Individual speech recognition in noise, the audiogram and more: Using automatic speech recognition (ASR) as a modelling tool

BIRGER KOLLMEIER*, MARC RENÉ SCHÄDLER, ANNA WARZYBOK, BERND T. MEYER, AND THOMAS BRAND

Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, Oldenburg, Germany

To characterize the individual patient's hearing impairment, a framework for auditory discrimination experiments (FADE, Schädler *et al.*, 2015) was extended here using different degrees of individualization. FADE has been shown to predict the outcome of both speech recognition tests and psychoacoustic experiments based on simulations using an automatic speech recognition (ASR) system which requires only few assumptions. It builds on the closed-set matrix sentence recognition test which is advantageous for testing individual speech recognition in a way comparable across languages. Individual predictions of speech recognition thresholds in stationary and in fluctuating noise were derived using the audiogram and an estimate of the internal detector noise ("level uncertainty"). Either "typical" audiogram shapes with or without a "typical" level uncertainty or the individual data were used for individual predictions. As a result, the individualisation of the level uncertainty was found to be more important than the exact shape of the individual audiogram to accurately model the outcome of the German matrix test in stationary or fluctuating noise for listeners with hearing impairment.

INTRODUCTION

Recent progress in computational modelling of the normal and impaired auditory system nurtures the hope that a better understanding is achieved of how hearing impairment affects speech communication in daily life. This will help to construct and assess more effective hearing devices. A first approach to provide a model framework which might be developed into an "objective yard stick" in rehabilitative audiology is considered here: The prediction of speech recognition thresholds (SRTs) in noise for an individual based on known audiological data (such as, e.g., the audiogram or measures of supra-threshold processing deficits). By comparing the predictions with the individual empirical SRTs, any special problems of the patient in understanding speech in noise other than explainable from his/her audiogram (such as, e.g., due to auditory neuropathy or more central or cognitive components of hearing impairment) may become obvious.

*Corresponding author: birger.kollmeier@uni-oldenburg.de

Proceedings of ISAAR 2015: Individual Hearing Loss – Characterization, Modelling, Compensation Strategies. 5th symposium on Auditory and Audiological Research. August 2015, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C. Dalsgaard, L. Tranebjærg, and T. Andersen. ISBN: 978-87-990013-5-4. The Danavox Jubilee Foundation, 2015.

Traditional modelling approaches for speech recognition are either based on predefined features (like an energy increase in a certain auditory band) or on instrumental measures that are calibrated using a set of reference thresholds (like the Articulation Index or Speech Intelligibility Index-based methods, see ANSI 1997; Meyer and Brand, 2013). More sophisticated approaches are based on psychoacoustical processing models (e.g., Holube and Kollmeier, 1996, Dau *et al.*, 1997; Jürgens and Brand, 2009), but require an "optimal detector" that possesses perfect prior knowledge about the to-be-recognized signals. The strong assumption of an optimal detector provides the model with an unfair advantage over human listeners that perform the same task, and may even weaken the need of an optimum auditory-system-inspired processing front end to achieve human-like performance, which, in turn, could be crucial to accurately model human sound perception.

An alternative way of predicting both sentence recognition thresholds and psychoacoustic performance using automatic speech recognition (ASR) without requiring a predefined reference or an optimal detector was recently proposed by Schädler *et al.* (2015; 2016). They predicted the outcome of the German matrix sentence recognition test (Kollmeier *et al.*, 2015) for different types of stationary background noise using Mel-frequency cepstral coefficients (MFCCs) as a front-end and whole-word Gaussian mixture/hidden Markov models (HMMs) as a back-end. By training and testing the ASR system with noisy matrix sentences on a broad range of signal-to-noise ratios (SNRs) they were able to predict SRTs for listeners with normal hearing with a remarkably high precision, outperforming SII-based predictions. In a second study, they extended the so-called simulation framework for auditory discrimination experiments (FADE) to successfully simulate basic psychoacoustical experiments as well as more complex Matrix sentence recognition tasks with a range of feature sets (front-ends). Schädler *et al.* (2015) concluded that the proposed FADE framework is able to predict empirical data from the literature with a single set of parameters, less assumptions compared to traditional modelling approaches, and without the need of an empirical reference condition.

The aim of the current study is to extend the FADE approach to model the effect of hearing impairment on speech recognition thresholds obtained with the German Matrix test in stationary and fluctuating noise. Therefore, different degrees of individualization for the model predictions were employed and compared with the empirical results for 99 normal-hearing and hearing-impaired listeners (198 ears).

METHODS

FADE approach

The simulation framework for auditory discrimination experiments (FADE) from Schädler *et al.* (2016) was used to simulate the outcome of the German Matrix test in a stationary and a fluctuating noise condition (see Schädler *et al.*, 2015, for details). The speech material consists of 120 recorded semantically unpredictable sentences with a fixed syntax (name-verb-number-adjective-object, like "Peter sees eight wet chairs".) For each word class, ten alternatives exist. The adaptively determined SRT

denotes the SNR that corresponds to 50%-words-correct performance. To obtain SRTs with FADE, an automatic speech recognizer (ASR) was trained and tested with noisy sentences on a broad range of SNRs (−24dB to +6dB), and the lowest SNR which resulted in 50%-words-correct recognition performance was interpolated and used as the predicted SRT. The ASR system used modified MFCCs as a front-end. On the back-end side, HMMs were used to model speech with whole-word models based on a "parametrically hearing-impaired" acoustical representation provided by the front-end. Hearing impairment was modelled in the front-end and implemented in the log Mel-spectrogram (logMS) from which the MFCC features were derived. A frequency-dependent attenuation was used to model an attenuation-loss (A) by clipping the amplitude values in each channel to the corresponding (interpolated) threshold from the audiogram. To model a supra-threshold distortion loss (D), a level uncertainty was implemented in the logMS by adding a Gaussian white noise with a standard deviation of u_L.

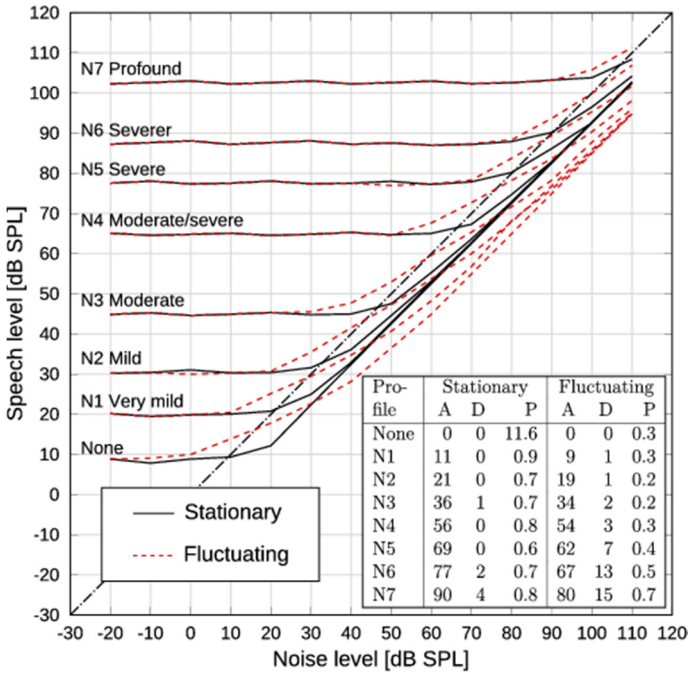
Audiological Data

Results from Brand and Kollmeier (2002) were used for comparing the predictions with empirical data. The data included measurements from 99 listeners (198 separately measured ears) ranging in age from 23 to 82 years (mean and standard deviation: 61.4 ± 13.2 years) and covering a broad range of hearing loss with the PTA varying from 0 to 80 dB HL (mean: 40.5 ± 16.1 dB HL). SRTs were obtained with the German matrix test in stationary ICRA1 and fluctuating ICRA5-250 noise. The ICRA5-250 noise is a speech-like modulated noise which simulates the long-term frequency spectrum and modulation properties of a single male speaker with silent intervals limited to 250 ms (Wagener *et al.*, 2006). The same noise condition was used in a study of Meyer and Brand (2013) with 113 listeners (of whom the 99 listeners considered here are a subgroup). They considered three extensions of the Speech Intelligibility Index (SII) for predicting SRT in stationary and fluctuating noise: A) original SII, B) considering frequency-independent level fluctuation of the noise, C) considering frequency-dependent level fluctuations of the noise, and D) considering frequency- dependent fluctuations of the speech and the noise.

RESULTS AND DISCUSSION

Audiogram-based predictions without suprathreshold distortions

Figure 1 shows the simulated SRTs for the 7 typical audiograms for flat and moderately sloping hearing loss defined by Bisgaard *et al.* (2010) as a function of the level of the stationary, test specific noise (solid lines). The simulations for the remaining 3 typical audiograms are not shown here to preserve the separability across curves. In general, the curves follow the pattern proposed by Plomp (1978) who separated an "Attenuation" component (A) from a "Distortion" component (D) of the hearing loss to derive the SRT as a function of noise level (NL). A power-law additivity parameter P was also introduced here to better reflect the fluctuating noise condition:



Pro-file	Stationary A	D	P	Fluctuating A	D	P
None	0	0	11.6	0	0	0.3
N1	11	0	0.9	9	1	0.3
N2	21	0	0.7	19	1	0.2
N3	36	1	0.7	34	2	0.2
N4	56	0	0.8	54	3	0.3
N5	69	0	0.6	62	7	0.4
N6	77	2	0.7	67	13	0.5
N7	90	4	0.8	80	15	0.7

Fig. 1: Speech recognition thresholds (SRTs) for the German matrix sentence test in the test-specific, stationary noise condition as a function of the noise level from simulations with FADE (solid lines). The curves correspond to different grades of hearing impairment based on the seven standard audiograms for flat and moderately sloping hearing loss from Bisgaard *et al.* (2010). The dashed lines show the same results for the fluctuating ICRA5-250 noise. The embedded table reports the attenuation (A) and distortion (D) components (in dB) and the power coefficient P of the best-fitting Plomp curves.

$$SRT_{Plomp}=10\log_{10}\left(10^{\frac{(A+D)*P}{10}}+10^{\frac{(NL+D)*P}{10}}\right)/P \qquad \text{(Eq. 1)}$$

For a given hearing loss, the SRT in quiet is dominated by A+D (horizontal part of the curves). With increasing noise level NL, a transition region (controlled by P) occurs until a constant SNR at SRT is achieved across a wide range of noise levels which reflects the D-value. The A-, D- and P- values fitted to the simulated curves using the Plomp (1978) formula for the different typical audiograms are given in the insert table in Fig. 1. Note that most of the variation across the typical audiograms are captured by the variation in the "Attenuation" component, whereas only the more severe hearing losses require an additional "Distortion" component which also reflects some deviation of the audiogram shape from the standard speech spectrum.

To test the non-individualized SRT predictions based on the audiogram alone (i.e., without suprathreshold processing impairment), Fig. 2A displays the predictions from the "typical" audiograms in Fig. 1 for the individual SRT in stationary ICRA1-noise. The SRT predictions obtained by interpolating across the 10 prototype audiograms

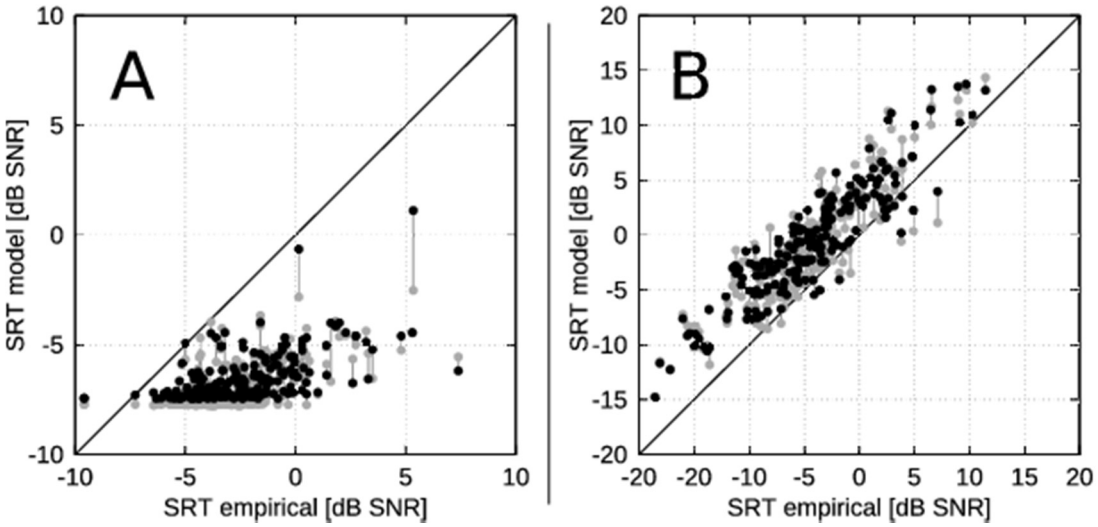
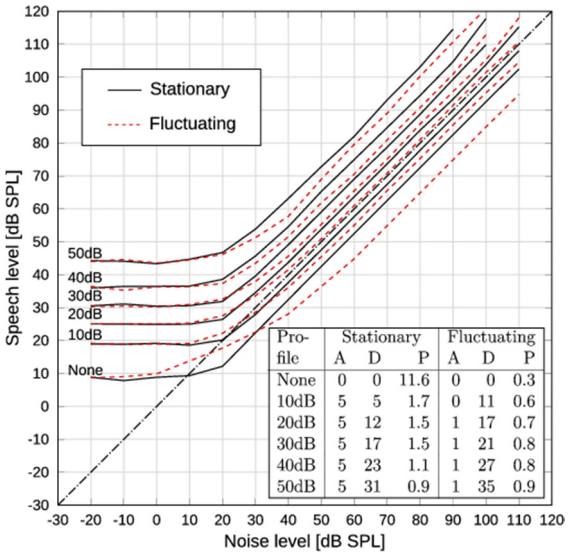


Fig. 2: Modelled speech recognition thresholds (SRTs) for 198 ears from 99 subjects plotted against the empirical data (x-axis). Panel A: Stationary noise, non-individualized predicted SRT (black dots) obtained from the respective best-fitting typical audiogram compared to the individually simulated SRT (grey dots) using the FADE approach with the individual audiogram data. Panel B: Fluctuating noise, predicted SRT from typical audiogram (black dots) vs. individually simulated (grey dots) taking into account the individual level uncertainty u_L estimated from the stationary noise condition.

(black dots) are plotted against the empirical values (given on the x-axis). For comparison, the individualized FADE simulations are given as grey symbols using the individual audiogram. The connection lines between the predicted values (that require only a very small computational load) and the simulated values (that are computationally expensive) indicate already a high coincidence in SRT prediction between both methods. However, neither method is able to model the empirical SRT in stationary noise in a satisfactory way since the large spread in the empirical data (ranging from −9 dB to + 7dB in SNR) is not reflected in the predictions based on the audiogram alone.

Modelling suprathreshold distortion as level uncertainty

Figure 3 displays the simulated SRT using the FADE approach for a normal audiogram with a set of fixed "level uncertainty parameter" u_L-values in order to model an increasing amount of supra-threshold distortions. Note that the curves exhibit a parallel shift to higher SRT values with increasing parameter u_L which is very similar to the effect of the D-parameter of the Plomp model. However, an increase by 10 dB in the level uncertainty parameter u_L does not translate directly into an equally-spaced increase of the D-parameter fitted to the curves in Fig. 3 (see inlaid table in Fig. 3): At low and high u_L-values the largest resulting difference in D for a 10-dB step in u_L is observed, whereas in the midrange the simulations exhibit a higher robustness against an increase in level uncertainty.



Pro-file	Stationary A	Stationary D	Stationary P	Fluctuating A	Fluctuating D	Fluctuating P
None	0	0	11.6	0	0	0.3
10dB	5	5	1.7	0	11	0.6
20dB	5	12	1.5	1	17	0.7
30dB	5	17	1.5	1	21	0.8
40dB	5	23	1.1	1	27	0.8
50dB	5	31	0.9	1	35	0.9

Fig. 3: Speech recognition thresholds (SRTs) in the test-specific, stationary noise condition as a function of the noise level from simulations with FADE for different values of level uncertainty u_L. The dashed lines show the corresponding results for the fluctuating ICRA5-250 noise. The embedded table reports the attenuation (A) and distortion (D) components (in dB) and the power coefficient P of the best-fitting Plomp curves according to Eq. 1.

Combining individualization of audiogram and suprathreshold distortion correction

To assess the effect of including a distortion correction based on estimates of the level uncertainty parameter u_L into the modelling of the SRT data, Table 1 shows the correlation coefficients (Pearson's R^2) between modelled SRTs for stationary and fluctuating noise and the empirical data. *Predictions* indicate an interpolation method based on computations for the 10 typical audiograms only, while *simulations* refer to computations performed for each individual audiogram. The individual suprathreshold distortion effect was not individually computed with the whole FADE approach, but rather estimated in two ways:

- For the "typical" estimate of the level uncertainty parameter u_L, a group of at least 5 and up to 32 listeners, characterized by the same "typical" audiogram, was considered. Their deviation between prediction and empirical SRT was averaged either for the stationary or for the fluctuating noise. This deviation in SNR was converted into a u_L value using the relation shown in Fig. 3, thus leading to the "typical stationary noise-based" or "typical fluctuating noise-based" individualization of u_L. The predicted or simulated SRT was obtained as before, but corrected by an appropriate SRT shift read out from the respective curve in Fig. 3.
- The "individual" estimate of u_L was determined from the individual deviation between modelled and empirical result in the stationary noise condition and then used to correct for suprathreshold distortions in fluctuating noise and vice versa. Note that estimating the "typical" u_L values from the stationary or fluctuating

noise condition provides approximately the same prediction accuracy in both cases (i.e., both for predicting the stationary and the fluctuating noise data) as indicated by the very similar R^2. This suggests that the individual distortion effect is estimated to be very similar for both types of noises – which is a desired property for a universally applicable parameter characterizing the impaired individuals performance. Using the "typical" audiogram and distortion correction already outperforms the SII prediction accuracy for the fluctuating noise case. In the stationary noise case, the individual distortion correction is required to outperform the SII predictions.

		Stationary noise			Fluctuating noise		
Model	Distortion correction	R^2	B	RMS	R^2	B	RMS
FADE prediction	none	0.31	-4.1	4.6	0.48	-4.6	6.1
	typical stat.-based	0.44	0.0	1.9	0.57	3.7	5.4
	typical fluc.-based	0.42	-1.9	2.7	0.56	0.0	3.8
	individual stat.-based	-	-	-	0.78	3.4	4.3
	individual fluc.-based	0.63	-1.6	2.3	-	-	-
FADE simulation	none	0.36	-4.3	4.7	0.57	-4.5	5.9
	typical stat.-based	0.49	-0.1	1.8	0.63	3.8	5.2
	typical fluc.-based	0.46	-2.0	2.7	0.63	0.1	3.5
	individual stat.-based	-	-	-	0.83	3.8	4.5
	individual fluc.-based	0.70	-1.9	2.4	-	-	-
SII version A		0.55	-	-	0.24	-	-
SII version B		0.59	-	-	0.42	-	-
SII version C		0.51	-	-	0.42	-	-
SII version D		0.35	-	-	0.52	-	-

Table 1: Statistical analysis of the predicted/simulated speech recognition thresholds (SRTs). Pearson's correlation coefficients (R^2) are reported along with the root-mean-square (RMS) prediction error and the bias (B) for predicted/simulated SRTs with different distortion correction methods and SII-based predictions from Meyer and Brand (2013).

Overall, the highest prediction and simulation accuracy is achieved if not typical parameter sets, but individualized audiogram and u_L values are employed: Fig. 2B shows the individually modelled SRT in fluctuating noise using the individually obtained u_L estimates from the stationary noise condition either predicted from the typical audiogram data (black dots) or individually simulated (grey dots). The graph demonstrates the high prediction accuracy observed for the individualized suprathreshold distortion parameter u_L even if not an individualized, but typical audiogram is used.

CONCLUSIONS

The ASR-based, reference-free FADE approach can be used as a theoretical counterpart of the empirical Plomp (1978) model to quantitatively assess the effect of hearing impairment on SRTs in stationary and fluctuating noise.

Suprathreshold processing deficiencies can be modelled by the level uncertainty parameter u_L which should be individually determined for a high prediction accuracy.

The prediction accuracy achieved (expressed by Pearson's R^2) is much higher than the prediction accuracy achieved with modified and optimized SII-based measures (e.g., data presented by Mayer and Brand, 2013).

Hence, the FADE approach is not only more versatile and makes much less assumptions than the SII, but also yields much higher prediction accuracy.

ACKNOWLEDGMENTS

Supported by DFG (SFB/TRR 31 & EXC Hearing4all).

REFERENCES

ANSI (**1997**). *S3.5 Methods for Calculation of the Speech Intelligibility Index.* Standards Secretariat, Acoustical Society of America.

Bisgaard, N., Vlaming, M.S., and Dahlquist, M. (**2010**) "Standard audiograms for the IEC 60118-15 measurement procedure," Trends Amplif., **14**, 113-120.

Brand,T. and Kollmeier, B. (**2002**). "Vorhersage der Sprachverständlichkeit in Ruhe und im Störgeräusch aufgrund des Reintonaudiogramms," Proc. 5. Jahrestagung der Deutschen Gesellschaft für Audiologie, Zürich.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation: I. Detection and masking with narrow band carrier," J. Acoust. Soc. Am., **102**, 2892-2905.

Holube, I. and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am., **100**, 1703-1716.

Jürgens, T. and Brand, T. (**2009**). "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," J. Acoust. Soc. Am., **126**, 2635-2648.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M., Uslar, V.N., Brand, T., and Wagener, K.C. (**2015**). "The multilingual matrix test: principles, applications and comparison across languages - a review," Int. J. Audiol. Suppl., **54**, 3-16.

Meyer, R.M. and Brand, T. (**2013**). "Comparison of different short-term Speech Intelligibility Index procedures in fluctuating noise for listeners with normal and impaired hearing," Acta Acust. United Ac., **99**, 442-446.

Plomp, R. (**1978**). "Auditory handicap of hearing impairment and the limited benefit of hearing aids," J. Acoust. Soc. Am., **63**, 533-549.

Schädler, M., Warzybok, A., Hochmuth, S., and Kollmeier, B. (**2015**). "Matrix sentence intelligibility prediction using an automatic speech recognition system," Int. J. Audiol. Suppl., **54**, 100-107.

Schädler, M.R., Warzybok, A., Ewert, S.F., and Kollmeier, B. (**2016**). "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception," Submitted.

Wagener, K.C., Brand, T., Kollmeier, B. (**2006**). "The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners," Int. J. Audiol., **45**, 26-33.