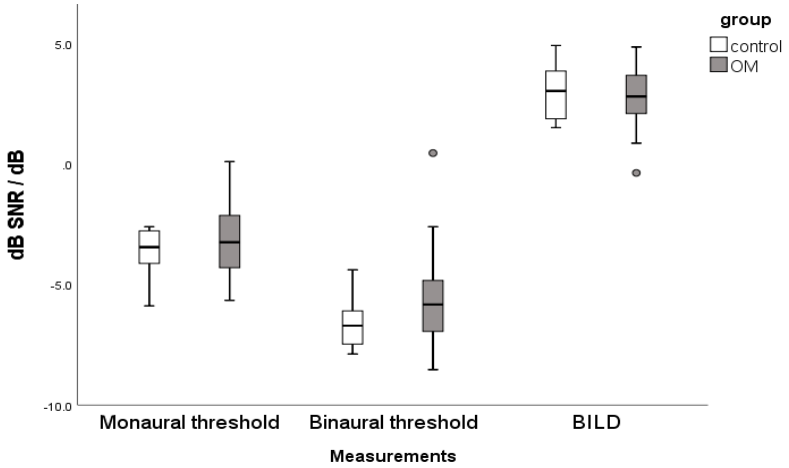


Fig. 1: Box plots showing the median, interquartile range and overall range of the monaural SRT, binaural SRT and BILD scores.

The mean binaural SRT was −5.7 dB SNR (SD: ±1.9 dB SNR) for the OM group and −6.6 dB SNR (SD: ±1.1 dB SNR) for the controls. On average, the BILD was reduced by 0.5 dB in the OM group. However, this difference was not significant ($t_{29} = 1.2$,

$p = 0.23$). Nor were there significant differences between the monaural ($t_{29} = -0.8$, $p = 0.43$) or binaural ($W = 139.0$, $p = 0.13$) SRTs of the two groups.

Speech reception with competing speech

The results of the speech-on-speech measurements are shown in Figure 2. The mean $S_fV_rV_l$, $S_fV_fV_f$ and spatial advantage scores for the control group were −4.7 dB SNR, 0.7 dB SNR and 5.4 dB (SDs: 2.3 dB SNR, 3.6 dB SNR and 3.9 dB, respectively). For the OM group, the corresponding values were −3.8 dB SNR, 2.7 dB SNR, and 6.5 dB (SDs: 2.8 dB SNR, 3.1 dB SNR and 3.0 dB, respectively). The statistical analyses revealed no significant group differences in mean $S_fV_rV_l$ ($t_{29} = -0.8$, $p = 0.41$), $S_fV_fV_f$ ($t_{29} = -1.6$, $p = 0.11$) or spatial advantage ($W = 150.0$, $p = 0.28$) scores.

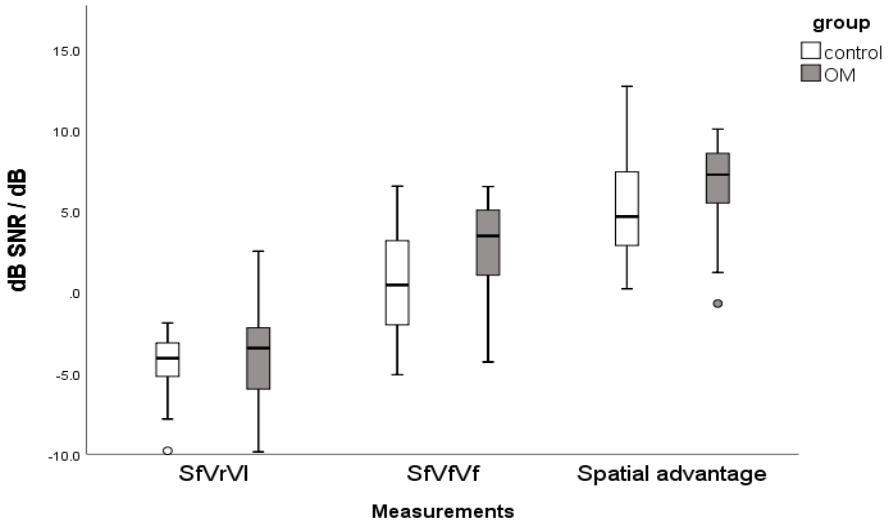


Fig. 2: Box plots showing the median, interquartile range and overall range of the $S_fV_rV_l$, $S_fV_fV_f$ and spatial advantage scores.

DISCUSSION

The aim of the current study was to characterize the longer-term effects of early-childhood conductive hearing loss on binaural hearing abilities. We used tone-in-noise, speech-in-noise and speech-on-speech stimuli with or without binaural (or spatial) differences between the competing signals to examine these effects on different levels of auditory processing. Preliminary analyses indicated higher (poorer) thresholds in conditions with binaural differences for the OM children compared to the controls. However, these results were not statistically significant.

In general, tone-in-noise detection performance tended to be poorer in the OM group compared to the control group. For instance, the BMLD scores of the controls were on average ~2 dB higher (better). The lack of a statistically significant difference between the BMLD mean scores of the OM and control groups observed here is consistent with some (Graydon *et al*., 2017) but not all (Hall *et al*. 1995b; Moore *et al*. 2003) previous studies. Regarding speech-in-noise performance, the results of the

present study suggest comparable BILD scores in children with or without a history of conductive hearing loss. Regarding the speech-on-speech measurements, the OM children required an SNR that was ~1 dB higher than that of the controls in the $S_fV_rV_l$ condition. However, the mean spatial advantage was similar for the two groups. This is in contrast to the findings of Tomlin and co-workers (Tomlin & Rance, 2014; Graydon *et al*., 2017) who observed reduced spatial advantage in OM children, even when hearing thresholds had returned to normal. One explanation for the inconsistent findings could be differences across studies in terms of OM history (e.g., age of onset or duration of middle-ear disease). For instance, Werker and Tees (2005) suggested that the first two years of life are the most critical for speech and language development. Consequently, auditory deprivation during this period may influence development the most. Tomlin and Rance (2014) also provided support for the idea that the age of onset influences the spatial processing abilities of OM children. They observed a significant correlation between the duration of OM and speech reception in the presence of two spatially separated speech interferers. Furthermore, the time interval since the occurrence of the last OM episode could play a role. Graydon *et al*. (2017) suggested that longer time intervals after the last OM episode in the case of their participants compared to those of Hall *et al*. (1995b) could be an explanation for why they did not observe an influence of OM on BMLD, whereas Hall *et al*. did. Hogan *et al*. (1996) also indicated that reduced BMLD scores in children with a history of OM improve over time as hearing thresholds return to normal. In the case of our study, the mean age of the children was 10.8 years. Thus, there may have been a rather long time interval between the last OM episode and the point in time when the measurements reported here were performed.

The data presented here constitute preliminary results from an ongoing study that eventually will include more participants. A larger sample size may well change the results of the statistical analyses reported here. Future work will also relate the psychoacoustic and speech perception measurements to information obtainable from the OM children's medical records. In this manner, it will be possible to investigate the influence of factors such as age of onset, time interval since the last OM episode and duration of early-childhood conductive hearing loss.

REFERENCES

Bennett, K. E., and Haggard, M. P. (**1999**). "Behaviour and cognitive outcomes from middle ear disease," Arch. Dis. Child., **80**, 28-35.

Cameron, S., Brown, D., Keith, R., Martin, J., Watson, C., and Dillon, H. (**2009**). "Development of the North American Listening in Spatialized Noise-Sentences Test (NA LISN-S): Sentence Equivalence, Normative Data, and Test-Retest Reliability Studies," J. Am. Acad. Audiol., **20**(2), 128-146. doi: doi.org/10.3766/jaaa.20.2.6.

Gardner, B., & Martin, K. (**1994**). "HRTF measurements of a KEMAR dummy-head microphone," Massachusetts Institute of Technology Media Lab Perceptual Computing, Cambridge, MA.

Graydon, K., Rance, G., Dowell, R., & Van Dun, B. (**2017**). "Consequences of early conductive hearing loss on long-term binaural processing," Ear Hear., **38**(5), 621-627. doi: 10.1097/AUD.0000000000000431.

Hall, J. W. 3rd, Grose, J. H., Pillsbury, H. C. (**1995**b). "Long-term effects of chronic otitis media on binaural hearing in children," Arch. Otolaryngol. Head Neck Surg., **121**, 847-852. doi: 10.1001/archotol.1995.01890080017003.

Hartley, D. E. H., Adams, C. V., Hogan, S. C., and Moore, D. R. (**2001**). "Temporal resolution and language abilities in children with a prospectively studied history of otitis media with effusion (OME)," Abs. Assoc. Res. Otolaryngol., **24**, 173.

Hoffman-Lawless, K., Keith, R. W., and Cotton, R. T. (**1981**). "Auditory processing abilities in children with previous middle ear effusion," Ann Otol Rhinol., **90**, 543-545. doi:10.1177/000348948109000606.

Hogan, S. C., Meyer, S.E. and Moore, D.R. (**1996**). "Binaural unmasking returns to normal in teenagers who had otitis media in infancy," Audiol. Neuro-Otol., **1**, 104-111. doi: doi.org/10.1159/000259189.

Keogh, T., Kei, J., Driscoll, C., Cahill, L., Hoffmann, A., Wilce, E., and Marinac, J. (**2005**). "Measuring the ability of school children with a history of otitis media to understand everyday speech," J. Am. Acad. Audiol., **16**(5), 301-311. doi: doi.org/10.3766/jaaa.16.5.5.

Koiek, S., Nielsen, J. B., Kjærbæk, L., Baltzer, M. G., and Neher, T. (**2020**). "Development of a Danish test material for assessing speech-in-noise reception in school-age children," Proc. ISAAR, **7**, 413-420.

Kollmeier, B., (**1996**). "Computer-controlled speech audiometric techniques for the assessment of hearing loss and the evaluation of hearing aids," *Psychoacoustics, Speech and Hearing Aids*. Singapore: World Scientific, 57-68.

Moore, D.R., Hartley, D.E.H., and Hogan, S.C.M. (**2003**). "Effects of otitis media with effusion on central auditory function," Int. J. Pediatr. Otorhinolaryngol., **67** (Suppl. 1) S63-S67. doi: doi.org/10.1016/j.ijporl.2003.08.015.

Nielsen, J. B., & Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol., **50**(3), 202-208, doi: doi.org/10.3109/14992027.2010.524254.

Tomlin, D., & Rance, G. (**2014**). "Long-term hearing deficits after childhood middle ear disease," Ear Hear., **35**(6), e233-e242, doi: 10.1097/AUD.0000000000000065.

Werker, J. F., and Tees, R. C. (**2005**). "Speech perception as a window for understanding plasticity and commitment in language systems of the brain," Dev. Psychobiol., **46**, 233-251. doi: doi.org/10.1002/dev.20060.

Analysis of a forward masking paradigm proposed to estimate cochlear compression using an auditory nerve model and signal detection theory

JENS THUREN LINDAHL, GERARD ENCINA-LLAMAS* AND BASTIAN EPP
Hearing Systems section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark

The healthy human auditory system has a large dynamic range. An "active mechanism", presumably due to the electromotility of the outer hair cells in the cochlea, leads to level-dependent amplification of basilar membrane (BM) vibration and a compressive BM input/output function. Different methods for estimating this compressive function based on behavioural forward masking have been suggested. These methods are based on the assumption that BM processing can be isolated from the response of the overall system and that the forward masking onto the probe is different for on- and off-frequency maskers. In the present study, a computational model of the auditory nerve (AN) in combination with methods from signal detection theory was used to test these assumptions. The simulated AN response was quantified in terms of rate and synchrony for different AN fibre types. Contribution of different tonotopic regions to overall sensitivity to the stimuli were analysed. The results show that on- and off-frequency maskers produce similar forward masking onto the probe. The simulation results suggest that the estimate of compression based on the behavioural experiment cannot be derived from sensitivity at the level of the AN but requires additional contributions, consistent with physiological studies.

INTRODUCTION

The healthy mammalian auditory system is remarkable in its ability to handle a large dynamic range of incoming sounds while having high sensitivity to low-intensity sounds. An underlying "active mechanism", presumably due to the electromotility of the outer hair cells (OHC) in the cochlea, leads to level-dependent amplification of basilar membrane (BM) vibrations and a compressive input/output (I/O) function over a narrow BM region (Ruggero *et al.*, 1997). Damage of OHCs leads to a hearing impairment (i.e., elevation of hearing threshold). Hence, compression is a proxy for the state of the OHCs and therefore for the health of the inner ear. No direct measurement of their state is currently possible in humans. Various methods have been developed to estimate BM compressive function based on behavioural forward masking paradigms (e.g., Nelson *et al.*, 2001). Important assumptions underlying these behavioural methods are, among others, that contributions of

*Corresponding author: encina@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

cochlear processing can be isolated from the behavioural response (the overall system), that sensitivity is dominated by the information at the tonotopic place corresponding to the probe frequency and that tonal maskers presented at on- and off-frequencies (near or away of the characteristic place of the tonal probe, respectively) produce different forward masking effect onto the probe (Jepsen and Dau, 2011). In the present study, a computational model of the auditory nerve (AN) in combination with methods from signal detection theory (SDT) was used to test the assumption that the "BM I/O function" derived using behavioural measures reflects the compressive growth of the BM at the characteristic place (on-frequency) corresponding to the probe frequency. The following assumptions were made: 1) Perceptual outcome measures are formed using the information present at the level of the AN in the form of spike rate and synchrony. 2) The information needed to resolve the task behaviourally can be encoded anywhere in the AN. 3) All three types of AN fibres (low, medium and high spontaneous rate, SR) can contribute to the perceptual outcome measure. 4) The information extracted in the behavioural outcome must be encoded at the level of the AN, as this is a mandatory stage.

METHODS

The experimental paradigm used in a behavioural study (Jepsen and Dau, 2011) was used to simulate the response of the AN. The AN model of Zilany *et al.* (2014) was used, implemented as described by Encina-Llamas *et al.* (2019).

Experimental paradigm

The stimulus was a tonal masker with a duration of 200 ms (5-ms rise/fall cosine ramps), a gap of varying length and a probe tone with a duration of 20 ms (10-ms rise/fall Hann window ramps with no steady portion) with a frequency of either 1000 or 4000 Hz. After computing the model threshold for each frequency, the probe level was fixed at 10 dB sensation level (SL). Thresholds in the AN model were calculated for each characteristic frequency (CF) comparing driven rate with SR using a non-parametric permutation test for equality of the means with a significance level of $\alpha = 1\%$. The masker was presented either at the probe frequency (on-frequency) or at a frequency 0.6 times that of the probe (off-frequency). The masker was presented at levels from 10 to 110 dB SPL in steps of 10 dB. The gap was set to 2, 5 and 10 to 100 ms in steps of 10 ms. First a block with the masker only was presented to the model followed by a block with the same masker plus the probe (see Fig. 1 top).

Two listeners from Jepsen and Dau (2011) were simulated: a normal-hearing (NH) listener represented by the averaged NH audiogram; and one mildly hearing-impaired (HI, listener HI01) with thresholds of 20 and 45 dB hearing level (HL) at 1 and 4 kHz, respectively. Hearing impairment, assuming a combination of ⅔ of OHC and ⅓ of inner hair cell (IHC) dysfunction, was implemented in the model by adjusting the parameters that control OHC gain (*cohc*) and IHC transduction (*cihc*).

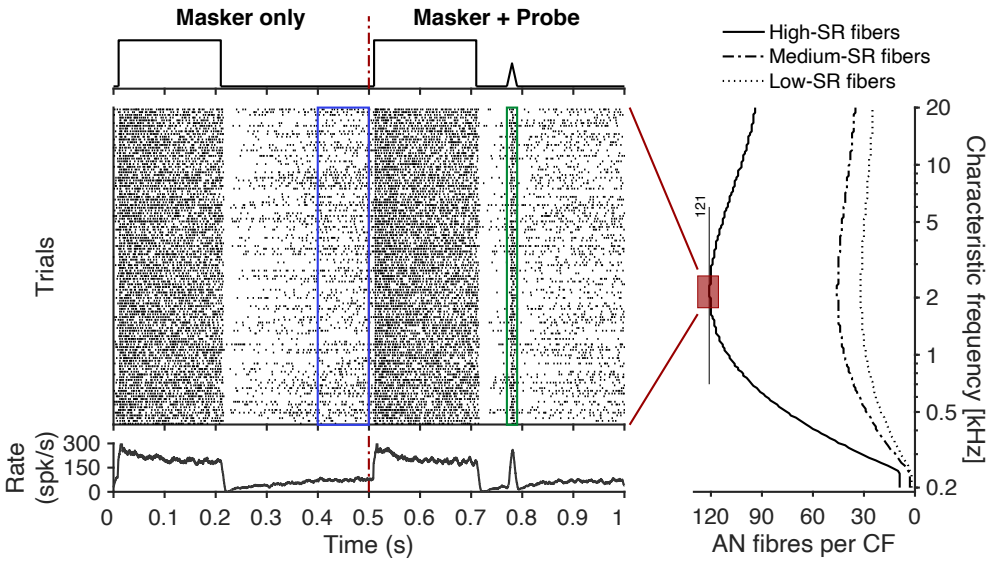


Fig. 1: Top: Stimulus paradigm. The masker was presented at the probe frequency (on-frequency) or at 0.6 times the probe frequency (off-frequency). The probe was presented following the second masker with a temporal gap of varying length. Middle: Trials of simulated raster plots for one CF. The number of simulated AN fibres per CF of each type differed across CFs (right panel). The rectangles indicate the analysis windows for the estimation of the spontaneous activity (blue) and the probe response (green). Bottom: Resulting post-stimulus time histogram (PSTH) derived from the response of all simulated nerve fibres at this CF; showing the onset response, adaptation, the reduction in rate following masker offset and the response to the probe.

Simulation and analysis using signal detection theory

The AN response was simulated for neurons with CFs from 200 Hz to 20 kHz, discretised into 200 CFs (cochlear segments) equally spaced along a logarithmic axis. The receptor potential of a single IHC was simulated for each CF. The number and type of nerve fibres at each CF followed the distribution used by Encina-Llamas *et al.* (2019), who considered a total number of 32000 AN fibres, based on physiological data from human temporal bones. At each CF and for each independently computed AN fibre, the resulting spike trains were stored and used to compute spike rate r_L and spike synchrony $\rho(f)$. These metrics were computed over two different time windows in the simulation: one centred on spontaneous activity (blue rectangle in Fig. 1 middle) and the other centred on the probe response (green rectangle in Fig. 1 middle). Spike rate (Eq. 1) was calculated as the number of spikes N_{spikes} normalised by the length of the analysis window L. Spike synchrony ρ (Eq. 2) was calculated as the vector sum of spike times projected onto a phasor rotating on the unit circle in the complex plane at the probe frequency f.

$$r_L = \frac{N_{\text{spikes}}}{L} \qquad \text{(Eq. 1)} \qquad \rho(f) = \left| \frac{1}{N} \sum_{n=1}^{N} e^{i2\pi f t_n} \right| \qquad \text{(Eq. 2)}$$

For each CF and fibre type, distributions of rate and synchrony were obtained for each window of analysis. The resulting distributions were used to calculate the sensitivity index d' using their mean and variance values, as shown in Eq. 3 (Jones, 2016).

$$d' = \frac{\mu_{\text{probe}} - \mu_{\text{SR}}}{\sqrt{\frac{1}{2}(\sigma^2_{\text{probe}} + \sigma^2_{\text{SR}})}} \qquad \text{(Eq. 3)} \qquad d'_{\text{comb}} = \left[\sum_{m=1}^{2} \sum_{k=1}^{\text{CF}} \sum_{l=1}^{3} {}^{\text{SR}_l}_{\text{CF}_k} d'^{\,2}_{\text{m}} \right]^{\frac{1}{2}} \qquad \text{(Eq. 4)}$$

All AN fibres were treated as independent sources of information. Information in the form of d' was combined across fibre type, across CF and across metric (rate and synchrony). When combining across CF, three different integration bandwidths were considered: 1) a narrow range of CFs equivalent to 1/6-octave band, (d'_{narrow}); 2) a bandwidth centred at the probe CF to cover the region where direct measurements of BM motion suggest a compressive growth (Ruggero *et al.*, 1997) (lower bound 1/2-octave below CF of the probe and upper bound 1/3-octave above the CF equal to the probe frequency, $\left(d'_{\text{compr}}\right)$); and 3) a range of 2-octaves of simulated CFs centred at the probe frequency $\left(d'_{\text{wide}}\right)$. The final combined sensitivity d'_{comb} for each group of parameters was calculated as shown in Eq. 4, with $\text{SR} = 1, 2, 3$ denoting the AN fibre type (i.e., high-, medium- and low-SR fibres, respectively; assuming the distribution based on the physiology of the cat), CF_k denoting the CF index, and $\text{m} = 1, 2$ denoting the two derived metrics (rate and synchrony, respectively).

RESULTS

Figure 2 shows heatmaps of rate (panel A) and sensitivity (panel B) evaluated in the analysis window of the probe (green vertical rectangle in Fig. 1) as function of CF and masker level for the NH simulations, and the combined sensitivity d'_{comb} as a function of masker level (panel C) for a fixed gap length of 30 ms. For simplicity, only high-SR fibres were considered in this example.

At low masker levels, the rates were highest in a narrow CF region near on-frequency (within the band of 1/6-octave shown by the green dashed lines) and decreased with increasing masker level. The region of reduced rate after masker offset increased with increasing masker level. The off-frequency masker clearly showed an asymmetric spread of reduced rate towards higher CFs. Sensitivity also reduced with increasing masker level. The combined sensitivity strongly reduced over a range of masker levels of about 10-15 dB. Such drops in combined sensitivity occurred at low masker levels for the on-frequency masker and at much higher levels for the off-frequency masker. The combined sensitivity did not depend on the bandwidths of CF integration.

Panels A and D in Fig. 3 show sensitivity as a function of gap length for a fixed masker level of 60 dB SPL for the NH and HI listeners, respectively. Panels B and E show simulated temporal masking curves (TMC). Panels C and F show the corresponding

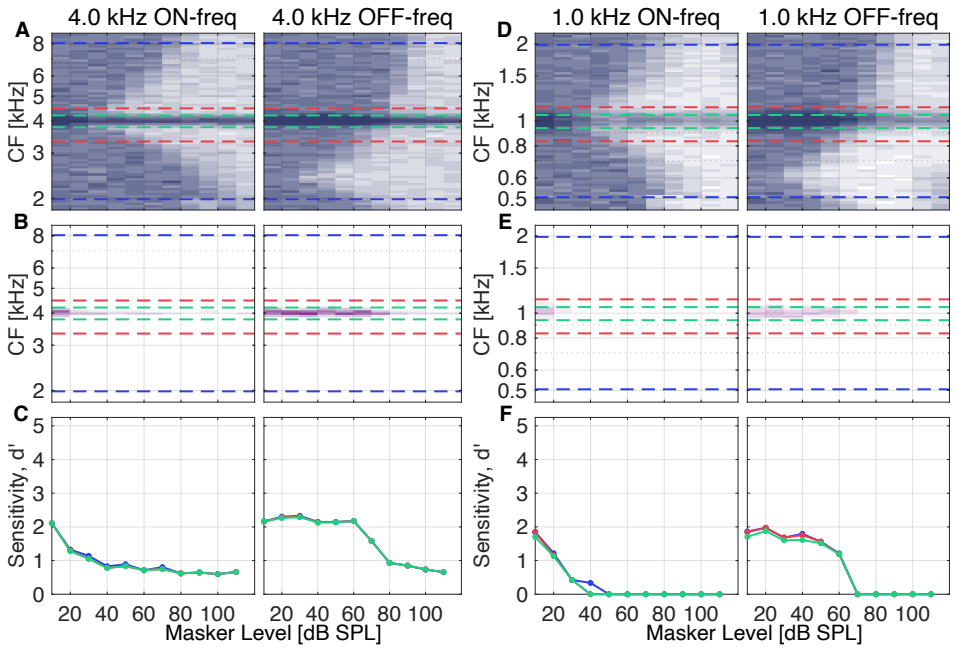


Fig. 2: Simulation results for the NH listener as a function of masker level for a fixed gap length of 30 ms. Panels A-C show results for a probe frequency of 4 kHz and panels D-F for a probe frequency of 1 kHz. Panels A and D and B and E show heatmaps of rate and sensitivity, respectively, as a function of CF and masker level. Panels C and F show combined sensitivity for different CF integration bands (green: d'_{narrow}, red: d'_{compr} and blue: d'_{wide}). Results show simulations considering only high-SR fibres.

derived "BM I/O" functions for the NH and the HI listener, respectively. The simulated "BM I/O" functions did not match the experimental data. In both cases, close to linear growth was predicted, in contrast to the behavioural data(black circles), because the TMCs for the on- and off-frequency maskers (panels B and E in Fig. 3) grew close to parallel. This finding was independent of the bandwidth of integration used to calculate the combined sensitivity. For the HI listener, the on- and off-frequency simulated TMCs (panel E) were shifted towards higher masker levels, also growing close to parallel, and more closely spaced. The broadening of the AN excitation for the HI listener (panel B) led to different results in the calculated combined sensitivity d'_{comb} for the different bandwidths of integration, resulting in higher values with increasing bandwidth of integration.

DISCUSSION

The simulation results show the reduction of AN response on SR activity across CFs due to the presence of the maskers (see panels A and D in Fig. 2). In agreement with direct BM measures (Ruggero *et al.*, 1997), the spread of such reduction is asymmetric towards higher CFs. For high masker levels, the response of high-SR fibres saturated

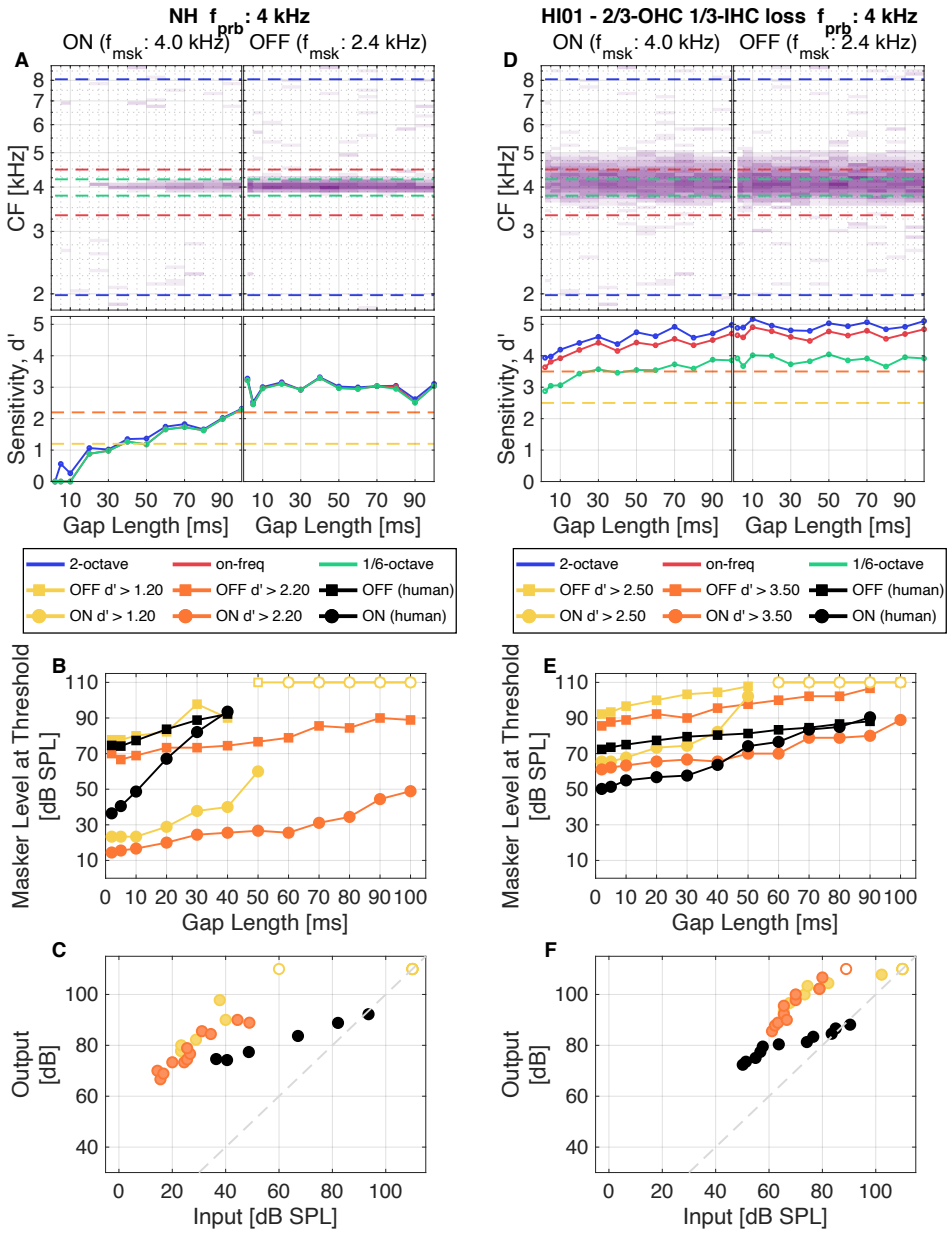


Fig. 3: Panels A and D show combined sensitivity as a function of gap length for NH and HI01, respectively, using a masker level of 60 dB SPL and a probe frequency of 4 kHz, with on- and off-frequency maskers. Dashed lines indicate integration bandwidths. Panels B and E show corresponding simulated TMC curves for two sensitivity criteria. Panels C and F show corresponding "BM I/O function" derived from the simulated TMC curves. In panels B, C, E and F, open symbols show ceiling responses using the maximum masker level of 110 dB SPL, and black solid lines and symbols show human results from Jepsen and Dau (2011), for comparison. The grey-dashed line indicates linear growth.

and hence the derived rate and sensitivity showed a plateau. The saturated rate of high-SR fibres limits the masking effect, because increasing the physical level of the masker does not lead to a change in the neuronal representation of the masker for this type of AN fibre. For the NH simulation, the results for the d'_{comb} did not show an effect of the bandwidth of integration (see panels C and F in Fig. 2), due to the narrow on-CF excitation produced by the low intensity probe. For the HI listener, the d'_{comb} grew with increasing integration bandwidth (see panel D in Fig. 3), consistent with the effect on the AN tuning curves (i.e., less sharply tuned tips) after reduced local (on-frequency) gain due to OHC damage (Liberman and Dodds, 1984). The intensity of the probe must be then increased to maintain excitation at 10 dB SL, leading to a larger populations of AN neurons encoding probe information and potentially contributing to detection. Even for the mild hearing impairment, the probe representation exceeded the definition of on-frequency bandwidth, d'_{compr}. In all cases, d'_{comb} was dominated by the high-SR fibres, as the probe was presented at 10 dB SL. The inclusion of synchrony had little effect on the results, probably due to the short duration of the probe.

The simulated TMCs did not match the behaviourally measured TMCs of Jepsen and Dau (2011). In contrast to the behavioural data, the simulated on- and off-frequency masking curves grew close to parallel, leading to almost linear growth in the derived "BM I/O functions". This implies that the effect of the tonal masker on the neuronal activity post-masker (and therefore on the probe) is essentially the same regardless of the frequency of the masker (on- vs off-frequency). Direct physiological recordings in AN neurons demonstrated that this is true particularly for the case of high-SR fibres (Yates *et al.*, 1990). Because the probe is kept at a very low stimulus level (10 dB SL), the encoding of the probe is dominated by high-SR fibres, which are not affected by the BM nonlinearity, and therefore cannot reflect the compressive growth. The AN model of Zilany *et al.* (2014) may not fully capture the effect of forward masking at the level of the AN, as this has been improved in a newer version of the model. Nevertheless, the model simulations seemed to generally agree with physiological evidence indicating that there is not sufficient forward masking at the level of the AN to account for the behavioural data, and that additional processing, which may include inhibitory mechanisms at the level of the inferior colliculus (IC), are needed to account for the behavioural responses (Nelson *et al.*, 2009). Hence, cochlear compression estimates measured using psychoacoustics might be strongly influenced by processing at these higher stages, compromising their power to infer only BM responses.

CONCLUSIONS

A model of the AN combined with SDT methods based on the three types of AN fibres, multiple CFs and rate and synchrony cues could not explain the behaviourally derived "BM I/O functions". Encoding of the probe is dominated by high-SR fibres due to the use of low intensity probes. BM nonlinearities are not reflected in the high-SR fibres (Yates *et al.*, 1990), which are limited by their saturation at high intensities and their SR at lower stimulus levels. This suggests that behaviourally derived "BM I/O

functions" with the current stimulation paradigm do not reflect the compressive growth of the BM, even though this occurs previous to the AN. Furthermore, physiological data suggest that contributions from higher neural stages are needed to account for the behavioural data (Nelson *et al.*, 2009). And even more, in the case of mild threshold elevation due to OHC damage, CFs outside the narrow frequency range where the healthy BM velocity was found to grow compressively with input level carry probe information, and may contribute to the perceptual outcome.

ACKNOWLEDGEMENTS

This work was funded by the Novo Nordisk Foundation grant NNF17OC0027872.

REFERENCES

Encina-Llamas, G., Harte, J. M., Dau, T., Shinn-Cunningham, B., and Epp, B. **(2019)**. "Investigating the effect of cochlear synaptopathy on envelope following responses using a model of the auditory nerve", J. Assoc. Res. Otolaryngol., **20** (4),363–382. doi: 10.1007/s10162-019-00721-7.

Jepsen, M. L., and Dau, T. **(2011)**. "Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss", J. Acoust. Soc. Am., **129** (1),262–281. doi: 10.1121/1.3518768.

Jones, P. R. **(2016)**. "A tutorial on cue combination and Signal Detection Theory: Using changes in sensitivity to evaluate how observers integrate sensory information", J. Math. Psychol., **73**,117–139. doi: 10.1016/j.jmp.2016.04.006.

Liberman, M. C., and Dodds, L. W. **(1984)** "Single-neuron labeling and chronic cochlear pathology. III. Stereocilia damage and alterations of threshold tuning curves", Hear. Res., **16**(1),55–74. doi: 10.1016/0378-5955(84)90025-X.

Nelson,D. A., Schroder, A. C., and Wojtczak, M. **(2001)** "A new procedure for measuring peripheral compression in normal-hearing and hearing-impaired listeners", J. Acoust. Soc. Am., **110**(4),2045–2064, 2001. doi: 10.1121/1.1404439.

Nelson, P. C., Smith, Z. M., and Young, E. D. **(2009)** "Wide-dynamic-range forward suppression in marmoset inferior colliculus neurons is generated centrally and accounts for perceptual masking", J. Neurosci., **29**(8),2553–2562. doi: 10.1523/JNEUROSCI.5359-08.2009.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. **(1997)** "Basilar-membrane responses to tones at the base of the chinchilla cochlea", J. Acoust. Soc. Am., **101**(4),2151–2163. doi: 10.1121/1.418265.

Yates, G. K., Winter, I. M., and Robertson, D. **(1990)** "Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range", Hear. Res., **45**(3),203 – 219. doi: 10.1016/0378-5955(90)90121-5.

Zilany, M. S. A., Bruce, I. C., and Carney, L. H. **(2014)** "Updated parameters and expanded simulation options for a model of the auditory periphery", J. Acoust. Soc. Am., **135**(1),283–286. doi: 10.1121/1.4837815.

Physiological correlates of masking release

HYOJIN KIM[1,*] AND BASTIAN EPP[1]

[1] *Hearing Systems Section, Technical University of Denmark, DK-2800 Lyngby, Denmark*

Masking release is one example of auditory object segregation where the masked threshold of a target sound decreases in the presence of beneficial cues. Two such cues are comodulation and interaural phase disparity (IPD) underlying the phenomena of comodulation masking release (CMR) and binaural masking level difference (BMLD) respectively. While the effect of these cues have been shown in behavioral studies, little is known about the underlying physiological mechanisms of masking release. In this study, we postulated an "internal signal-to-noise ratio (iSNR)" that reflects neuronal representation of a masked tone. As the proxy for iSNR, we investigated the applicability of late auditory evoked potentials (LAEPs). We added an onset asynchrony cue with comodulation and IPD cues. Results showed that onset asynchrony had a negative effect on CMR while it did not affect BMLD. The P2 component of the vertex LAEPs was suggested to be an objective measure of iSNR. This will provide us information about whether temporal contexts affect the neuronal representation of CMR and BMLD at the level of the auditory cortex.

INTRODUCTION

Our auditory system has the remarkable ability of sound object segregation. In a simple case, where S is a tone and M is a masker, the task for our auditory system can be defined as the detection of a masked-tone by separating the tone (S) from the masker (M). According to the power-spectrum model of masking (e.g. Fletcher, 1940), the detection threshold is correlated to a certain constant signal-to-noise ratio ($k = S/M$) that is based on the physical intensity of the stimulus. This model cannot explain a masking release where the detection threshold decreases by adding cues to the stimulus with identical power spectra. To account for the effect of beneficial cues without changes in the power spectrum, this model can be reformulated in terms of an internal signal-to-noise ratio (iSNR) at the cortex level. As a masked tone ($S+M$) is transmitted through the auditory system, it will have a neuronal representation of S_i + M_i at the cortex level.

$$S + M \rightarrow S_i + M_i \rightarrow \textit{behavioral measures}$$

Assuming that there exists a mapping between internal representations ($S_i + M_i$) and behavioral measures (e.g. audibility measure, masked thresholds), it is possible to

*Corresponding author: hykim@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

predict the behavioral outcome from the internal representation. To achieve this, an objective measure of iSNR is required. Masking release has been investigated on various levels. At the single-cell level, correlates of masking release induced by co-modulation (comodulation masking release, CMR) was found in the cochlear nucleus (CN) (Pressnitzer *et al.*, 2001). The neural responses indicated the suppression of a comodulated masker and the enhancement of a tone at the first neural stage after the cochlea. Masking release induced by interaural phase difference (binaural masking level difference, BMLD) was suggested to be processed at the inferior colliculus (IC) (Jiang *et al.*, 1997) that is located upstream of the CN. At the cortical level, Epp et al. (2013) investigated the neural representation of a masked tone with comodulation and IPD cues using EEG. This result suggested that the neuronal representations of these cues are combined at the level of the auditory cortex, supporting the idea of bottom up processing and a superposition of masking release. They found that the late auditory evoked potential P2 can be an objective measure of the audibility of the stimulus (iSNR). The amplitude of P2 was correlated with the individual level above masked threshold rather than to the physical signal-to-noise ratio (SNR) of the stimulus (Epp *et al.*, 2013). Contrary to this idea, CMR was found to be reduced by the streaming effect, which is assumed to be a higher-level process (Dau *et al.*, 2005; Grose *et al.*, 2009). This has only been shown by psychoacoustical experiments measuring masked thresholds. Only few studies investigated with physiological experiments (e.g. the mismatch negativity (MMN)), however, no study has found a neural correlate of the streaming effect on masking release (Verhey *et al.*, 2012). In addition, there is no study regarding the effect of streaming on BMLD. Hence, it remains unclear how streaming affects combined CMR and BMLD (the streaming effect on masking release). We postulate that the streaming effect is related to temporal information processed at the level of the CN. As the IPD cue is likely processed at the level of the IC, we hypothesize that BMLD will not be affected by streaming (Figure 1). In this study, we investigate whether the streaming effect is: i) a result of bottom up processing; ii) merely additional neuronal processing after summation of neuronal representation of CMR and BMLD.

As an extension of the study by Epp et al. (2013), we added onset asynchrony as a grouping cue to induce the streaming effect. We first measured masked thresholds for each condition, $TH_m([condition])$. The level of the tone was set to the same level above the individual masked threshold (e.g. 15dB + $TH_m([condition])$) with a fixed level of the noise. If case i) is true, we hypothesize that the P2 amplitude will be the same for all conditions where the level of the tone was set to the same level above masked threshold in each condition (Figure 2a). The same P2 amplitude indicates the same iSNR for all conditions despite different SNRs of the stimulus for each condition. If the second case ii) is true, however, the P2 amplitude measured at masked thresholds will be higher as the detrimental effect of the streaming effect would require a higher level of neuronal representation of CMR to achieve the same audibility (Figure 2b).

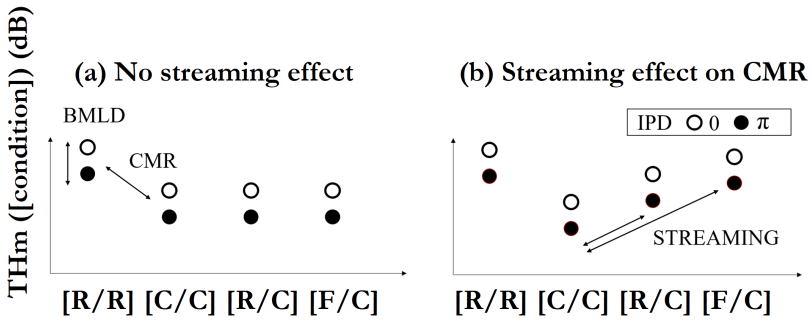


Fig. 1: Hypotheses about psychoacoustic experiment results. (a) CMR and BMLD are independent of the temporal context of the masker bands. (b) CMR are reduced as flanking bands (FBs) and signal-centred bands (SCB) are grouped separately by the onset asynchrony cue (the streaming effect) while there is no streaming effect on BMLD

METHODS

The same stimuli were used for the behavioural and the electrophysiological experiment (Figure 3). Masker conditions were designed based on Grose *et al.* (2009). For each condition, different first masker (Masker 1) and second masker (Masker 2) intervals were used. Four different masker conditions were used (Figure 4): (a) [R/R]: both masker intervals had uncorrelated envelope fluctuations; (b) [C/C]: both masker intervals had comodulated envelopes; (c) [R/C]: the first masker interval had uncorrelated envelope fluctuations and the second masker interval had comodulated envelope fluctuations; (d) [F/C]: the first masker interval had comodulated flanking bands (FBs) and the second masker interval had comodulated FBs and signal-centred bands (SCB). The masker consisted of five narrow-band noises with a width of 20 Hz, centered at 460 Hz, 580 Hz, 820 Hz, 940 Hz (FBs) and at 700 Hz (SCB). The bandwidth and center frequency of each noise band were chosen to maximize CMR (Grose *et al.*, 2009). The total duration of the signal was 700 ms including 20 ms on- and offset ramps. The first masker interval was gated on for 500 ms followed by a second masker of duration 200 ms. The target tone had a frequency of 700 Hz and was presented with the second masker interval. The target tone had an interaural phase difference of either 0 or π. The stimuli were digitally generated with a sampling frequency of 48 kHz.

The stimuli were presented using ER-2 headphones. For the psychoacoustical experiment, a modular framwork for running psychoacoustic experiments and computational perception models (AFC) software package for MATLAB was used (Ewert, 2013). An adaptive and three-alternative forced choice procedure was used with a one-up, two-down rule (Levitt, 1971). The listener was asked to choose the interval with the tone. During the EEG experiment, the stimuli were presented while participants were

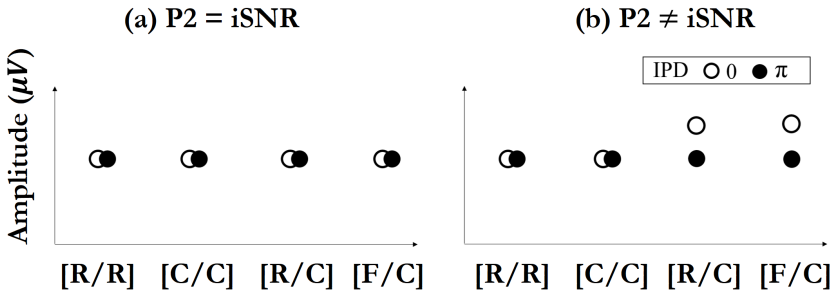


Fig. 2: Hypotheses about electrophysiological experiment results. (a) The P2 component of the LAEP reflects iSNR. (b) The P2 component of the LAEP cannot reflect iSNR and only reflect the summed neuronal representation of CMR and BMLD.

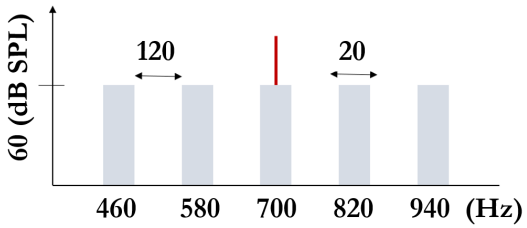


Fig. 3: Spectra of the stimulus. A target tone (700 Hz) was presented with a set of narrow-band masker bands: One signal centered band (SCB) and four flanking bands (FBs). Thresholds were measured individually and used to adjust the levels for the EEG experiment to set equal levels above masked threshold.

watching a silent movie. Late auditory evoked potentials (LAEPs) were measured using a 144 channel EEG amplifier (g.Tec HiAmp research) with active electrodes. A conductive gel was used to reduce the impedance of electrodes. Electrodes with an impedance higher than 10 kΩ were excluded from the analysis. The reference electrode was placed close to the mastoid (P8) and the region of interest was the central position (Cz). The data analysis was performed using FieldTrip (Oostenveld *et al.*, 2011). The EEG data were partitioned into epochs from -300 to 1200 ms relative to the onset of the masker. Each epoch was low pass filtered with a cut-off frequency of 20 Hz. Detrending, base line correction and weighted averaging (Riedel *et al.*, 2001) were applied to increase the signal-to-noise ratio. Trials containing signals exceeding 100 μV in any channel were rejected as artifacts.

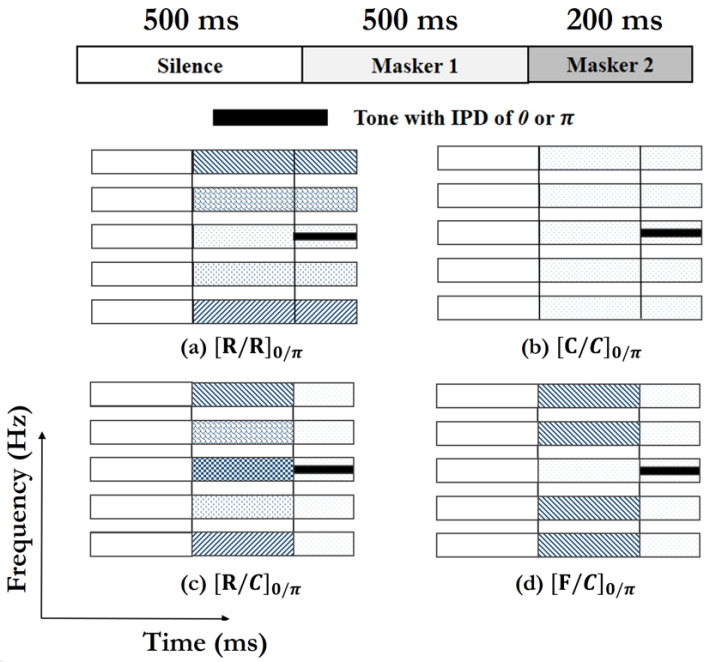


Fig. 4: Schematic spectrograms of the stimulus. Each block represents a noise band. The thick black line represents a tone. Tone is presented either with IPD of 0 or π. RAN: All noise bands have random envelope fluctuations. COM: All noise bands have the same envelope fluctuations (comodulated). FCOM: Only the flanking bands are comodulated.

RESULTS AND DISCUSSION

Comodulation masking release, CMR

$$CMR_* = THm([R/R]) - THm([*/C]) \qquad \text{(Eq. 1)}$$

Here, * stands for one of three masker types (RAN, FCOM, COM).

Figure 5 shows the results of the psychoacoustic experiment. In the diotic condition (Figure 5, circle), CMR_C was observed for all listeners. The CMR_R and CMR_F are smaller than CMR_C. In the streaming conditions ([R/C], [FC/C]), we postulate that the auditory system grouped masker bands into separate objects due to their uncorrelated intensity fluctuations during the first masker (Masker 1). Reduced CMR indicates that the comodulation cue is not beneficial when masker bands were separated before the comodulation cue is provided.

Binaural masking level difference, BMLD

$$BMLD_{masker} = THm([condition]_0) - THm([condition])_\pi \qquad \text{(Eq. 2)}$$

When an IPD of π was introduced, BMLD was observed for all conditions (Figure 5, cross). BMLD was almost constant, except for the [F/C] condition. For participant A - C, $BMLD_{F/C}$ was larger than in the other conditions. This might suggest that there might be individual differences in the effects of temporal contexts to BMLD. Therefore, it remains to be seen which pattern of results is dominant over a larger cohort of listeners.

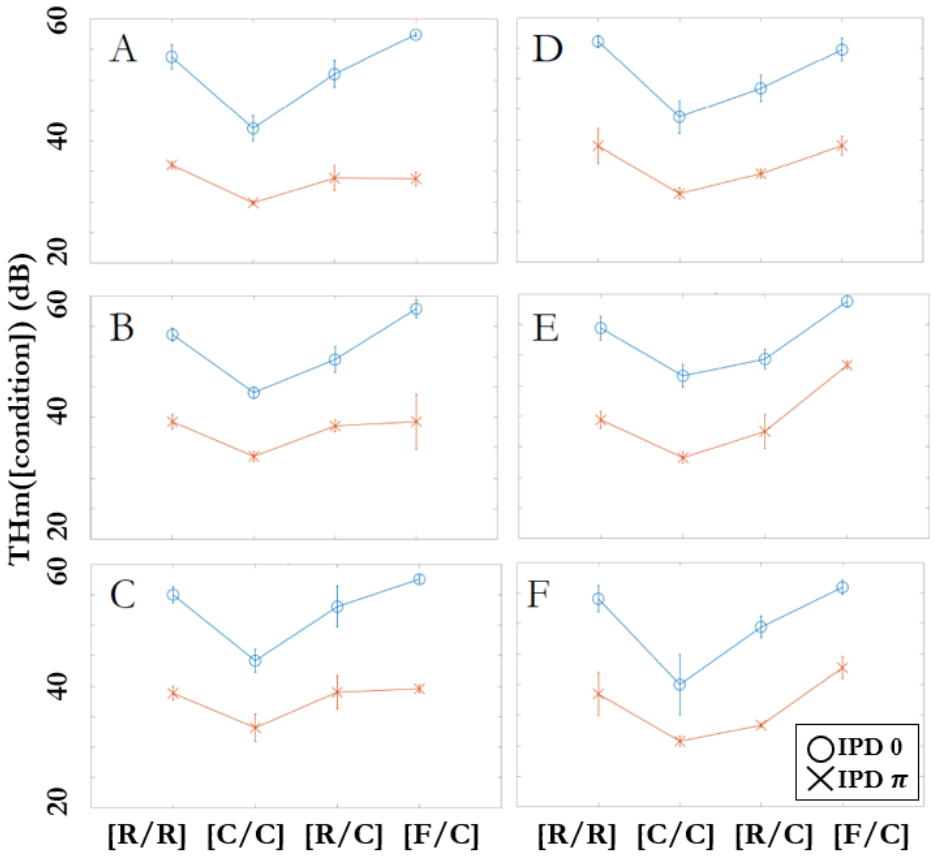


Fig. 5: Masked threshold for the different stimulus conditions for diotic (circle) and dichotic (cross) presentation of the tone. Thresholds measurements were repeated three times.

Onset response

To check for the presence of an onset response evoked by the onset of the different masker intervals, event-related potentials (ERPs) for the conditions were analyzed. Figure 6 and 7 show the grand average ERPs for the conditions with two identical masker intervals ([R/R] and [C/C]) and two different masker intervals ([R/C] and

[F/C]), respectively. Both masker groups evoked a typical onset response at the start of the first masker, caused by the change in stimulus energy. The onset response decayed after about 400 ms, which shows a suitable scaling of the masker-only time interval before the onset of the tone. No such onset was found at the transition from the first to the second masker interval (vertical line). These results also confirm the suitability of the masker design to study the response to the tone independent of the masker onset response in the first masker interval.

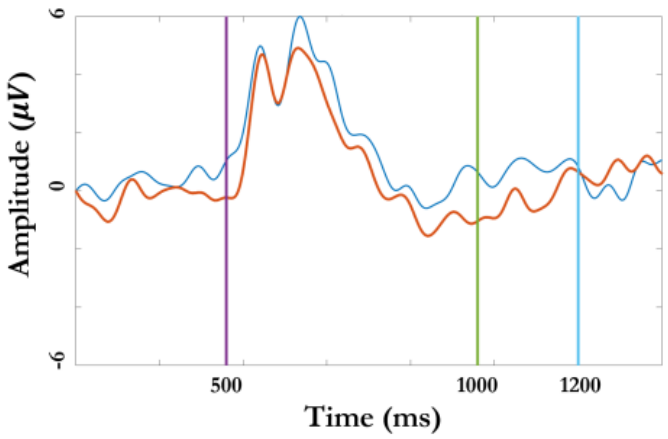


Fig. 6: Average ERPs of [R/R] (thin line) and [C/C] (thick line) conditions. Masker 1 onset (500 ms), Masker 2 onset (1000 ms) and offset (1200 ms).

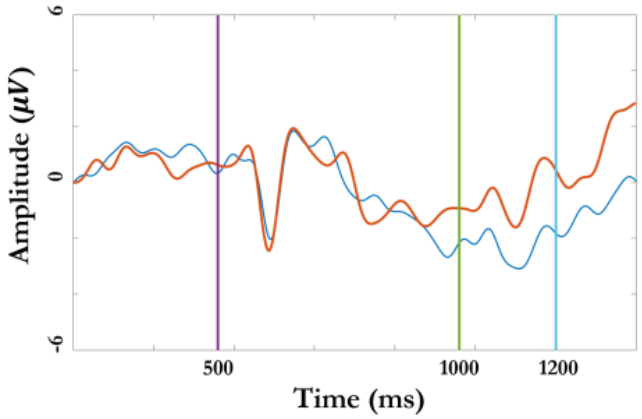


Fig. 7: Average ERPs of [R/C] (thin line) and [F/C] (thick line) conditions. Masker 1 onset (500 ms), Masker 2 onset (1000 ms) and offset (1200 ms).

CONCLUSION

In this study, we designed the stimuli to investigate the streaming effect on CMR and BMLD. The effect of streaming on CMR was observed as shown in previous

studies. There was no streaming effect on BMLD in all conditions except [F/C] condition where there were individual differences. To avoid overlapping between evoked potentials induced by Masker 1 and Masker 2, Masker 1 needs to be 500 ms in length. The preliminary data confirms the applicability of the design. Additional data will provide more conclusive results for the effect of streaming on BMLD and correlations between psychophysics and electro-physiology.

REFERENCES

Dau, T., Ewert, S. D., and Oxenham, A. J. (**2005**), "Effects of concurrent and sequential streaming in comodulation masking release," Auditory signal processing (Springer), 334–342.

Epp, B., Yasin, I., and Verhey, J. L. (**2013**), "Objective measures of binaural masking level differences and comodulation masking release based on late auditory evoked potentials," Hear. Res., **306**, 21–28.

Ewert, S. D. (**2013**), "AFC—A modular framework for running psychoacoustic experiments and computational perception models," Proc. AIA-DAGA, 1326–1329.

Fletcher, H. (**1940**), "Auditory patterns," Rev. Mod. Phys., **12**(1), 47.

Grose, J. H., Buss, E., and Hall III, J. W. (**2009**), "Within-and across-channel factors in the multiband comodulation masking release paradigm," J. Acoust. Soc. Am., **125**(1), 282–293.

Jiang, D., McAlpine, D., and Palmer, A. R. (**1997**), "Responses of neurons in the inferior colliculus to binaural masking level difference stimuli measured by rate-versus-level functions," J Neurophysiol., **77**(6), 3085–3106.

Levitt, H. (**1971**), "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am., **49**(2B), 467–477.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (**2011**), "FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," Comput. Intel. Neuroscience., **2011**, 1–9.

Pressnitzer, D., Meddis, R., Delahaye, R., and Winter, I. M. (**2001**), "Physiological correlates of comodulation masking release in the mammalian ventral cochlear nucleus," J. Neurosci., **21**(16), 6377–6386.

Riedel, H., Granzow, M., and Kollmeier, B. (**2001**), "Single-sweep-based methods to improve the quality of auditory brain stem responses Part II: Averaging methods," Zeitschrift fur Audiologie, **40**(2), 62–85.

Verhey, J. L., Ernst, S. M., and Yasin, I. (**2012**), "Effects of sequential streaming on auditory masking using psychoacoustics and auditory evoked potentials," Hear. Res., **285**(1-2), 77–85.

A comparison of two measures of subcortical responses to ongoing speech: Preliminary results

FLORINE L. BACHMANN[1]*, EWEN N. MACDONALD[1], AND JENS HJORTKJÆR[1,2]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark (DTU), 2800 Kgs. Lyngby, Denmark*

[2] *Danish Research Centre for Magnetic Resonance, Centre for Functional Diagnostic Imaging and Research, Copenhagen University Hospital, 2650 Hvidovre, Denmark*

Neural responses in the auditory brainstem and midbrain are traditionally obtained with repetitions of basic stimuli such as clicks and tones. However, two different methods to measure subcortical responses to ongoing speech with non-invasive electroencephalography (EEG) have recently been published: one based on regularised linear regression (Maddox and Lee, 2018), and the other based on cross-correlation (Etard *et al.*, 2009; Forte *et al.*, 2017). Here, we compare these two methods using the same EEG data set. For both measures, we found prominent peaks in the response functions at latencies consistent with wave V of the auditory brainstem response (ABR; mean latency: 8.19 and 5.97 ms, respectively). The peak response latencies in individual participants were correlated between the regression approach and conventional click-evoked auditory brainstem responses (click-ABRs), suggesting a common underlying neural source. However, similar correlations were not found between the two speech-based methods, nor between the correlation approach and click-ABRs. This could arise from either differences in the methodologies or from variability in the measures.

BACKGROUND

Comparing neural processing at different stages of the auditory pathway provides a deeper understanding of the auditory system. Generally, this interplay has been investigated using different stimuli. Brainstem responses are traditionally investigated with basic stimuli such as tones or clicks, but cortical activity has also been assessed with ongoing speech. Using complex stimuli such as ongoing speech to measure responses at subcortical processing stages could shed light on different facets of early speech processing. Furthermore, it would offer the possibility of simultaneously observing neural responses at the subcortical and cortical level to different speech features using the same ongoing speech stimulus. Recent research suggests that this could be possible. Two independent research groups published two different

*Corresponding author: flbach@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

approaches for measuring subcortical responses to ongoing speech. Maddox and Lee (2018) used a regularised linear regression approach, and Forte *et al.* (2017) used cross-correlation to assess the association between features of the continuous speech stimulus and the recorded EEG. Both groups reported a peak in their response functions, with a latency similar to that of wave V of the auditory brainstem response (ABR; 6.17 $\pm$ 0.31 ms and 9.3 $\pm$ 0.7 ms, respectively). Maddox and Lee (2018) further compared their response derived from ongoing speech with a classical click-evoked auditory brainstem response (click-ABR) and found high correlations for both peak latencies and amplitudes. Although differences in the two methods exist, the similar morphology of the estimated response functions could indicate that they measure equivalent aspects of the brainstem response to speech. The present study compares these two methods to one another based on the same data set, and to a classical ABR measurement.

METHODS

Data acquisition

Participants listened to an audio book while their neural activity was recorded with an electroencephalogram (EEG) system. Fourteen (7 female) young (M_{age} = 23.12 $\pm$ 2.411) native Danish speakers participated in the study. All participants had pure-tone thresholds better than 20 dB hearing level in both ears (measured at standard audiometric frequencies: 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, 6 kHz, and 8 kHz). Each participant provided written informed consent, and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

Measurements were conducted in a soundproof, electrically shielded listening booth. Participants were seated in a comfortable chair in front of a computer screen. Experiment presentation and data acquisition were controlled from outside the booth. The audio book was presented at 65 dB SPL through ER-2 insert earphones (Etymotic Research), with a sampling frequency of 44.1 kHz. The audio book consisted of the beginning of the Danish version of Lord of the Flies by William Golding, read by a male narrator. Longer pauses in the audio book were restricted to 450 ms, and the audio book was cut into trial segments of 50 s duration. To ensure that participants attended the story, three multiple-choice questions were asked after every trial. For each segment, one of the three comprehension questions was presented to the participant prior to listening to the segment. The experiment consisted of 36 trials, and answer accuracy was above 80% for all participants ($M_{correct} = 90.74\% \pm 4.49\%$). To get used to the experimental procedure, participants completed a short training session consisting of two trials before starting the experiment. Data from the training session were not included in the analysis.

To compare speech EEG recordings with standard ABRs, basic click-ABR responses were obtained after the speech experiment. A 10 Hz click train with alternating

polarities was presented at 93 dB peak-to-peak equivalent SPL (to a 1 kHz sinusoid) for five minutes, resulting in 3000 click repetitions. No jitter was applied to the click train.

The EEG was recorded using the Active Two system (BioSemi) with a sampling rate of 16384 Hz. Electrical potentials were measured from 32 scalp electrodes placed according to the 10-20 system, and 4 external electrodes placed on the left and right mastoid bones, as well as over and below the right eye to measure the electrooculogram (EOG).

Analysis

The EEG data was pre-processed using Matlab (MathWorks) and the Fieldtrip toolbox (Oostenveld *et al.*, 2011). Pre-processing of the EEG data was identical for both speech-EEG methods and the click-ABR. It entailed high-pass filtering at 1 Hz to exclude slow electrode drifts and re-referencing to the average of the two mastoid electrodes, after which the mastoid channels were discarded. Noisy EEG channels were further identified through visual inspection and discarded (on average, 0.36 channels per participant were discarded). All analyses reported here focused on electrode Pz. Both speech-based approaches were computed twice, once for the audio segment that was heard during the respective EEG recording (corresponding audio), and once for all other audio segments, which had been presented at another time during the study (random audio). After the analyses described below, the responses at Pz for each participant were averaged over trials. From this calculated individual response, the largest local maximum between 1 and 11 ms was identified as the response peak, and the respective time point as peak latency or delay. For one participant, no clear local maximum between 1 and 11 ms could be identified with the regularized linear regression approach. This participant was therefore excluded from average latency estimations for the regression approach, and all correlations entailing the regression approach. All filters applied to the audio and the EEG in the common pre-processing such as at later processing stages were applied both in forward and reverse direction, compensating for filter delays.

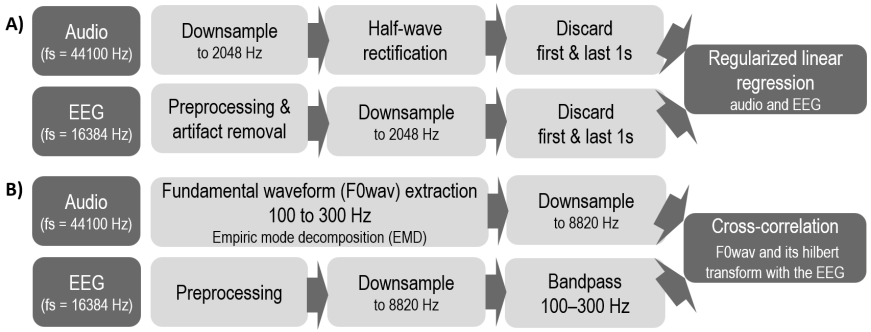


Fig. 1: Schematic description of the analysis pipelines of the (A) regression approach, and (B) correlation approach.

Regularized linear regression approach

The analysis pipeline for the subcortical regularised linear regression approach is depicted in Fig. 1 A, and is based on Maddox and Lee (2018). This analysis is similar to techniques used for measuring speech entrainment at the cortical level (Hjortkjær *et al.*, 2018; Lalor *et al.*, 2009).

In a first step, the audio was down-sampled to 2048 Hz. To account for cochlear processing, and for better comparison with the EEG data, half-wave rectification was applied to the audio. The EEG signal was first pre-processed and artefacts removed, after which signals were down-sampled to 2048 Hz. Muscle and eye movement artefacts were identified as extreme values of the z-scored EEG and the EOG channels, respectively, using individual cut-offs (average cut-off: z-value of 5.07 for muscle artefacts, and 0.57 for eye artefacts). The EOG channels were discarded thereafter. Both EEG recordings and audio from the affected time points were not considered in the analysis.

The first and the last second were discarded from both the audio and the EEG signal, and the two pre-processed signals were then fed into a ridge regression analysis. Using the Telluride Decoding Toolbox (Akram *et al.*, 2017), a forward model was computed for a ridge parameter of $\lambda = 2^{12}$. Time lags between -10 and 23 ms were considered for this analysis. The resulting regression weights or temporal response function (TRF) map from the time-lagged audio stimulus linearily to the EEG response, and characterises the stimulus-evoked neural response (Fuglsang *et al.*, 2017; Ding and Simon, 2012b; Ding and Simon, 2012a; Lalor *et al.*, 2009). The TRF recorded at electrode Pz was up-sampled to the original recording sampling rate of 16384 Hz, and interpreted as the response.

Cross-correlation approach

Figure 1 B shows the analysis pipeline for the cross-correlation approach, which was conducted similarly to how Forte *et al.* (2017) applied it. In contrast to the regression approach, the fundamental waveform of the audio signal was extracted prior to the analysis. The fundamental waveform was extracted according to Kegler (2019), using empirical mode decomposition. No half-wave rectification was applied in the process. The fundamental waveform was restricted between 100 and 300 Hz and down-sampled to 8820 Hz. The EEG recording was pre-processed, but unlike the regularized linear regression approach, no further artefact rejection was applied for the cross-correlation approach. The EEG signal was also down-sampled to 8820 Hz, and then band-pass filtered between 100 and 300 Hz to offer a fair comparison between audio and EEG recording.

The cross-correlation of the EEG recorded at Pz and the fundamental waveform, and the imaginary part of its Hilbert transform, were then computed and interpreted as the real and imaginary part of a complex correlation function, respectively. Time lags between -60 to 60 ms were considered. The computed cross-correlation functions

were up-sampled to the original recording sampling rate of 16384 Hz, and the magnitude peak of the complex correlation function was identified.

Click-ABRs

Classical ABRs at electrode Pz were computed from EEG recordings to the click stimuli. After pre-processing, the EEG data recorded for every click was defined as one trial, and aligned. Trials with voltages exceeding 50 μV were interpreted as including artefacts and excluded from further analysis (3.56 $\pm$ 7.19 trials excluded on average). The ABR data were analysed without down-sampling.

RESULTS

Results from all three compared methods showed response peaks with latencies below 10 ms, consistent with a brainstem or midbrain origin of the response. For the click-ABR, wave V occurred at a latency of 6.69 $\pm$ 0.28 ms. On average, the peak responses were observed at earlier times with the cross-correlation approach (5.97 $\pm$ 1.45 ms), than with the regression approach (8.19 $\pm$ 0.46 ms; Fig. 2).

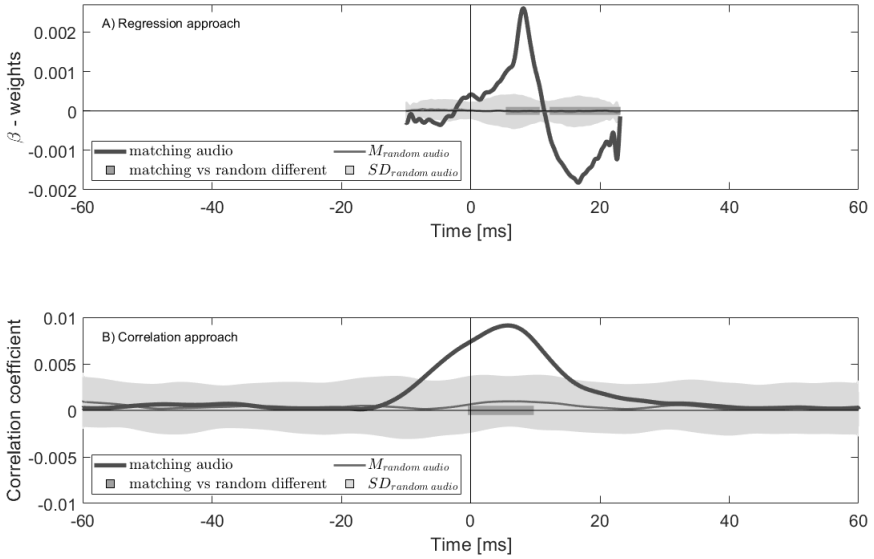


Fig. 2: Comparison of subcortical responses to ongoing speech computed with the two methods: (A) temporal response function (TRF; $\lambda = 2^{12}$), (B) correlation function. Both responses were calculated at electrode Pz. Due to computational limitations, only time lags between -10 and 23 ms were considered for (A).

A two-sided two-sample t-test between the average responses to the matching and the random audio was conducted at every time point and significant differences were observed for both approaches (from 5.49 to 10.62 ms such as from 12.27 ms on for the regression approach and from -0.37 to 9.70 ms for the correlation approach; $\alpha = 0.05$,

no correction applied). The regression approach produced a sharper response peak compared to the cross-correlation approach.

The range of the response peaks was similar across the three methods (Fig. 3 A). Paired-sample t-tests with Bonferroni correction were conducted pairwise between all methods. Latencies obtained with the regression approach were significantly different from both those measured with the correlation approach, and click-ABR wave V (both $p < 0.001$). Latencies from the correlation approach and the click-ABR did not differ significantly ($p = 0.071$).

The average peak obtained with the three different methods for each individual were compared (Fig. 3 B and C). Pearson's correlation coefficient was computed for all three comparisons. The correlations between the regression and the correlation approach and the other two approaches did not yield significance ($p = 0.363$ and $p = 0.136$, for regression and click-ABR respectively). However, latencies of the regression approach and the click-ABRs were highly correlated ($\rho = 0.844$; $p < 0.001$ after Bonferroni correction).

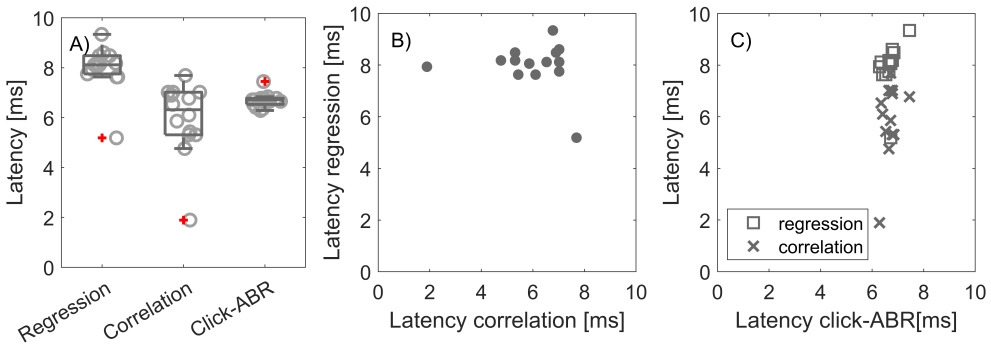


Fig. 3: Comparison of the time lags between the different methods. (A) Box plot of latencies obtained with the different methods, (B) comparison of latencies between regression and correlation approach, (C) comparison of latencies from both regression and correlation approach with the click-ABR.

DISCUSSION

The purpose of the study was to compare two novel methods for deriving a subcortical response to ongoing speech using the same EEG data set. The response latencies obtained for both the regression and the correlation approach lay consistently within a similar range to that reported by the reference studies (Forte *et al.*, 2017; Etard *et al.*, 2009; Maddox and Lee, 2018), and are similar to the latencies of wave V in traditional click-ABRs, both observed here and in previous studies (Garrett and Verhulst, 2019; Maddox and Lee, 2018). Latencies obtained with the regression approach and the click-ABR were correlated. Taken together, our results confirm that both approaches measure aspects related to brainstem processing. Participants'

latencies were not significantly correlated between the two speech-based methods, nor between the correlation approach and the click-ABRs. This may be because the individual differences across the young normal-hearing listeners tested here were small relative to the variance in the latency estimation inherent in each method. If the three methods measure equivalent aspects of the subcortical response, then significant correlations might be observed in studies that include more participants with a broader age range and/or listeners with hearing loss.

Significant differences were observed for average latencies between both speech-based methods, such as between the regression approach and the click-ABRs. Given the differences across methods, there are several possible explanations for this result which are still consistent with the three methods measuring similar aspects of the auditory brainstem response. First, the artefact rejection procedure varied between the three methods. For the click-ABR, trials that exceeded 50 μV were excluded, whereas for the regression approach, a statistical analysis of EOG and EEG data was used to identify muscle and eye movement artefacts. For the cross-correlation approach, no artefact rejection was applied. These differences may have contributed to the observed differences in relative latency. In addition, the latency introduced by the analysis window used to extract F0 in the cross-correlation approach may differ from that in the peripheral auditory system, biasing the results from this approach.

In the present study, only the latency of the peak response was considered across the three methods. However, other aspects, such as peak amplitude, may be of interest for investigating group differences in sub-cortical processing. Thus, further work is needed to compare these methods using other metrics and to investigate how robust the two approaches are.

SUMMARY

The regularized linear regression approach of Maddox and Lee (2018) and the cross-correlation approach of Etard *et al.* (2009) were applied to the same EEG data to derive a subcortical response to speech. The response latencies of both measures were similar to each other and to that of a traditional click-ABR approach.

ACKNOWLEDGEMENTS

This work was partially financially supported by the Sonova Holding AG, and J.H. was supported by the Novo Nordisk Foundation, synergy grant NNF17OC0027872 (UHeal). The authors would like to thank Jonatan Märcher-Rørsted for his support with parts of the analysis, and Rikke Skovhøj Sørensen for conducting the audiometric testing of the participants.

REFERENCES

Akram, S., A. de Cheveigné, P. U. Diehl, E. Graber, C. Graversen, J. Hjortkjaer, N. Mesgarani, L. Parra, U. Pomper, S. Shamma, J. Simon, M. Slaney, and D. Wong (**2017**). *Telluride Decoding Toolbox*. https://github.com/neuromorphs-2017-decoding/telluride-decoding-toolbox.

Ding, N. and J. Z. Simon (**2012a**). "Emergence of neural encoding of auditory objects while listening to competing speakers," Proc. Natl. Acad. Sci. U.S.A., **109** (29), 11854–11859, doi: 10.1073/pnas.1205381109.

Ding, N. and J. Z. Simon (**2012b**). "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," J. Neurophysiol., **107** (1), 78–89, doi: 10.1152/jn.00297.2011.

Etard, O., M. Kegler, C. Braiman, A. E. Forte, and T. Reichenbach (**2009**). "Decoding of selective attention to continuous speech from the human auditory brainstem response," NeuroImage, **200**, 1–11, doi: 10.1016/j.neuroimage.2019.06.029.

Forte, A. E., O. Etard, and T. Reichenbach (**2017**). "The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention," eLife, **6**, e27203, doi: 10.7554/elife.27203.001.

Fuglsang, S. A., T. Dau, and J. Hjortkjær (**2017**). "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," NeuroImage, **156**, 435–444, doi: 10.1016/j.neuroimage.2017.04.026.

Garrett, M. and S. Verhulst (**2019**). "Applicability of subcortical EEG metrics of synaptopathy to older listeners with impaired audiograms," Hear. Res., **380**, 150–165, doi: 10.1016/j.heares.2019.07.001.

Hjortkjær, J., J. Märcher-Rørsted, S. A. Fuglsang, and T. Dau (**2018**). "Cortical oscillations and entrainment in speech processing during working memory load," Eur. J. of Neurosci., 1–11, doi: 10.1111/ejn.13855.

Kegler, M. (**2019**). *Fundamental waveforms extraction*. https://github.com/MKegler/fundamental_waveforms_extraction.

Lalor, E. C., A. J. Power, R. B. Reilly, and J. J. Foxe (**2009**). "Resolving precise temporal processing properties of the auditory system using continuous stimuli," J. Neurophysiol., **102** (1), 349–359, doi: 10.1152/jn.90896.2008.

Maddox, R. K. and A. K. Lee (**2018**). "Auditory brainstem responses to continuous natural speech in human listeners," eNeuro, **5** (1). doi: 10.1523/eneuro.0441-17.2018.

Oostenveld, R., P. Fries, E. Maris, and J.-M. Schoffelen (**2011**). "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," Comput. Intell. and Neurosci., **2011**, 1, doi: 10.1155/2011/156869.

Provoking and minimising potentially destructive binaural stimulation effects in auditory steady-state response (ASSR) measurements

SAM DAVID WATSON[1,*], SØREN LAUGESEN[2] AND BASTIAN EPP[1]

[1] *Hearing Systems section, Department of Health Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Interacoustics Research Unit, DK-2800 Lyngby, Denmark*

An aided sound-field auditory steady state response (ASSR) has the potential to be used to verify the quality of fit of hearing aids on infants. Each aided ear should ideally be tested independently, but it is suspected that binaural testing may be used by clinics to reduce test time. This study simulates 'clinically conceivable' dichotic ASSR sound-field conditions to examine the risk of making false judgements due to unchecked binaural effects. Unaided ASSRs were recorded with a clinical two channel EEG system for 15 normally hearing subjects using a three-band CE-Chirp^{TM} stimulus. It was found that the noise corrected power of a response harmonic can be reduced by up to 10 dB by introducing large ITDs equal to half the time period of the stimulus envelope. This could lead to concluding that a hearing aid fitting is poor, even though the fitting would have passed separate monaural ASSR tests (false referral). No effect was detected for simulated lateralisations of the stimulus, which is beneficial for a proposed aided ASSR approach. Full-scalp ASSR recordings show distinct SNR reductions and topographical changes in response to the large ITDs, and demonstrate the vulnerability of ASSR to montage and inter-subject variation. Findings suggest that multi-harmonic detectors could make binaural measurements robust to artificial reductions of response harmonics cause by large ITDs.

INTRODUCTION

The auditory steady-state response (ASSR) can be used to assess the hearing of non-responsive subjects such as infants, and by extension, when the stimulus is presented as a sound field, assess individually the quality-of-fit of the hearing aid(s) (HA) once adapted to a patient (Rance, 2008). Stimuli presented above the estimated aided ASSR thresholds, monaurally, should provoke a response which is greater than the noise floor by a predetermined F-test ratio to be considered present. The presence of a response then infers that the HA fit is good.

There is, however, the potential that clinics wish to test binaurally, as it has the possibility not only to reduce the testing time by half compared to two independent

*Corresponding author: sadaw@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

monaural tests but also improves test transparency to parents (Mehta *et al.*, 2019). Despite the apparent benefits binaural testing also has the possibility to introduce interference effects, as a nominally diotic presentation may be disturbed (e.g. by head movement) to produce in reality a dichotic stimulus. In the worst case this may produce artificial test false-negatives, leading to unnecessary additional referrals.

Subsequently, this study considers if it is possible that a single perturbed binaural measurement could lead to a referral whereas the preferred two individual monaural measurements would not.

There are few previous studies investigating plausible clinical dichotic scenarios; Zhang & Boettcher (2008) found that separately increasing interaural time and level differences (ITDs and ILDs) caused positive and negative binaural interaction components (BICs) respectively on 80 Hz click ASSR, while Narayanan (2018) found that a monaural ASSR to a 4 kHz octave band chip carrier can be significantly reduced by a lateralised presentation due to the subject's head shadow. Riedel & Kollmeier (2002) used 15 Hz click auditory brainstem responses (ABR) and found that laterally perceived stimuli produce lower wave V amplitudes than those perceived centrally. Much larger ITDs may also occur in the cases of unilateral digital aid users, where signal processing may delay the arrival of sound ($\approx 5 - 10$ ms Kates, 2008) to the aided ear compared to the unaided. The current study assesses how large ITDs as well as simulated 'realistic' combinations of ITDs and ILDs effect ASSR amplitudes.

METHODS

Two-channel EEG

Fifteen (seven female) normally hearing (symmetric $<$ 20 dB HL 250 Hz - 4 kHz) young (mean 23.2 years, $\pm$ 2.20) listeners participated in the measurements with the two-channel EEG. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (ref. H-16036391). The stimulus, presented in a soundproof booth, unaided, over Eytmotic ER-1 insert phones, consisted of three modified CE-Chirp^{TM} chirp trains: a double octave width 707 Hz centered (40 Hz rate band), and two single octave width 2 kHz and 4 kHz (90 Hz rate bands). Stimuli were presented at a nominal broadband free-field level of 65 dB SPL with each chirp train scaled to match the equivalent band power of the Internation Speech Test Signal (ISTS) (Holube *et al.*, 2010); except in the case of the lateralised stimuli where frequency banded gains and attenuations were applied derived from Denk *et al.* (2018) behind-the-ear (BTE) head related transfer functions (HRTFs), shown in Figure 2.

The stimulus conditions consisted of a reference diotic condition and seven dichotic conditions. The dichotic stimuli were of three types: (1) rate-specific ITD resulting in envelope anti-phase of either the 40 Hz rate band ('Inv 40', see Figure 1) or the 90 Hz rate band ('Inv 90'), (2) interaural inverse polarities with no effect on the envelopes ('Inv Pol'), (3) lateralised with realistic combinations of ITDs and ILDs to simulate

incidence of the stimulus from ±45 degrees (ITD: 354 µs) and ± 90 degrees (ITD: 688 µs) on the azimuthal plane. Data collection for each condition continued until a Bonferroni corrected F-test threshold (corresponding to a p-value of 0.01) was reached for all three response bands in both channels, or until 15 minutes had passed.

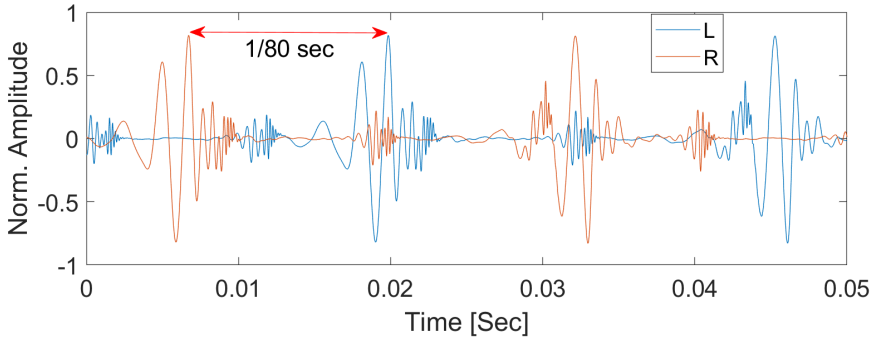


Fig. 1: Stimulus waveforms for left and right channels in the Inv 40 condition. An ITD of $\approx 1/80$ s was introduced to place the envelopes of the 40 Hz rate band into interaural antiphase)

ASSRs were measured using an adapted clinical two-channel EEG system (Interacoustics Eclipse) connected to an RME Fireface UC soundcard, with a vertex - mastoid montage. The data were recorded and processed in custom software using MATLAB. The first two harmonics of the corresponding repetition rate in the averaged frequency domain representation were considered. Reported ASSR levels were noise corrected (NC) by subtracting the mean level of the surrounding 20 frequency bins of each respective response bin, avoiding other response bins or known particularly contaminated bins (e.g., line noise, radio bands).

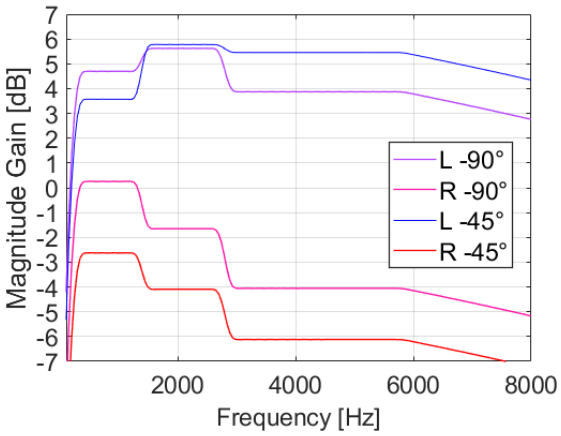


Fig. 2: Frequency band specific gains or attenuations, derived from Denk *et al.* (2018), used to simulate head shadow and baffle effects for stimuli incident from the left; $-45°$ and $-90°$.

The resulting ASSR amplitudes were compared by fitting mixed linear models, implemented in R using the 'lme4' (Bates *et al.*, 2015) package, followed by planned pairwise analysis of variance (ANOVA) comparisons between each condition and the diotic reference with 'lmerTest' (Kuznetsova *et al.*, 2017).

64-channel scalp topography

The Inv 40 condition and the diotic reference were repeated using a 64 channel (BioSemi Active Two) full scalp montage. Four subjects participated in this experiment; two that showed large changes and two that showed small changes between these two conditions in the preceding experiment. The data were low-pass filtered at 1638 Hz, and then stored. Offline, the raw data epochs were high and low pass filtered at 20 Hz and 400 Hz respectively using forward-backward filtering with 2^{nd} order Butterworth filters, in addition to notch filters at 50 Hz, 100 Hz, and 150 Hz. Epochs were automatically and manually screened for muscle activity artifacts and affected epochs were removed. Channels were re-referenced to the grand mean of all EEG channels. The scalp topography was calculated from the signal-to-noise ratios (SNRs), with consistently noisy channels replaced with data interpolated from spatial neighbours using the MATLAB cubic interpolation. Most processing was implemented using the FieldTrip toolbox (Oostenveld *et al.*, 2011) for MATLAB.

RESULTS

Two-channel EEG

Figure 3 shows the noise-corrected ASSR amplitudes for the two-channel recording divided by harmonic and band response. Only the Inv 40 condition produced responses significantly lower than the corresponding Diotic: in the first harmonic of the 707 Hz band, and in the second harmonics of the 2 kHz and 4 kHz bands. First harmonic responses are generally greater than second harmonic, and 707 Hz band responses are generally greater than the other two bands. This is supported by the significant factors Harmonic and Stimulus Frequency in Table 1.

Factor	ASSR NC Power Level	
	F statistics	p
Subject (random)		<0.0001***
Condition	F(9,1467.4) = 45.6	<0.0001***
Stimulus Freq.	F(2,1467.1) = 123	<0.0001***
Harmonic	F(1,1467.1) = 496	<0.0001***
Channel	F(1,1467.4) = 1.83	0.176
Condition: Stim. Freq.	F(18,1467.0) = 0.565	0.925
Condition: Harmonic	F(9,1467.0) = 1.47	0.154
Stim. Freq.:Harmonic	F(2,1467.1) = 3.95	0.0195*
Stim. Freq.:Channel	F(2,1467.0) = 4.21	0.0151*
Cond.:Stim. Freq.: Harm	F(18,1467.1) = 8.73	<0.0001***

Table 1: ASSR NC level F-test statistics derived from a mixed linear models fit to the processed data from the main 'phase 2' experiment. Type III Analysis of Variance with Satterthwaite's method. Significance codes: 0 - ***, 0.001 - **, 0.01 - *, 0.05 - .

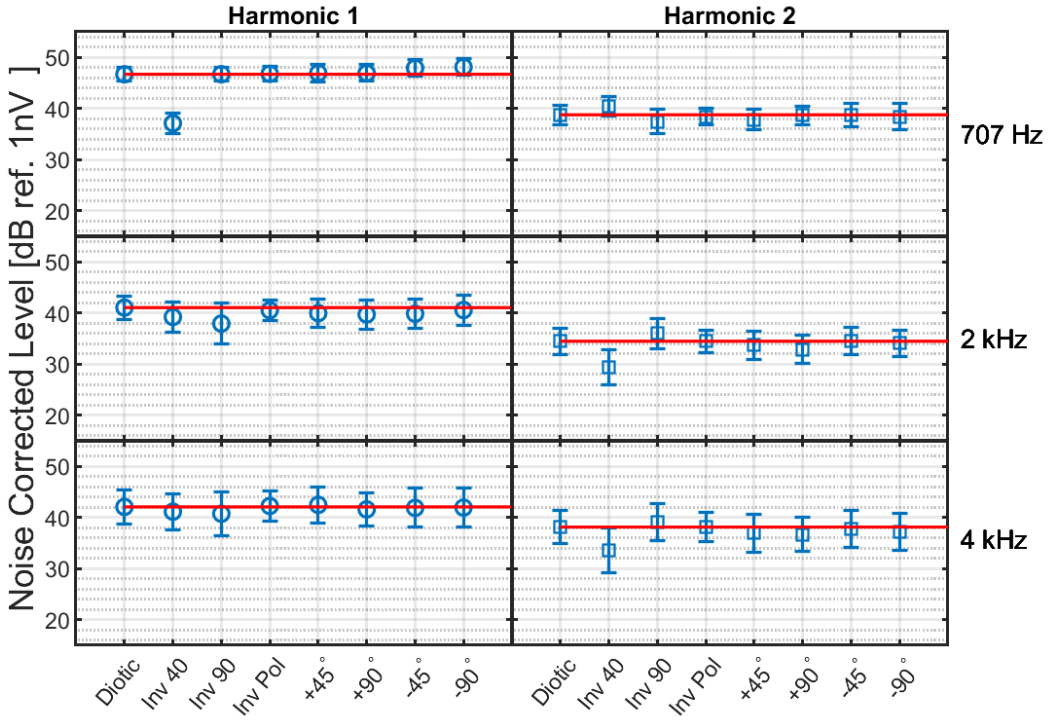


Fig. 3: Noise corrected level of the ASSR for each stimulus condition, shown separately for each stimulus frequency band and response harmonic. Means were taken across subjects and both channels. Error bars represent 95% confidence intervals. The red horizontal lines correspond to the mean level of the relevant reference diotic condition.

Table 1 summarises the F-test statistics from the model fit to the ASSR NC level data. The strong random effect of Subject indicates significant variation among subjects, as expected (Laugesen *et al.*, 2018). The significant factor Condition and three way interaction indicates that the Inv 40 condition has a varied effect depending on the response harmonic and band. The planned pairwise comparisons confirms that the NC ASSR amplitudes were significantly lower in the Inv 40 condition in the first harmonic 707 Hz band (-9.9 dB, $p < 0.0001$) than the corresponding Diotic condition response. Furthermore, amplitudes were also significantly reduced in the Inv 40 Condition in the second harmonic, 2 kHz and 4 kHz bands, (-6.8 dB, $p < 0.0001$ & -5.7 dB, $p = 0.000432$), compared to the corresponding Diotic condition response. No other modifications of the diotic stimulus led to a significant change in the ASSR amplitude.

64-channel scalp topography

Figure 4 shows the 707 Hz band SNR topography for all 64 electrodes for the diotic (left column) and the Inv 40 condition (right column).

In the diotic condition, all listeners had a maximum of the SNR along the vertex-nasion axis with a minimum in often both hemispheres. The location of these minima though were variable. In the Inv 40 condition a reduction of the average SNR was found across all listeners. Inversely now all listeners showed a minimum of the SNR along the vertex-nasion axis, with maxima now occurring reliably lateralised but varying from frontal (subject 8) to occipital (subject 17). It is interesting to note that the vertex electrode in the two-channel EEG montage was for most subjects located in the vicinity of a maximum in the diotic condition and close to a minimum in the Inv 40 condition.

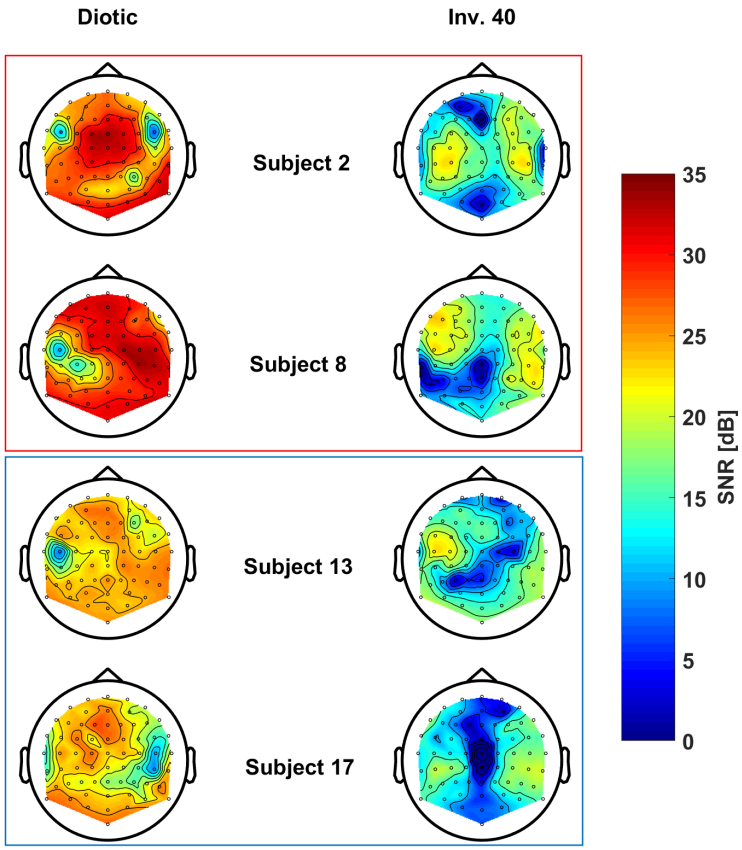


Fig. 4: Mean 707 Hz band SNR topography of the 64 channel scalp montage separately plotted for four subjects during diotic (left column) and Inv 40 (right column) stimulation. The red and blue enclosing boxes indicate the subjects demonstrating large and small reductions in ASSR level from diotic to Inv 40 stimulation respectively.

The SNR patterns in the Inv 40 condition were similar across all four subjects, but during the diotic condition there was a distinction between two groups. Subjects showing a large reduction (red) do so because they start with overall greater SNRs in response to the diotic stimulation than those who show a smaller reduction (blue).

DISCUSSION

No effect of the lateralised stimuli was found. This may be because lateralisation of stimuli has no consequence for ASSR production. Alternatively, as Zhang & Boettcher (2008) found, the BICs in response to ILD and ITD may be opposite or too small to detect, resulting in no observable net effect. This is encouraging for the clinical implementation of sound-field ASSR, as it seems normal frontal incidence of the stimulus does not need to be maintained. This is in opposition to the ABR behaviour reported in Riedel & Kollmeier (2002).

The reduction in response level in the second harmonic compared to the first, and in the 2 kHz and 4 kHz bands compared to the 707 Hz band match the findings of Laugesen *et al.* (2018). The difference between bands is attributed to the varied repetition rates and chirp frequency widths.

The Inv 40 condition had the effect of reducing the 707 Hz band response in the first harmonic, but also resulted in the reduction of the 2 kHz and 4 kHz bands second harmonic responses. Further analysis reveals that the Inv 40 condition also places the envelopes of the second harmonics of the two 90 Hz bands into antiphase, so this response reduction effect seems related to interaural envelope phase. This would suggest that binaural testing on unilateral HA users, which the Inv 40 condition simulates, could artificially suppress some response harmonics and result in good HA fittings being pronounced poor. The chance of this occurring could be minimised or eliminated by using a response detector which always considers multiple harmonics. It is currently unexplained why significant reductions in ASSR are seen under the Inv 40 envelope anti-phasic condition, but none is seen in the similar Inv 90 condition.

The significant factor Subject, along with Figure 4, suggests that during diotic stimulation there is a large inter-subject variation in the response SNR across the head. Furthermore, even in these ideal conditions low SNR areas on the scalp develop, and that their exact location cannot be precisely predicted. This demonstrates the vulnerability of HA validation by ASSR to montage choice as well as diversity in individual's response amplitudes.

In conclusion, if performing binaural tests, it should be avoided for unilateral HA fittings, which may lead to significant artificial response suppression and therefore some false referrals. Naturally however, monaural tests will always give more clarity as to the quality-of-fit of each separate HA, since a single well fitting aid in a binaural test would still produce an ASSR.

ACKNOWLEDGEMENTS

The work of the first author was supported by the William Demant Foundation.

REFERENCES

Bates, D., Martin, M., Bolker, B., & Walker, S., (**2015**). "Fitting Linear Mixed-Effects Models Using lme4", J. Stat. Softw., **67**(1), 1-48.

Denk, F., Ernst, S. M. A., Ewert, S. D., and Kollmeier, B., (**2018**). "Adapting hearing devices to the individual ear acoustics: Database and target response correction functions for various device styles", Trends Hear., **22**, 1-19.

Holube, I., Fredelake, S., Vlaming, M., Kollmeier, B., (**2010**). "Development and analysis of an International Speech Test Signal (ISTS)", Int. J. Audiol., **49**(12), 891-903.

Kates, J. M., (**2008**). *Digital Hearing Aids*, Plural Publishing, Inc., San Diego, California, USA.

Riedel, H., Kollmeier, B., (**2002**). "Auditory brain stem responses evoked by lateralized clicks: Is lateralization extracted in the human brain stem?", Hear.s Res., **163**, 12-26.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B., (**2017**). "lmerTest Package: Tests in Linear Mixed Effects Models", J. Stat. Softw., **82**(13), 1-26.

Laugesen, S., Rieck, J. E., Elberling, C., Dau, T., & Harte, J. M., (**2018**). "On the Cost of Introducing Speech-Like Properties to a Stimulus for Auditory Steady-State Response Measurements", Trends Hear., **22**, 1-11.

Mehta, K., Mahon, M., Watkin, P., Marriage, J., & Vickers, D., (**2019**). "A qualitative review of parents' perspectives on the value of CAEP recording in influencing their acceptance of hearing devices for their child", Int. J. Audiol., **58**(7), 401-407.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J., (**2011**). "FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data", Comput. Intel. Neurosc., **2011**, 1-9.

Rance, G. (**2008**). "Part A - The Role of Auditory Steady-State Responses in Fitting Hearing Aids", *The Auditory Steady-State Response: Generation, Recording, and Clinical Application*, Plural Publishing, Inc. San Diego, California, USA, 241 - 258.

Zhang, F., & Boettcher, F. A., (**2008**). "Effects of interaural time and level differences on the binaural interaction component of the 80 Hz auditory steady-state response", J. Am. Acad. Audiol., **19**(1), 82-94.