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. © The Authors. 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:

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

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

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

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

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

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).

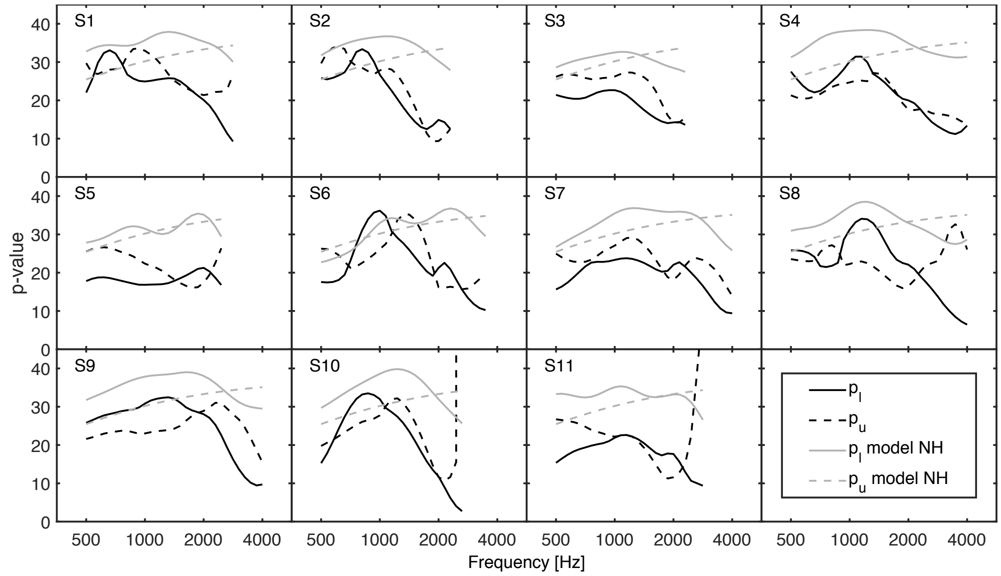


Fig. 1: Black lines show estimated values of p_l (solid lines) and p_u (dashed lines). Grey lines show model predictions for normal-hearing subjects.

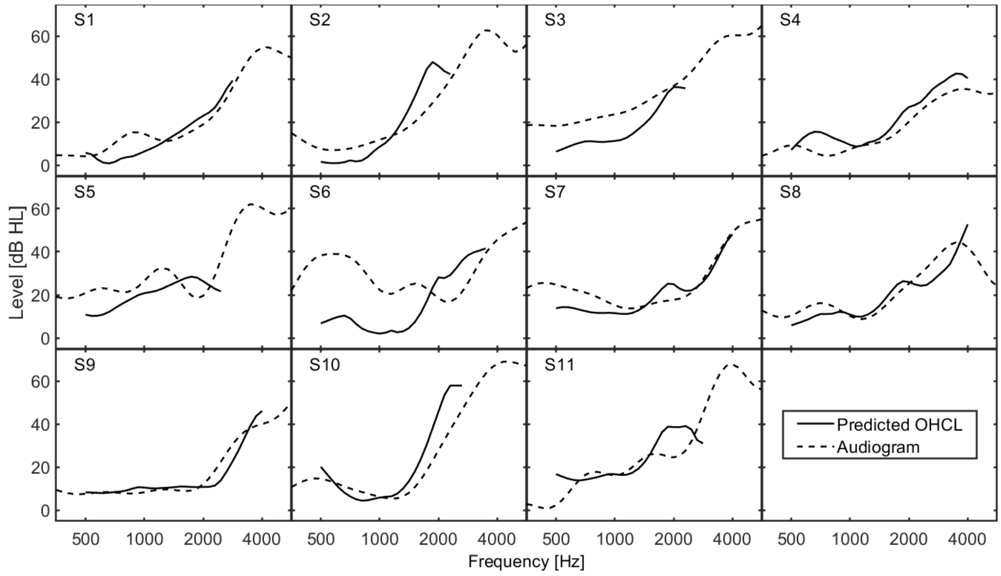
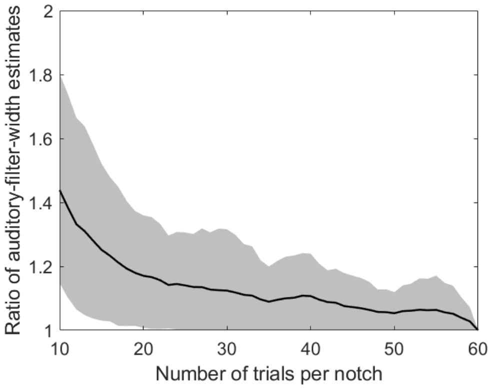


Fig. 2: Solid lines show OHCL values derived from p_l and p_u using the model of Moore and Glasberg (2004). Dashed lines show the audiometric thresholds.

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.



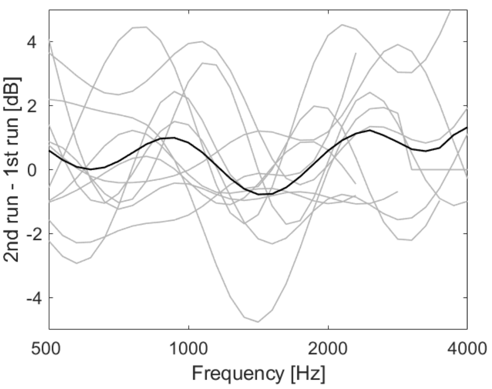


Fig. 3 (left): Ratio between estimated auditory-filter width after *n* trials per notch and the final estimate, plotted as a function of *n*. The inverse was taken if the ratio was smaller than 1. The solid line shows the geometric mean across subjects and the grey area shows the geometric standard deviation.

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.

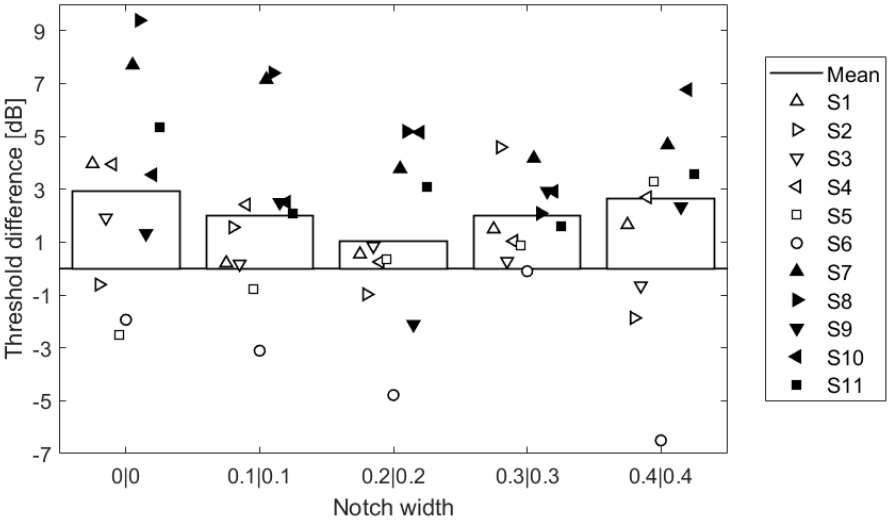


Fig. 5: Difference between the thresholds at 1.4 kHz obtained using the 2-up/1-down procedure and the BAL method for the five symmetric notches. Bars show the mean across subjects and symbols show individual results.

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.