DTU Library

Evaluation of a Fast Method to Measure High-Frequency Audiometry Based on Bayesian Learning

Casolani, Chiara; Borhan-Azad, Ali; Sørensen, Rikke Skovhøj; Schlittenlacher, Josef; Epp, Bastian

[Link back to DTU Orbit](#)

Original Article

Evaluation of a Fast Method to Measure High-Frequency Audiometry Based on Bayesian Learning

**Chiara Casolani[1], Ali Borhan-Azad[1], Rikke Skovhøj Sørensen[1],
Josef Schlittenlacher[2], and Bastian Epp[1]** (iD)

Abstract
This study aimed to assess the validity of a high-frequency audiometry tool based on Bayesian learning to provide a reliable, repeatable, automatic, and fast test to clinics. The study involved 85 people (138 ears) who had their high-frequency thresholds measured with three tests: standard audiometry (SA), alternative forced choice (AFC)-based algorithm, and Bayesian active (BA) learning-based algorithm. The results showed median differences within ± 5 dB up to 10 kHz when comparing the BA with the other two tests, and median differences within ± 10 dB at higher frequencies. The variability increased from lower to higher frequencies. The BA showed lower thresholds compared to the SA at the majority of the frequencies. The results of the different tests were consistent across groups (age, hearing loss, and tinnitus). The data for the BA showed high test–retest reliability (>90%). The time required for the BA was shorter than for the AFC (4 min vs. 13 min). The data suggest that the BA test for high-frequency audiometry could be a good candidate for clinical screening. It would add reliable and significant information without adding too much time to the visit.

Introduction

Audiometry is the first test performed when screening a patient for potential hearing loss or hearing disorders. Standard audiometry (SA) is usually performed in the interval between 125 and 8000 Hz in steps of 1 octave or 0.5 octave. However, the audible frequencies for human hearing span the interval between 20 and 20,000 Hz (Purves et al., 2001). Previous research suggested that frequencies above 8000 Hz can add valuable information for early detection of hearing loss (Wang et al., 2021). For example, when hearing loss is driven by ototoxic drugs, an early detection and change in the drug prescription can avoid the spread of the damage (Chauhan et al., 2011). High-frequency audiometry (HFA) can also be useful in combination with SA to identify severe acoustic traumas and therefore diagnose noise-induced hearing loss at an early stage (Ahmed & Dennis, 2001; Büchler et al., 2012; Mehrparvar et al., 2011). In addition to this, a correlation between high-frequency hearing loss and the strength and laterality of tinnitus was found, suggesting a causality

between the phenomena (Vielsmeier et al., 2015). Despite the reported benefits that the HFA-added information would give, it is not a current practice in the clinics to perform the test. The reason might be related to a time factor that also links to the higher cost of having to spend more time on a single patient. The audiometry test is, in fact, usually performed by an audiologist or a clinician who presents the sound through the audiometer and waits for the feedback of the patient who is instructed to push a button only in case they heard the sound (Bess & Humes, 2008). Increasing evidence shows, however, that assessment of sensitivity at high frequencies might be a proxy for

---

[1]Auditory Physics Group, Hearing Systems section, Department of Health Technology, Technical University of Denmark, Lyngby, Denmark
[2]Speech Hearing and Phonetic Sciences, University College, London, UK

Corresponding Author:
Bastian Epp, Auditory Physics Group, Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Denmark.
Email: bepp@dtu.dk

detection of hidden hearing loss and potentially, more specifically, a synaptic damage in the inner ear (Liberman et al., 2016). These frequencies are the fastest to deteriorate due to the loss of low spontaneous rate fibers (Furman et al., 2013; Kobel et al., 2017).

The goal of this study is to assess an HFA test based on a Bayesian active (BA) learning algorithm (Schlittenlacher et al., 2018) and to compare it to (a) the standard HFA (SA) performed with the audiometer, and (b) an HFA test based on an alternative forced choice (AFC) procedure (Ewert, 2013). BA tests enjoy increasing popularity in the hearing research community and further variants for assessing the audiogram up to 8 kHz have been proposed (Cox & De Vries, 2021; Heisey et al., 2018; Song et al., 2015). There are two aspects that we want to investigate: (i) the test–retest reliability for each of BA, AFC, and SA in the frequency range between 8 and 16 kHz, and (ii) to compare the differences in the hearing threshold estimates across all three tests. According to previous studies, the variability of the pure tone audiometry (PTA) tends to increase for the high frequencies. The error stays, however, within ± 10 dB for more than 90% of the population for each frequency tested (Frank, 2001; Schmuziger et al., 2004). Studies in children (aged 4–13 years old) showed similar results (Beahan et al., 2012).

The advantage of an automated test (e.g. BA or AFC) over a test requiring an additional operator is that the test progresses automatically based on the input from the participant. Nonautomated tests such as SA depend on an additional operator to modify the relevant parameters. The screening procedure in a clinic could, therefore, be optimized to free time for additional care of the participant. Another advantage of the BA over AFC or SA is the increased ability of BA to capture potential details of the audiogram. An evaluation of discrete frequencies is not sensitive to variations in between two frequencies. The BA selects the frequency increment based on an information criterion which increases the sensitivity to variability in narrow frequency intervals. In principle, AFC and SA could provide similar information, but at the cost of increased measurement time. Based on the underlying criterion to reduce uncertainty, the result of this test could then provide an, in principle, continuous and optimal estimate of the audiometric threshold within any selected frequency interval.

The data of this study will provide the required information if, and to which extent, BA-based methods can replace or supplement current adaptive methods to measure audiometry. If application of the BA-based method proves useful, it will allow for more extensive data collection in both research and audiology without the need for extensive additional time resources.

Materials and Methods

Three tests were performed: (a) standard manual HFA with an AC40 audiometer (Interacoustics) that will be referred

to as "SA", (2) HFA with an AFC procedure (Ewert, 2013) that will be referred to as "AFC." Finally, (c) the HFA based on a BA-learning algorithm (adaptation for a high frequency of the software developed for standard frequencies by Schlittenlacher et al., 2018) that will be referred to as "BA." The order of the three tests has been randomized for each participant. Each test was carried out twice in order to assess the test–retest reliability.

Participants

We recruited 85 participants (of age min: 20, max: 78, mean: 42 years), resulting in a total of 138 tested ears for this study. The participants were partially recruited from an existing database, and partially newly recruited. The only recruiting criterion for participants already in our database was a maximum threshold of 60 dB hearing level (dBHL) at 8 kHz to decrease the probability of increased thresholds above 8 kHz. This criterion was not imposed on newly recruited participants without an available standard audiogram (below 8 kHz). For participants with only one ear satisfying the inclusion criterion that ear was tested. For the other participants, the left ear was tested first, followed by the right ear if time permitted (and neither ear was blocked by cerumen). Out of the 85 participants, 52 had normal hearing up to 8 kHz, and 33 showed a threshold higher than 20 dBHL for at least one frequency below or equal to 8 kHz. Before the tests, the participants' ears were examined with otoscopy, and a few general questions about hearing in noise, tinnitus, and hyperacusis were asked to assess their subjective perception of their own hearing. Participants with tinnitus were tested with warble tones in the SA and AFC tests (see Table 1 for distributions of the stimuli presented to the listeners).

To reveal the potential effect of age within the participant group, participants were separated into groups "younger than 40" and "40 and Over." This divider is close to the mean age of the participant group. Given the heterogeneity of the participants, this comparison will only reveal any potential

Table 1. Stimuli Used in the Tests for the Ears Tested Included in the Experiments and Location of the Test.

	Test and stimulus (tone/warble)			Location	
	SA	AFC	BA	DTU	BBH
All	110/28	110/28	138	110	28
Under 40	71	71	71	59	12
40 and Over	67	67	67	51	16
NH	84	84	84	70	14
HL	54	54	54	40	14

Note. SA = standard audiometry; AFC = alternative forced choice; BA = Bayesian active; DTU = Technical University of Denmark; BBH = Bispebjerg Hospital. The accumulated entries for test and stimulus ("All") denote the ratio of pure tones/warble tones.

strong effects of age, but will not be sensitive to any specific aspects of aging.

All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

Apparatus

The experiments were performed at two locations (Technical University of Denmark [DTU]; and Bispebjerg Hospital [BBH], DK) and with two different apparatus. The SA was performed with a clinical audiometer (Interacoustics, AC40) and headphones (Radioear, DD450). The AFC and BA experiments were performed in both locations with a portable setup. The portable consisted of a soundcard (RME Fireface UCX), a custom-made headphone amplifier optimized for listening experiments (Sonible hpa2), a portable PC running Windows (HP Zbook), and a pair of headphones (Sennheiser HDA200). The portable setup was calibrated each day before testing.

Tests

The SA and the AFC were performed using pure tones with frequencies of 8, 9, 10, 11.2, 12.5, 14 and 16 kHz. For listeners suffering from tinnitus, pure tones were replaced by warble tones with a frequency modulation of 5%. For the SA, a 1-up-2-down, ascending method was used by the operator where two responses out of three presentations at a single level were needed to determine the threshold (Hughson–Westlake procedure; see also ASHA, 2005). The BA was applied in the same frequency interval between 8 and 16 kHz, but without predefined intermediate frequencies. All tests were performed in a soundproof booth. In the SA, the participant was supposed to press a button whenever a sound was audible. In the AFC, the participant was asked to choose the interval out of three that they believed contained the target sound (and to guess at random if they could not hear it). This was done by clicking on the "1," "2," or "3" button in a graphical user interface implemented in MATLAB (MATLAB, 2021). In the BA, the participant was supposed to answer either "Yes" or "No" to the question "Did you hear the tone?" This test was also implemented in MATLAB (MATLAB, 2021) using the GPML toolbox (Rasmussen & Nickisch, 2010).

The AFC followed an adaptive 1-up 2-down rule. The threshold was determined after four reversals. The initial step size was 8 dB. After each upper reversal, the step size was decreased to 4 dB and finally to 2 dB. The last four reversals at the smallest step size were averaged to determine the threshold. The starting level was set to 70 dB sound pressure level (SPL) and was increased in case the participant chose the wrong interval until a maximum level of 80 dB SPL. All signals were generated with a duration of 0.5 s including 25 ms raised-cosine ramps at onset and offset with a sampling frequency of 48 kHz.

The BA consisted of 50 trials. Each trial contained three tone pulses. Three pulses were chosen to decrease the chance for the participant to confuse it with a potential steady tinnitus. Each pulse had a duration of 250 ms with a rise/fall time of 20 ms and 100 ms silence between pulses. All signals were generated with a sampling frequency of 48 kHz. The first tone was presented at 60 dB SPL at 12 kHz. The maximum level was 80 dB SPL. To quantify bias effects due to lapses in attention, five silent trials (in addition to the 50 trials) were randomly presented.

The starting level for each test was based on piloting data. The selected levels provided a clearly audible signal in most cases. An impact of this difference in the starting levels cannot be excluded but is unlikely because of the applied adaptive procedure and the rather small difference in starting levels.

The BA was based on Schlittenlacher et al. (2018), but with a length scale of 0.5 octave. The length scale defines how much an estimate is affected by data gathered at adjacent frequencies, and is used in the kernel of the underlying Gaussian process (for details, see Schlittenlacher et al. 2018). In summary, the algorithm uses the available data to determine the frequency/level combination that maximally reduces uncertainty. Following the first tone, frequencies with a distance of 0.25 octave are tested in order to obtain an initial estimate of the threshold at these frequencies. At each frequency, at least one positive ("I heard the tone") and one negative answer ("I did not hear the tone") were collected. Thereafter, a continuous function is fitted using a Gaussian process. The fit is updated after each trial to provide a probabilistic estimate of the detection probability for each frequency and level. The BA aims to maximize the information in each trial to cover the desired spectral range. In the final step, it determines the threshold is the level where the detection probability is 50%. The use of an initial grid is not mandatory, but makes the algorithm more robust, and overall, the algorithm leads to a reduction in measurement time (Casolani et al., 2020).

It should be noted that the three tests target different points on the psychometric function for the detection threshold. The SA defines the threshold as the lowest level heard after a bracketing procedure controlled by the operator of the audiometer. With the procedure used in this study (two responses out of three presentations at a single level needed to determine the threshold), this targets an off-center point above 50% on the psychometric function. Other common procedures used in SA lead to a 50% point. While there might be a small difference, the expected difference can be assumed small in the context of clinically relevant precision. The AFC with the used 1-up-2-down procedure targets the 71% point on the psychometric function. Hence, this might lead to a higher threshold estimate than that of the standard test, in addition to systemic biases between methods, which together amounted to about 2 dB between BA and SA up to 8 kHz in Schlittenlacher et al. (2018). The BA returns the 50% contour across the evaluated frequency interval.

All tests were performed two times for each frequency and ear. The average of these two runs was used for data analysis. The data from two ears of one participant were treated as independent data sets. Some listeners fulfilled the inclusion criterion only for one ear. This resulted in a number of data points lower than twice the number of participants.

Statistical Evaluation

A Kruskal–Wallis test included in MATLAB (MATLAB, 2021) was applied to evaluate potential statistical differences in threshold across groups or methods. Statistical differences between two measurements of the threshold at the same frequency using the same method were evaluated using an equivalency test. This test consisted of a Wilcoxon-signed rank test implemented in MATLAB (Cardillo, 2023). The criterion for the rejection of the null hypothesis (no difference) was based on the calculated Z-value at a significance level of 5%.

Results

Figure 1 shows a boxplot of the difference between the thresholds measured by the BA and SA by frequency. For the BA results, only thresholds under 80 dB SPL, the maximum presentation level, were considered. At all frequencies, the median difference was within ± 10 dB, in agreement with previous studies exploring the PTA variability at high frequencies (Frank, 2001; Schmuziger et al., 2004). At five of the seven test frequencies (8, 9, 10, 12.5, and 14 kHz), the median difference was within ± 5 dB. In general,
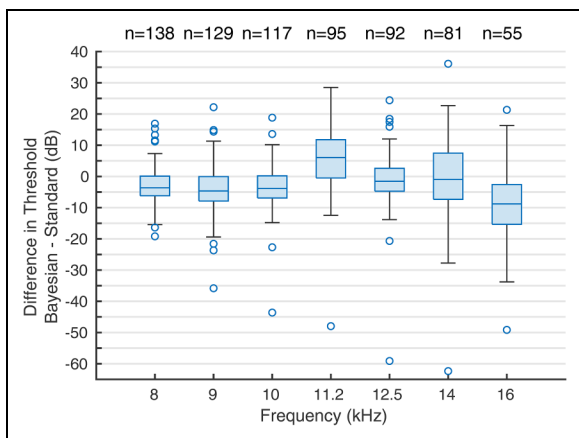
the quantiles were broader for the higher frequencies compared to the lower frequencies. The BA showed lower thresholds compared to the SA at all frequencies except 11.2 kHz.

Figure 2 shows a boxplot of the difference between the thresholds measured by the BA and AFC tests by frequency. Since each measurement was repeated twice, for each ear, the difference between the mean thresholds recorded by each test was considered. As in Figure 1, at all frequencies, the median difference was within ± 10 dB. At five of the seven test frequencies (8, 9, 10, 12.5, and 16 kHz), the difference was within ± 5 dB. The quantiles were broadest at 11.2 and 14 kHz, where also the median of the differences was greater. The BA test returned higher thresholds than the AFC at all frequencies except 9 and 16 kHz. On average, the difference was 2.3 dB, and the standard deviations (*SD*s) were between 4.9 and 9.4 dB. The SDs were lower than those for the differences between the BA and SA.

Figure 3 shows a boxplot of the difference between the thresholds measured by the AFC and SA by frequency. For each ear, the difference between the mean thresholds recorded by each test was considered. As in the previous results, at all frequencies, the median difference was within ± 10 dB. At four of the seven frequencies (8, 9, 12.5, and 16 kHz), it was within ± 5 dB. Frequency had little effect on quantile width. The threshold estimates of the AFC were systematically lower than those of the standard test at all frequencies.

Figure 4 shows a box plot of the difference between the median thresholds measured by the three tests. For each ear, the difference between the median thresholds measured by each test was considered. When comparing the BA to

Figure 1. Boxplot of the difference between the thresholds measured by the BA and SA by frequency. The figure shows the median, upper (0.75), and lower (0.25) quantiles, and the minimum and maximum values that are not outliers. The circles represent outliers. The numbers above the boxes show the number of ears for which a response was measured at each frequency.
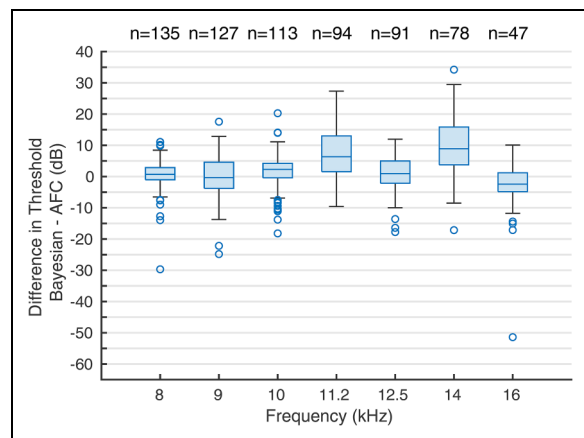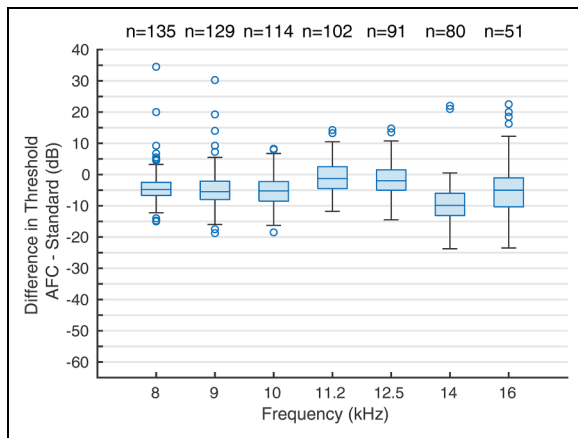*Note.* BA = Bayesian active; SA = standard audiometry.

Figure 2. Boxplot of the difference between the thresholds measured by the BA and AFC tests by frequency. The figure shows the median, upper (0.75), and lower (0.25) quantiles, and the minimum and maximum values that are not outliers. The circles represent outliers. The numbers in the boxes show the number of ears for which a response was measured at each frequency.
Note. BA = Bayesian active; AFC = alternative forced choice.

Figure 3. Boxplot of the difference between the thresholds measured by the AFC and SA by frequency. The figure shows the median, upper (0.75), and lower (0.25) quantiles, and the minimum and maximum values that are not outliers. The circles represent outliers. The numbers in the boxes show the number of ears for which a response was measured at each frequency. *Note.* AFC = alternative forced choice; SA = standard audiometry.

the SA, the median difference was within ±5 dB for all groups except *40 and Over* and *Hearing Loss*, for which it was within ±11 dB. Comparing the BA to the AFC, the median difference was within ±5 dB for all groups. Comparing the AFC to the SA, the median difference was within ±5 dB only for the groups *Under 40* and *Normal Hearing*. Comparing both the Bayesian and AFC to the SA, the quantiles were narrower for the groups *Under 40* and *Normal Hearing* than they were for *40 and Over* and *Hearing Loss*. Tinnitus had a lesser effect on quantile width. In general, the quantiles for the BA and AFC comparison were narrower than those for the two comparisons to the SA. Additionally, the differences were more consistent.

For all groups, the BA and AFC tests estimated lower thresholds than the SA. When only considering the underlying psychometric functions, the AFC would be expected to show higher thresholds compared to the BA and SA tests. The BA test estimated higher thresholds than the AFC. Based on the psychometric function, the AFC would be expected to have a higher threshold than the BA, but possibly compensated for by previously reported systemic bias in BA.

Table 2 shows the difference between the first and second trials of all tests for each frequency as percentages that were within ±5, ±10, and >± 10 dB. Percentages tended to be lower for higher frequencies, consistent with previous studies' frequencies (Frank, 2001; Schmuziger et al., 2004). Percentages were generally highest for the SA, and lowest for the BA.

Figure 5 shows a box plot of the difference between the first and second trials of the BA test. The medians were all within ±1.5 dB. No significant differences were found ($p >$ .05 for all frequencies, Kruskal–Wallis test, see also

Table 3). The p-values were not corrected for multiple testing. The quantiles were narrow, with a slight broadening toward the highest frequencies. This suggests high test–retest reliability. Figures 6 and 7 show box plots of the differences between the first and second trials of the AFC and the SA for comparison. Table 4 shows the intraclass correlation coefficients (ICCs) for each test, overall and split by age, hearing status, and tinnitus status. For all tests and groups, the ICC was above 0.90, except for the *Under 40* and *NH* groups in the BA test, for which it was above 0.85. Even though the ICC was lower for these groups, the BA shows high reliability in general. The ICCs for the AFC and SA were all above 0.90, indicating that these tests were also highly reliable. The ICC for SA was comparable with previously found values in a comparison of self-administered audiometry (Bastianelli et al., 2019).

A Kruskal–Wallis test was used to investigate if there was a significant difference in threshold between the different tests SA, BA, and AFC. First, the analysis was performed with a focus on the listener. Listeners were pooled into one group ("All") or split into different groups ("Under 40," "40 and Over," "NH," "HL," "No Tinnitus," and "Tinnitus"). Then it was evaluated if the variable "test" had a significant influence on the median threshold across frequency. If the p-value was above the threshold, pairwise comparisons were made for each pair of the values of the "test." Secondly, the analysis was performed with a focus on frequency. Here, all listeners were pooled. It was evaluated if the variable "test" had an influence on the threshold at each frequency. If the p-value was above the threshold, pairwise comparisons were made for each pair of the values of the "test." Bonferroni correction was applied to the initial threshold for the p-value of .05 to correct for multiple comparisons (seven listener groups and seven frequencies). Hence, the final criterion for significance was $\alpha = .05/14 = .0036$.

Table 5 shows the results of the statistical evaluation. Values in bold indicate significance. A significant influence of "test" was found for the groups "40 and Over" and "HL." The test result revealed differences between thresholds measured by BA versus SA and thresholds measured by AFC versus SA. It is interesting to note that the members of these groups largely coincide. No difference was found for the threshold measured with AFC and BA.

Also, a significant influence of "test" was found for most of the frequencies. Only at 8 kHz no influence of "test" was found. And for the pairwise comparison of BA and AFC, only the threshold at 14 kHz was different, while for the other comparisons, all (AFC–SA) or all but one (BA–SA) frequencies showed significantly different thresholds.

In summary, the statistical tests indicate a bias in the measured threshold dependent on the test used. This result is in line with previous results. In practice, such a bias can easily be compensated for once known with proper precision.
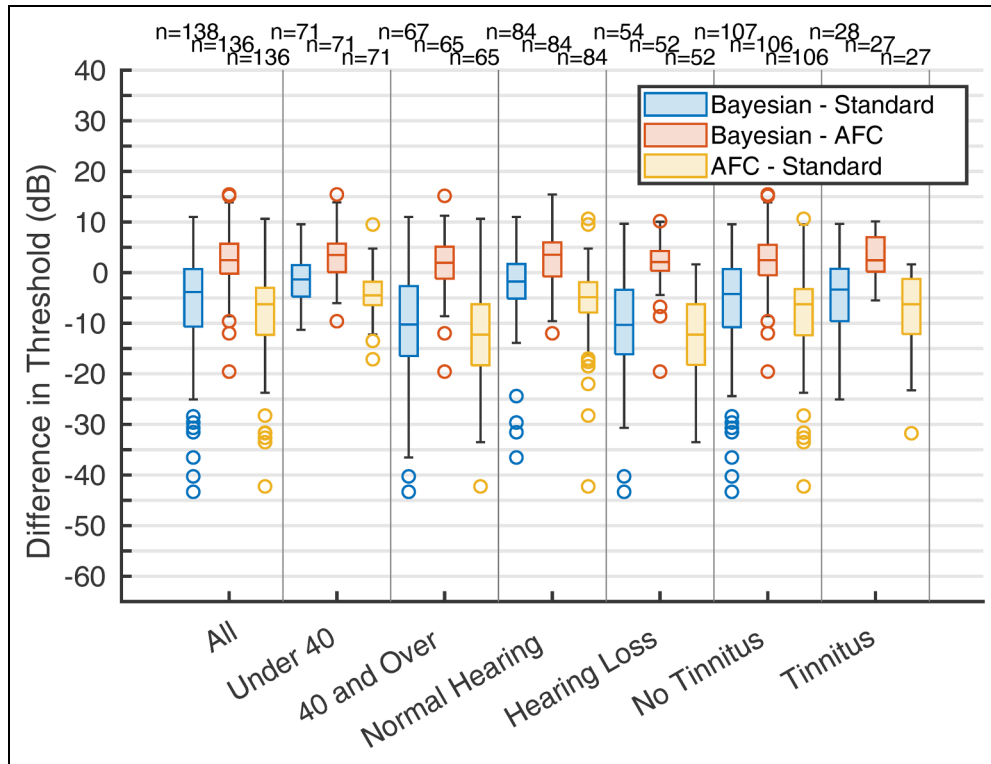
Figure 4. Boxplot of the median difference between the thresholds measured by the three tests. The result for the whole population is shown, alongside the population split by age, hearing status, and tinnitus status. The figure shows the median upper (0.75) and lower (0.25) quantiles, and the minimum and maximum values that are not outliers. The circles represent outliers. The numbers in the boxes show the number of ears in each group for which a response was measured.

Table 2. Percentage of Threshold Differences Between First and Second Session Collapsed Over all Ears That Were Within ±5 dB, ±10 dB, and >±10 dB at Each Frequency.

	Frequency (kHz)						
BA							
Threshold difference	8.0	9.0	10.0	11.2	12.5	14	16
±5 dB	83	85	80.9	81.5	77.8	74.7	63
±10 dB	95.6	94.5	92.2	93.5	93.3	89.3	89.1
>±10 dB	4.4	5.5	7.8	6.5	6.7	10.7	10.9
AFC							
Threshold difference	8.0	9.0	10.0	11.2	12.5	14	16
±5 dB	89.9	86	85.6	88.2	88.6	83.3	85.4
±10 dB	98.4	98.3	98.2	98	100	93.6	97.9
>±10 dB	1.6	1.7	1.8	2	0	6.4	2.1
SA							
Threshold difference	8.0	9.0	10.0	11.2	12.5	14	16
±5 dB	97.8	97.1	96.4	97.8	97	97.4	96
±10 dB	99.3	98.6	99.3	99.3	98.5	100	96
>±10 dB	0.7	1.4	0.7	0.7	1.5	0	4

Note. BA = Bayesian active; AFC = alternative forced choice; SA = standard audiometry. Only ears were included where a threshold could be measured in both sessions.

Discussion

The thresholds estimated by each of the three tests were overall similar when compared both across frequencies and across different listener groups. When considering the difference between each participant's thresholds at each frequency, the median differences between tests were nearly all within ±10 dB. This is comparable to the variability of PTA at high frequencies shown in previous studies (Frank, 2001; Schmuziger et al., 2004). The majority of the differences were within ±5 dB, well under this variability and equal to what is considered the variability for SA below 8 kHz. This indicated that both the AFC and the BA are reasonably comparable to the SA. The fraction of listeners where no threshold could be measured in the automated procedures (AFC and BA) was, however, higher than for the SA.

On an individual level, each participant's median threshold measured by BA and AFC is within ±5 dB for all groups. This shows great similarity and hence interchangeability of the two methods. When dividing the results according to different listener groups (age, hearing loss, and tinnitus), both the BA and AFC show differences in the range ±10 dB for all groups except *40 and Over* and
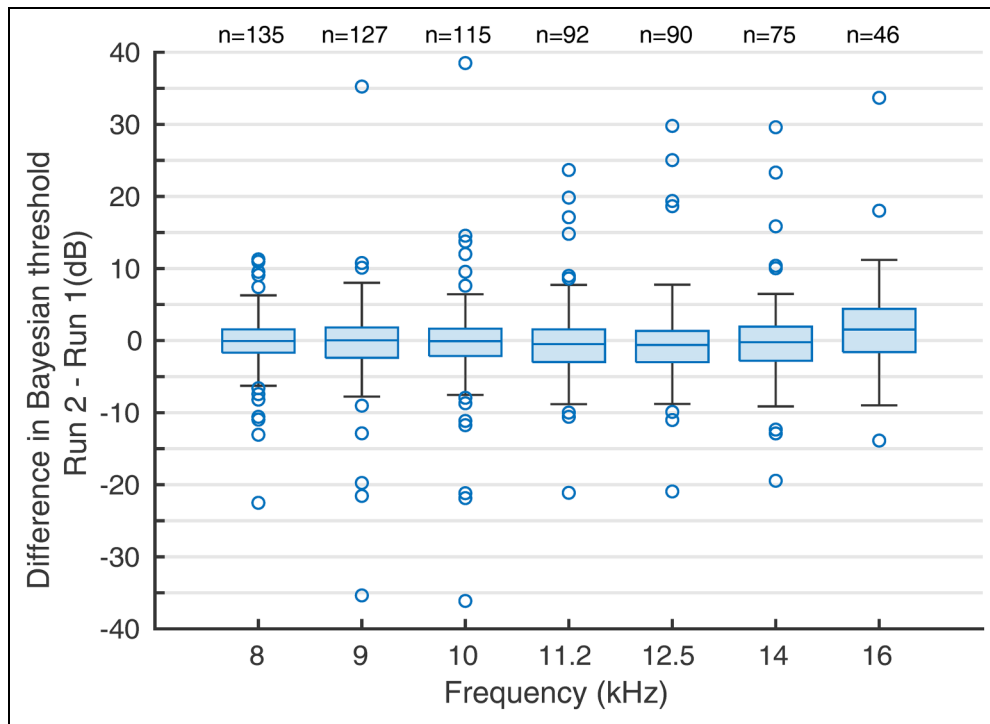
Figure 5. Boxplot of the median difference between the thresholds measured in the first and second runs of the Bayesian active (BA) test. The result for the whole population is shown, alongside the population split by age, hearing status, and tinnitus status. The figure shows the median upper (0.75) and lower (0.25) quantiles, and the minimum and maximum values that are not outliers. The circles represent outliers.

Table 3. Statistical results of the difference between the first and second measurements of the Bayesian active (BA) test (Kruskal–Wallis).

BA

Frequency (kHz)	8.0	9.0	10.0	11.2	12.5	14	16
z-value	0.14	0.44	0.64	1.32	1.45	0.77	1.54
P-value (two-tail)	.88	.66	.52	.19	.15	.44	.12
Mean of differences	−0.07	0.03	−0.08	−0.50	−0.60	−0.22	1.54
Confidence interval	[−4.95, 4.55]	[−4.07, 3.97]	[−5.44, 4.77]	[−4.50, 3.69]	[−4.96, 3.94]	[−4.53, 5.63]	[−4.45, 5.41]

Hearing Loss relative to the SA. For these two groups, the difference falls outside of ± 10 dB. To decide which of these methods is more useful in diagnosis, these results should be compared to objective measures of hearing aid fitting or the analysis of the variability in outcome measures aimed to quantify the effects of "hidden hearing loss" (Liberman et al., 2016). The differences to the gold standard of SA indicate an interesting source of information along this direction. There are a number of outliers in the results of the AFC and the BA. None of the listeners reported that they lost focus during the experiment, but several listeners mentioned that the experiments became tiring. This might be a factor of the relatively small number, but notable outliers of up to 60 dB. One might speculate that monitoring the listeners' vigilance might reduce the number of outliers. It might also be possible to automatically detect and account for trials in the procedure where the level adjustments are likely due to the lack of vigilance of the listener.

Overall, the two automated methods (BA and AFC) provide lower threshold estimates compared to the SA. One possible explanation for this result is that the definition of threshold differs between tests. The SA and the BA reflect 50% of the psychometric function, while the AFC measures the 71% point. However, with this consideration, the AFC should provide a higher threshold than the SA and the BA. In clinical practice, this difference might, however, be irrelevant since it only amounts to 1 dB up to 8 kHz (Leek, 2001). Another potential source for bias is the paradigm underlying the different methods. The SA used a go/nogo task in the SA, and the BA and the AFC are forced-choice paradigms. This might lead to differences in the slopes of the underlying psychometric functions for the different methods. One more
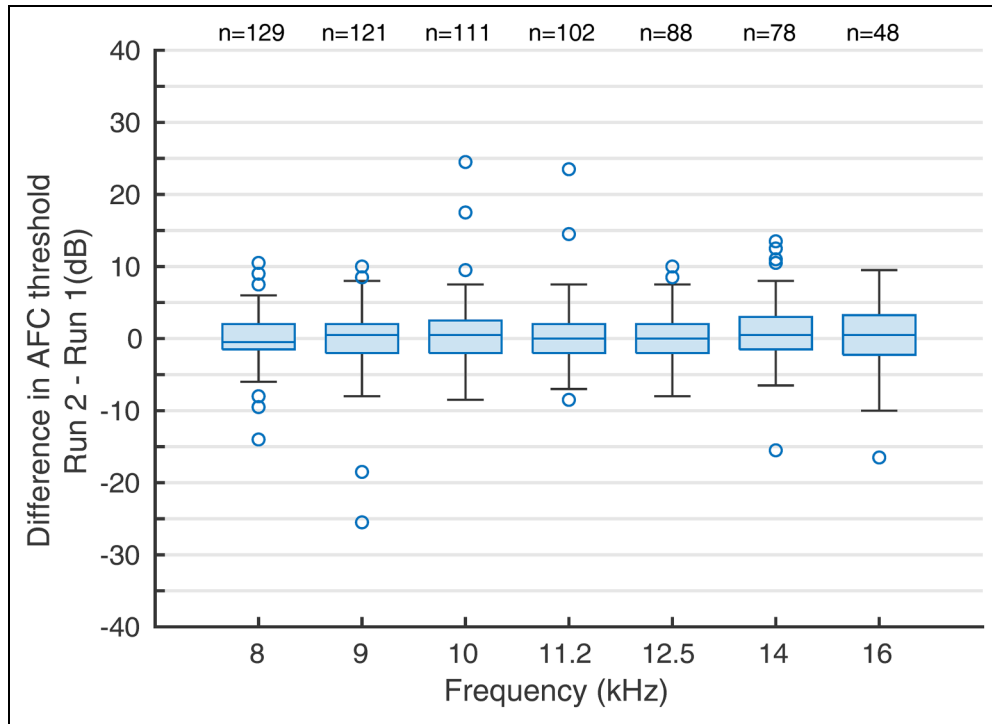
Figure 6. Boxplot of the median difference between the thresholds measured in the first and second runs of the alternative forced choice (AFC) test. The result for the whole population is shown, alongside the population split by age, hearing status, and tinnitus status. The figure shows the median upper (0.75) and lower (0.25) quantiles, and the minimum and maximum values that are not outliers. The circles represent outliers.
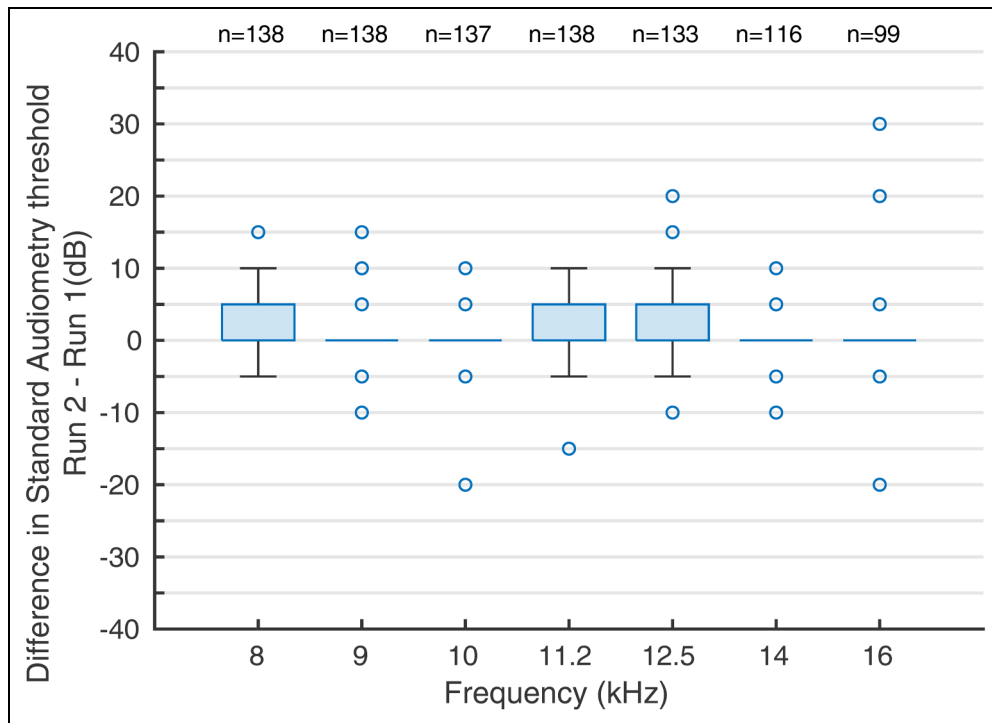
Figure 7. Boxplot of the median difference between the thresholds measured in the first and second runs of the standard audiometry (SA) test. The result for the whole population is shown, alongside the population split by age, hearing status, and tinnitus status. The figure shows the median upper (0.75) and lower (0.25) quantiles, and the minimum and maximum values that are not outliers. The circles represent outliers.

Table 4. Test–Retest Intraclass Correlation Coefficients.

	BA	AFC	SA
All	0.9361	0.9830	0.9931
Under 40	0.8816	0.9671	0.9635
40 and Over	0.9373	0.9630	0.9907
NH	0.8644	0.9396	0.9791
HL	0.9502	0.9748	0.9868
No tinnitus	0.9533	0.9804	0.9926
Tinnitus	0.9641	0.9779	0.9899

Note. BA = Bayesian active; AFC = alternative forced choice; SA = standard audiometry.

Table 5. Test Comparison *P*-Values.

	All 3	BA–SA	BA–AFC	AFC–SA
Listener groups				
All	.0074	—	—	—
Under 40	.0065	—	—	—
40 and Over	**<.001**	**<.001**	.3591	**<.001**
NH	.0194	—	—	—
HL	**<.001**	**<.001**	.3119	**<.001**
No tinnitus	.0190	—	—	—
Tinnitus	.0952	—	—	—
Frequency (kHz)				
8	.0236	—	—	—
9	.0012	**<.001**	.8945	**.0024**
10	**<.001**	**<.001**	.3629	**<.001**
11.2	**<.001**	.0557	.0309	**<.001**
12.5	**<.001**	**<.001**	.5805	**<.001**
14	**<.001**	**.0012**	**<.001**	**<.001**
16	**<.001**	**<.001**	.9583	**<.001**

Note. BA = Bayesian active; AFC = alternative forced choice; SA = standard audiometry. The individual columns show the statistical effects when considering all methods ("All 3"), or the difference between the results obtained with the two methods.
Values in bold face indicate significant differences.

source for these differences might be deviations in the calibration method and the headphones. Both systems were calibrated using a coupler measurement system: the setup for the BA and the AFC on a daily basis, while the clinical equipment followed a standard clinical calibration interval. The contribution of this effect is presumably small and can be compensated for if required.

In the current study, no measures were taken to fixate the listeners' heads to the headphones to ensure constant headphone placement within and across tests. Headphone placement can have an impact on the results (e.g., Almeida et al., 2015). Such and similar factors increasing the variability in the data make it harder to evaluate the impact of the procedure to measure thresholds in quiet only. The current data therefore likely also includes some effects of headphone placement or movement. Because these differences are in the dimensions of differences between operator- and participant-placed headphones in audiometry (Almeida et al., 2015), this

might indicate an impact of the procedure which is smaller than, or of similar size to, headphone placement. Headphone placement is also a present factor in clinical practice. Hence, the variability in the present data might reflect variability expected in experiments with noncontrolled placement of headphones, as well as variability in ear canal acoustics.

The present study used different stimuli for the different listener groups. Listeners with tinnitus were tested with warble tones in the SA and AFC tests, while listeners without tinnitus were tested with pure tones in the SA and AFC tests. Previous studies showed only small differences when comparing pure tones with warble tones ($\pm 5\%$ and modulation rate of 5 per second) in the frequency range between 250 and 8000 Hz. These differences are likely irrelevant to the shape of the audiogram (Dockum & Robinson, 1975). A comparison between pulsed, continuous, and warble-tone stimuli in HFA showed very similar thresholds for pulsed and continuous tones. Differences were found between warble tones and continuous and pulsed tones. It was suggested that the underlying reason might be the low-frequency contents of a warble tone in the case of sloping hearing loss where the frequency of the highest sensitivity determines the threshold (Hamill & Haas, 1986). These observations might also be reflected in the present study. Methods using pure tones will not be sensitive to any steep local gradients in threshold microstructure. Also, physical factors such as headphone placements can change the transmission characteristics of the headphones and the ear canal and lead to effective steep local gradients in the audiogram. The Bayesian procedure might, however, partially capture such an effect because a steep slope will contain information and hence influence the selection of test frequency in the procedure. In case the thresholds of tinnitus listeners were biased by sweeping over a steep slope in the audiogram, this will be reflected in the standard deviation of the data and the comparison across methods. While the deviation introduced by the differences in stimuli might have increased the variability in the data, accommodating the difficulties of tinnitus listeners to actually do the task, by providing them the cue of a frequency modulation, might have compensated for another potential source of variability. In order to decide on an optimal combination of method and stimulus, additional dimensions of sources of variability need to be investigated in detail.

The rationale behind the use of the stimuli in the present study was to use the same type of stimuli as was used in previous studies using the same methods but in the normal frequency range. Schlittenlacher et al. (2018) used pulsed tones because these are easier to detect than pure tones for listeners with tinnitus. Assuming the same holds true for the extended frequency range considered in the present study, one could expect similar results to Schlittenlacher et al. (2018) for pure tones or warble tones. Schlittenlacher et al. (2018) also compared their Bayesian method with pulsed tones to
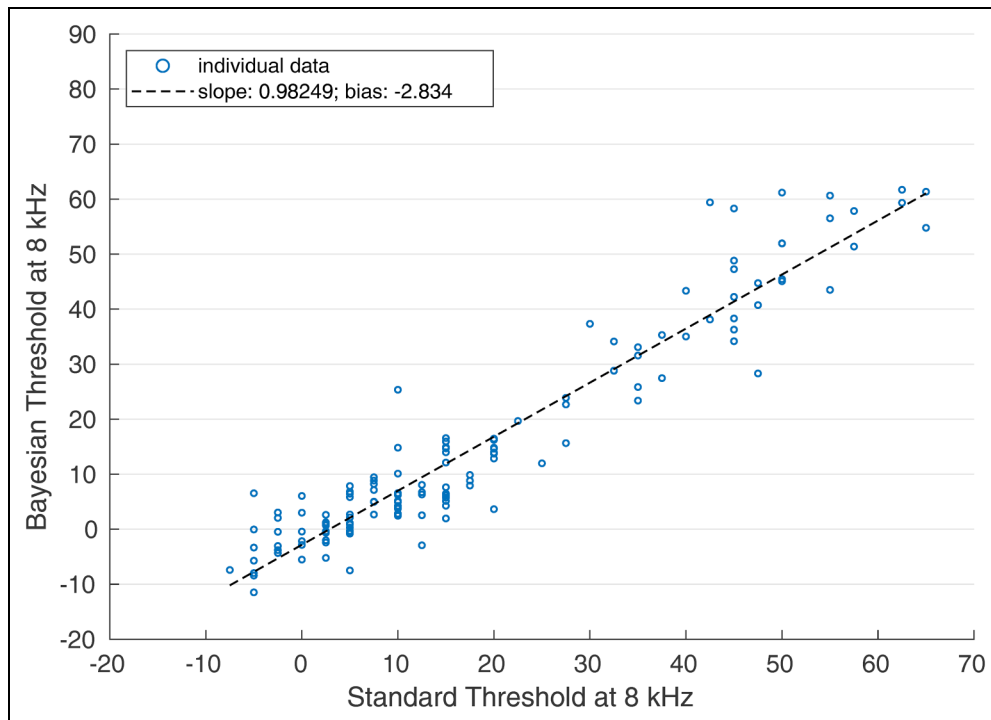
Figure 8. Scatter plot of the threshold at 8 kHz for the BA and SA. The dashed line shows a linear fit.
Note. BA = Bayesian active; SA = standard audiometry.

a manually operated audiometer with steady tones. Most importantly, the discrepancy between the two methods was rather constant across frequencies, which could be considered by a simple subtraction resulting in a definition for the reference of 0 dBHL adjusted to the method or stimuli. The difference was also rather small considering that not only slightly different stimuli were used but, more notably, different methods and different headphones. Hence, while the present study provides some confidence, a direct comparison of stimuli within each of the methods might give proof that the choice of stimuli plays a minor role.

On a statistical level, many significant differences were found in the comparisons between methods. The individual comparisons between the BA and the AFC with the SA show significant differences. The only significant difference between the BA and the AFC was found at 14 kHz. These results might reflect the systematically lower thresholds for the AFC and the BA relative to the SA. The ICC is high for all levels and all groups for all methods. Hence, this indicates a systemic bias separating the SA from the BA and the AFC.

Figure 8 shows a direct comparison of the thresholds at 8 kHz for the BA and the SA. It reveals a close relationship and a bias of about 2.8 dB toward the BA. This bias is comparable to the values reported by Schlittenlacher et al. (2018).

The main difference between these methods can be found in the time required to collect the data. The duration of the SA can vary from operator to operator based on experience and

training. Based on current recommendations about manual PTA (ASHA, 2005), one might expect the audiogram for one ear to take around 6–8 min for an experienced operator excluding preparatory actions such as visual inspection and instructions (8 frequencies, 1–2 s tone duration, and 10 presentations per frequency plus pauses between presentations). Variability induced by human intervention will be absent in automated tests such as the AFC or the BA. The time taken for each participant to complete the BA and the AFC was recorded. The time was not recorded for the SA, due to the large influence of the experimenter as a human factor. The Bayesian test took an average of 3 min and 40 s, while the AFC took an average of 12 min and 53 s. A small number of participants requested a break during the AFC trials. This break time is included in the test duration, as this reflects real-world situations. No listener requested a break during the BA. Hence, the Bayesian method clearly saves a significant amount of time while providing results comparable to those of the AFC. In addition, the BA provides an estimate across the whole interval of interest, while the AFC and the SA only provide a threshold at discrete points. It is, however, important to note that the BA is, such as the AFC and the SA, not sensitive to the presence of potential threshold microstructure which might be present in the frequency range between 8 and 16 kHz (Baiduc et al., 2014). Threshold microstructure is not part of clinical audiological assessment and hence this shortcoming of these three methods can be neglected in this context.

The population of hearing-impaired listeners used in the present study showed, from a clinical perspective, only mild hearing impairment. While a generalization of the findings toward the clinical population as a whole cannot be made, it can be assumed that these findings also hold for more severe hearing impairment. It is, however, possible that cognitive factors can play a role in the performance of automated procedures. These need to be identified and accommodated in the interaction with the participant.

Only looking at duration, the data show a benefit of the BA over SA and AFC to measure an audiogram in the high-frequency region. In addition to the time benefit, the next logical step could be to optimize the parameters of the BA to achieve an optimal balance between speed and precision. One main parameter of the BA is the scale over which the underlying Gaussian process estimates the next frequency step, combined with the initial grid defining the starting points of the procedure. A finer initial grid might provide additional information on a shorter frequency interval. But this benefit might be counteracted by the time required to measure the threshold at the initial grid. The precision in threshold at a given frequency in SA and AFC is mainly given by the chosen step size at which the intensity is varied. It might be possible to also optimize the adjustment of the presented level changes in the BA in order to achieve a precision that is similar to the AFC. Another relevant point that could be captured automatically is the effect of the ear canal acoustics and headphone placement on the results. An additional degree of freedom might identify the potential impact of ear canal acoustics by small variations in frequency, separate those from a potential threshold microstructure, and thereby maximize precision. These aspects of optimization are, however, out of the scope of the present study.

Conclusion

The assessment of the test based on Bayesian learning for audiometry at high frequencies showed a satisfying performance of the tool. The test–retest reliability was consistently high and the difference with the standard method was consistent with the error range for HFA found previously in other studies. Given the very short execution time of the test—on average around 4 min—this test could be a good solution to include the HFA in clinical screening procedures.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

ORCID iD

Bastian Epp ![ORCID] https://orcid.org/0000-0002-8062-0837

References

Ahmed, H. O., & Dennis, J. H. (2001). High-frequency (10–18 kHz) hearing thresholds: Reliability, and effects of age and occupational noise exposure. *Occupational Medicine*, *51*(4), 245–258. https://doi.org/10.1093/occmed/51.4.245

Almeida, B. P., Menezes, P. d. L., de Andrade, K. C. L., & Teixeira, C. F. (2015). Positioning of earphones and variations in auditory thresholds. *Brazilian Journal of Otorhinolaryngology*, *81*(6), 642–646. https://doi.org/10.1016/j.bjorl.2015.08.016

ASHA (2005). Guidelines for manual pure-tone threshold audiometry. https://www.asha.org/policy.

Baiduc, R. R., Lee, J., & Dhar, S. (2014). Spontaneous otoacoustic emissions, threshold microstructure, and psychophysical tuning over a wide frequency range in humans. *Journal of the Acoustical Society of America*, *135*(1), 300–314. https://doi.org/10.1121/1.4840775

Bastianelli, M., Mark, A. E., McAfee, A., Schramm, D., Lefrançois, R., & Bromwich, M. (2019). Adult validation of a self-administered tablet audiometer. *Journal of Otolaryngology – Head & Neck Surgery*, *48*(1), 59. https://doi.org/10.1186/s40463-019-0385-0

Beahan, N., Kei, J., Driscoll, C., Charles, B., & Khan, A. (2012). High-frequency pure-tone audiometry in children: A test–retest reliability study relative to ototoxic criteria. *Ear and Hearing*, *33*(1), 104–111. https://doi.org/10.1097/AUD.0b013e318228a77d

Bess, F. H., & Humes, L. (2008). *Audiology: The fundamentals* (4th ed.). Lippincott Williams & Wilkins.

Büchler, M., Kompis, M., & Hotz, M. A. (2012). Extended frequency range hearing thresholds and otoacoustic emissions in acute acoustic trauma. *Otology and Neurotology*, *33*(8), 1315–1322. https://doi.org/10.1097/MAO.0b013e318263d598

Cardillo, G. (2023). Wilcoxon. https://github.com/dnafinder/wilcoxon

Casolani, C., Harte, J. M., & Epp, B. (2020). Looking for objective correlates between tinnitus and cochlear synaptopathy. In *Proceedings of the International Symposium on Auditory and Audiological Research: Auditory Learning in Biological and Artificial Systems*, Vol. 7, 421–428.

Chauhan, R. S., Saxena, R. K., & Varshey, S. (2011). The role of ultrahigh-frequency audiometry in the early detection of systemic drug-induced hearing loss. *Ear, Nose and Throat Journal*, *90*(5), 218. https://doi.org/10.1177/014556131109000506

Cox, M., & De Vries, B. (2021). Bayesian pure-tone audiometry through active learning under informed priors. *Frontiers in Digital Health*, *3*, 723348. https://doi.org/10.3389/fdgth.2021.723348

Dockum, G. D., & Robinson, D. O. (1975). Warble tone as an audiometric stimulus. *Journal of Speech and Hearing Disorders*, *40*(3), 351–356. https://doi.org/10.1044/jshd.4003.351

Ewert, S. (2013). AFC – A modular framework for running psychoacoustic experiments and computational perception models. In *Proceedings of the international conference on acoustics AIA-DAGA 2013, Merano, Italy* (pp. 1326–1329). Deutsche Gesellschaft für Akustik (DEGA).

Frank, T. (2001). High-frequency (8 to 16 kHz) reference thresholds and intrasubject threshold variability relative to ototoxicity criteria using a Sennheiser HDA 200 earphone. *Ear and Hearing*, *22*(2), 161–168. https://doi.org/10.1097/00003446-200104000-00009

Furman, A. C., Kujawa, S. G., & Liberman, M. C. (2013). Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates. *Journal of Neurophysiology*, *110*(3), 577–586. https://doi.org/10.1152/jn.00164.2013

Hamill, T. A., & Haas, W. H. (1986). The relationship of pulsed, continuous, and warble extended-high frequency thresholds. *Journal of Communication Disorders*, *19*(3), 227–235. https://doi.org/10.1016/0021-9924(86)90012-2

Heisey, K. L., Buchbinder, J. M., & Barbour, D. L. (2018). Concurrent bilateral audiometric inference. *Acta Acustica United with Acustica*, *104*(5), 762–765. https://doi.org/10.3813/AAA.919218

Kobel, M., Le Prell, C. G., Liu, J., Hawks, J. W., & Bao, J. (2017). Noise-induced cochlear synaptopathy: Past findings and future studies. *Hearing Research*, *349*(Sp. Iss. SI), 148–154. https://doi.org/10.1016/j.heares.2016.12.008

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception and Psychophysics*, *63*(8), 1279–1292. https://doi.org/10.3758/BF03194543

Liberman, M. C., Epstein, M. J., Cleveland, S. S., Wang, H., & Maison, S. F. (2016). Toward a differential diagnosis of hidden hearing loss in humans. *PLoS One*, *11*(9), e0162726. https://doi.org/10.1371/journal.pone.0162726

MATLAB (2021). *9.11.0.1769968 (r2021b)*. The MathWorks Inc.

Mehrparvar, A. H., Mirmohammadi, S. J., Ghoreyshi, A., Mollasadeghi, A., & Loukzadeh, Z. (2011). High-frequency audiometry: A means for early diagnosis of noise-induced hearing loss. *Noise and Health*, *13*(55), 402–406. https://doi.org/10.4103/1463-1741.90295

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., & Williams, S. M. (2001). *Neuroscience* (2nd ed.). Sinauer Associates.

Rasmussen, C. E., & Nickisch, H. (2010). Gaussian processes for machine learning (GPML) toolbox. *The Journal of Machine Learning Research*, *11*, 3011–3015.

Schlittenlacher, J., Turner, R. E., & Moore, B. C. (2018). Audiogram estimation using Bayesian active learning. *The Journal of the Acoustical Society of America*, *144*(1), 421–430. https://doi.org/10.1121/1.5047436

Schmuziger, N., Probst, R., & Smurzynski, J. (2004). Test–retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Resarch ER-2 earphones. *Ear and Hearing*, *25*(2), 127–132. https://doi.org/10.1097/01.AUD.0000120361.87401.C8

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., & Barbour, D. L. (2015). Fast, continuous audiogram estimation using machine learning. *Ear & Hearing*, *36*(6), e326–e335. https://doi.org/10.1097/AUD.0000000000000186

Vielsmeier, V., Lehner, A., Strutz, J., Steffens, T., Kreuzer, P. M., Schecklmann, M., & Kleinjung, T. (2015). The relevance of the high frequency audiometry in tinnitus patients with normal hearing in conventional pure-tone audiometry. *BioMed Research International*, *2015*, 302515. https://doi.org/10.1155/2015/302515

Wang, M., Ai, Y., Han, Y., Fan, Z., Shi, P., & Wang, H. (2021). Extended high-frequency audiometry in healthy adults with different age groups. *Journal of Otolaryngology – Head and Neck Surgery*, *50*(1), 52. https://doi.org/10.1186/s40463-021-00534-w