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.

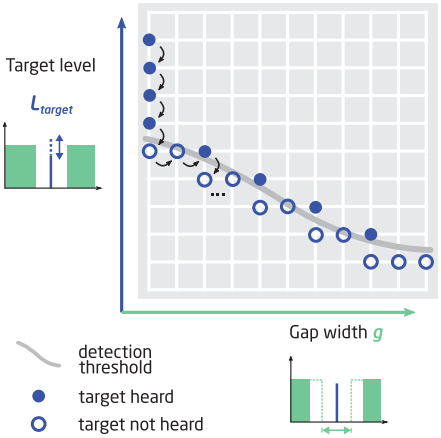

Fig. 1: Illustration of one run of the grid method. The track starts with a moderately high target level at zero gap width. Level is decreased until threshold is reached, after which gap is increased until the target can be heard again.

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.

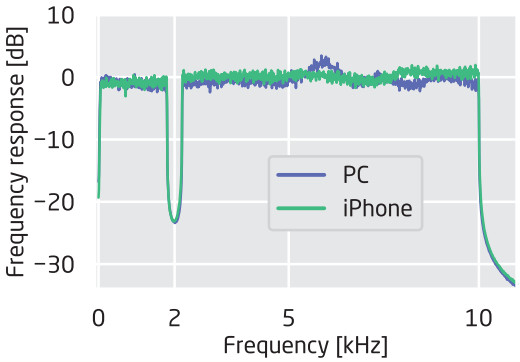

Fig. 2: Two spectra of a noise band (10–10000 Hz) with a spectral gap of 0.1 at 2 kHz, presented via the two different playback systems after compensating for differences in frequency responses by inverse filtering.

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

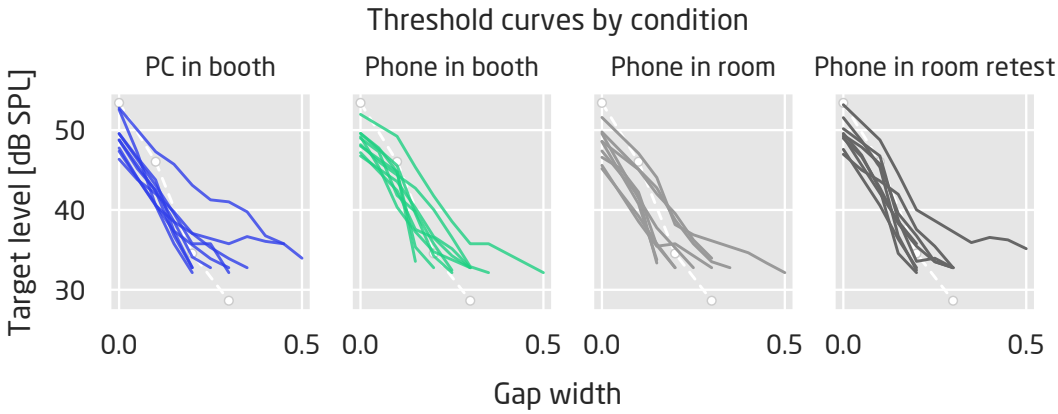

Fig. 3: The lines show the individual threshold curves. The white dashed line is the same in all figures and shows data from Rosen and Baker (1994)

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.

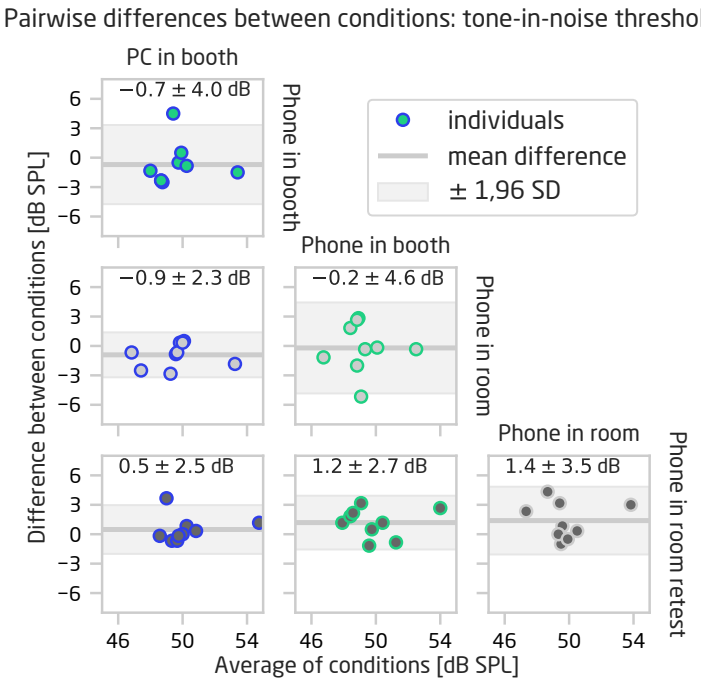

Fig. 4: Pairwise Bland–Altman plots visualizing test–retest repeatability for tone-in–noise thresholds (threshold at zero gap) between two conditions. Horizontal line shows the average difference in thresholds between two conditions, and the shaded area illustrates the 95% limits of agreement (mean ±1.96 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.

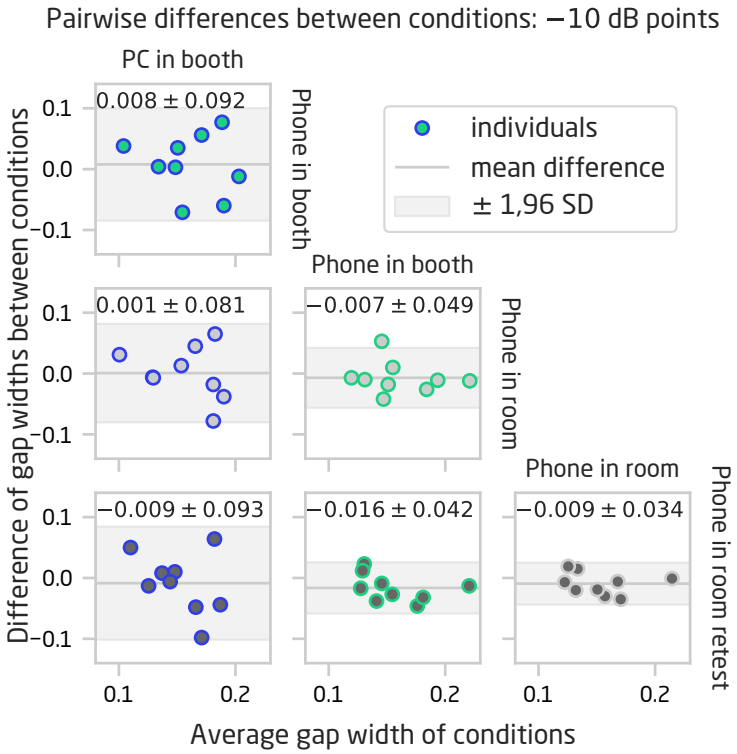

Fig. 5: Pairwise Bland–Altman plots visualizing test–retest repeatability for the −10 dB gap width, i.e. the interpolated gap width at which the threshold is 10 dB lower than at zero gap. Horizontal line shows the average difference in −10 dB points between two conditions, and the shaded area illustrates 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.