


Performance and reliability evaluation of an improved machine learning-based pure-tone audiometry with automated masking

Nicolas Wallaert^{1,2} | Antoine Perry^{2,3} | Sandra Quarino¹ | Hadrien Jean² |
Gwenaëlle Creff¹ | Benoit Godey¹ | Nihaad Paraouty² 

¹Department of Otorhinolaryngology–Head and Neck Surgery, Rennes University Hospital, Rennes, France

²R&D Department, My Medical Assistant SAS, Reims, France

³Laboratoire d'Informatique Signal et Image, Electronique et Télécommunications, ISEP Ecole d'ingénieurs du Numérique, Paris, France

Correspondence

Nihaad Paraouty, R&D Department, My Medical Assistant SAS, 5 Bis Cours Anatole France, 51100 Reims, France.
Email: paraouty@iaudiogram.com

Funding information

French National i-Nov Grant,
Grant/Award Number: DOS0127610/00;
Region Grand Est Deeptech-BPI France Grant

Abstract

Objective: Automated air-conduction pure-tone audiograms through Bayesian estimation and machine learning (ML) classification have recently been proposed in the literature. Although such ML-based audiometry approaches represent a significant addition to the field, they remain unsuited for daily clinical settings, in particular for listeners with asymmetric or conductive hearing loss, severe hearing loss, or cochlear dead zones. The goal here is to expand on previously proposed ML approaches and assess the performance of this improved ML audiometry for a large sample of listeners with a wide range of hearing status.

Methods: First, we describe the changes made to the ML method through the addition of: (1) safety limits to test listeners with a wide range of hearing status, (2) transient responses to cater for cochlear dead zones or nonmeasurable thresholds, and importantly, (3) automated contralateral masking to test listeners with asymmetric or conductive hearing loss. Next, we compared the performance of this improved ML audiometry with conventional and manual audiometry in a large cohort ($n = 109$ subjects) of both normal-hearing and hearing-impaired listeners.

Results: Our results showed that for all audiometric frequencies tested, no significant difference was found between hearing thresholds obtained using manual audiometry on a clinical audiometer as compared to both the manual and automated improved ML methods. Furthermore, the test–retest difference was not significant with the automated improved ML method for each audiometric frequency tested. Finally, when examining cross-clinic reliability measures, significant differences were found for most audiometric frequencies tested.

Conclusions: Together, our results validate the use of this improved ML-based method in adult clinical tests for air-conduction audiometry.

KEYWORDS

audiometry, automated test, Bayesian learning, contralateral masking, psychoacoustics

Abbreviations: GP, Gaussian process; ML, machine learning.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *World Journal of Otorhinolaryngology - Head and Neck Surgery* published by John Wiley & Sons Ltd on behalf of Chinese Medical Association.

INTRODUCTION

Nearly 20% of today's world population live with hearing loss (HL). Yet, hearing disorders remain widely undiagnosed and untreated, leading to central neural deficits, ranging from language acquisition impairments in children¹ to dementia and depression in older populations.^{2,3} In fact, hearing loss represents a major health and societal challenge, with substantial economic impact. One of the main reasons that hearing loss remains underdiagnosed today in developed countries is due to a lack of available time for hearing healthcare professionals and a limited number of trained hearing specialists: ENT doctors and audiologists. As recently pointed out by Lesica et al.,⁴ our hearing healthcare system still relies on labor-intensive procedures and this is particularly true for the hearing examination process, namely the conduction of audiometry tests. Although recent technological innovation in clinical applications through artificial intelligence and machine learning (ML) is starting to have a widespread impact in different clinical settings, including the detection of retinal diseases^{5,6} and medical imaging,^{7,8} they remain rarely applied in the hearing healthcare system (e.g., speech audiometry tests,^{9,10} audiogram interpretation,¹¹ and see Wasmann et al.¹²).

In most countries, clinical hearing assessment relies on one primary measure, pure-tone audiometry, carried out manually by an expert health practitioner.¹³ Pure-tone audiometry consists of estimating detection thresholds of pure tones presented in silence and at different frequencies, hence providing a pure-tone audiogram for each ear. The commonly used procedure is based on a modified version of the Hughson–Westlake procedure.^{14,15} Precisely, this measurement is performed frequency by frequency, with a first sound at a given frequency being presented to the listener at an audible intensity level. The intensity is then reduced in fixed increments until the listener no longer responds, that is, the listener no longer hears the stimuli. The intensity is then increased by a smaller increment until the listener responds again. This procedure is then repeated for several such reversals for each frequency and each ear.^{16–19}

To save time, automated measures²⁰ of pure-tone audiometry have been proposed. In fact, some automated methods are currently commercialized, for example, the Automated Method for Testing Auditory Sensitivity (AMTAS) (Grason-Stadler, Interacoustics).^{21,22} However, most commercialized automated methods have similar drawbacks as the manual conventional method: (1) they measure the audiogram for a discrete number of frequencies (sparse sampling may lead to incomplete hearing assessment at untested frequencies), (2) they use multiple tones presented at one frequency (leading to predictable tone detection), and (3) they use a minimum of 5 dB intensity steps (higher precision may improve hearing aid fittings).

To address those issues, several authors have recently described novel ways of measuring pure-tone audiometry through the use of ML.¹² Such ML approaches (e.g., the Machine Learning Audiogram [MLAG]) are based on both active sampling methods to rapidly estimate audiometric thresholds^{23,24} and probabilistic classification.²⁵ Such ML-based audiometry provides continuous audiogram thresholds in frequency, with 1-dB precision and uses test sounds that are not predictable. More recently, an online version of the MLAG has

also been developed and validated,²⁶ in addition to measures of bilateral audiograms in both ears simultaneously,²⁷ to further reduce testing time. In parallel, Schlittenlacher et al.²⁸ extended the work of Gardner et al.²³ by evaluating two distinct methods for continuous audiogram threshold estimation: Yes/No and Counting of tone pulses. The authors²⁸ also incorporated an omission rate to account for the listener's imperfect responses (i.e., misses and false alarms) leading to improved threshold estimates.

To ensure the safe and adequate use of such ML-based audiometry in routine clinical tests for adult listeners with hearing loss of different etiologies, including conductive, sensorineural, or mixed hearing loss of all degrees of severity from mild to severe losses, the current work extends the ML approach described in Schlittenlacher et al.²⁸ More precisely, we implemented the following key changes:

1. the addition of transient positive and negative responses to constrain the audiometry test phase, to safely test listeners with cochlear dead zones and severe losses (previously unaddressed),
2. the addition of automated contralateral masking rules to better account for air-bone gaps (ABGs) than what has been previously described,²⁹ and
3. general safety limits during automated testing to cater to a wide range of normal-hearing (NH) and hearing-impaired (HI) listeners in routine clinical settings.

Here, in the first study, we present the ML-based approach developed by Schlittenlacher et al.,²⁸ the Yes/No ML method), to measure pure-tone thresholds and highlight the different changes we applied in the development of our improved ML-based method in a computer software. Next, the performance of this improved ML audiometry with both manual and automated pure-tone audiometry tests was compared to conventional, well-established manual audiometry measures ("Reference") in a large cohort of both NH and HI listeners.

We also assessed the test-retest reliability of the automated and improved ML audiometry when run twice on the same group of subjects. For comparison, we also examined test-retest reliability in cross-clinical settings. Although no time gain is demonstrated for the listeners with the use of the automated and improved ML-based approach as compared to manual conventional audiometry measures, test automation allows key time gain for health professionals who no longer need to perform the audiometry tests. Overall, these results validate the use of our automated and improved audiometry approach in daily clinical settings.

STUDY 1: PERFORMANCE OF AN IMPROVED ML-BASED AUDIOMETRY

Methods

Subjects

The subjects ($n = 109$) tested here were of a wide age range (see Table 1) and were recruited from an audiology clinic in France. Only

TABLE 1 Age, gender, and hearing loss distribution of subjects.

Subject distribution	
Gender	
Male	<i>n</i> = 53 male (49%)
Mean ± 1 SD Age (Minimum-Maximum Age)	65.8 ± 14.3 years (23–87 years)
Female	<i>n</i> = 56 female (51%)
	60.2 ± 17.7 years (18–95 years)
All	<i>n</i> = 109 subjects
	62.9 ± 16.3 years (18–95 years)
Hearing status	
NH	<i>n</i> = 32 ears (15%)
	16 female, 16 male
	48.2 ± 18.1 years (21–75 years)
HI	<i>n</i> = 186 ears (85%)
	96 female, 90 male
	42.2 ± 15.3 years (18–95 years)
Hearing loss severity	
Mild hearing loss (PTA between 21 and 40 dB)	<i>n</i> = 100 ears (46%)
	49 female, 51 male
	63.6 ± 11.7 years (18–85 years)
Moderate hearing loss (PTA between 41 and 70 dB)	<i>n</i> = 78 ears (36%)
	42 female, 36 male
	67.8 ± 17.8 years (18–95 years)
Severe hearing loss (PTA between 71 and 90 dB)	<i>n</i> = 8 ears (4%)
	5 female, 3 male
	66.1 ± 6.9 years (55–72 years)
Hearing loss type	
Conductive hearing loss	4.3% of HI subjects
Mixed hearing loss	12.9% of HI subjects (with 5.4% asymmetric loss)
Sensorineural hearing loss	82.8% of HI subjects (with 20.4% asymmetric loss)

Abbreviations: HI, hearing impaired; ML, machine learning; NH, normal hearing; PTA, Pure-Tone Average.

adults were tested (>18 years) and all subjects spoke French fluently. No exclusion criteria based on the etiology of hearing loss was used, in order to not exclude any type of hearing loss. No subjects were excluded from the study. To have an overall balanced population of subjects, the past hearing status of subjects (when available) were used to ensure recruitment of subjects with different hearing loss severities. All subjects were fully informed of the goal of the study and provided written consent before their participation. The study was approved by the French Regional Ethics Committee (Comité de Protection des Personnes Est III; SI number: 22.03364.000107).

Lateralization test

Prior to the audiometry tests, participants were asked to indicate on which side they had the best hearing. Next, all subjects were tested

using a manual procedure, to test their lateralization with a Weber test. Subjects were equipped with an ossi-vibrator positioned on the forehead position and stimulated with pulsed pure tones at four frequencies: 0.5, 1, 2, and 4 kHz with a 1 s-long stimulus. For each frequency, once the experimenter obtains the first positive response from the subject, the intensity is increased by 15 dB to establish the lateralization. To do so, subjects are asked to say which ear the sound came from. The lateralization results for each frequency are stored manually in the software interface by the experimenter, and the results are used to automatically compute contralateral masking levels for air-conduction pure-tone audiometry (see Methods Section "Automated contralateral masking during automated audiometry procedure").

Audiometry tests

Next, for all 109 subjects, hearing thresholds were measured for both ears using three methods as follows:

1. Manual Reference audiometer,
2. Manual improved ML audiometer, and
3. Automated improved ML audiometer.

The order of presentation of the three tests was randomized. For each subject, pure-tone audiometry was measured for individual ears with either manual masking (for "Manual Reference" and "Manual improved ML-audiometer," see Methods section "Manual Audiometry") or an automated masking procedure (for "Automated improved ML-audiometer," see Methods section "Automated ML-Audiometry Procedure") presented to the nontest ear.

Material and calibration

All testing took place in an audiometric booth. The Reference audiometer used was a Natus–Otometrics Astera II diagnostic audiometer equipped with Sennheiser HDA 200 headphones mounted on Peltor earmuffs.

The second audiometer used was a computer software developed by My Medical Assistant SAS (iAudiogram). All stimuli were generated at a sampling frequency of 44.1 kHz and a resolution of 24 bits. The digital-to-analog conversion was performed by an audio interface without acoustic attenuation. Test stimuli, as well as contralateral masking stimuli, were presented via TDH 39 headphones mounted on Peltor earmuffs.

For the lateralization test, we used the Reference audiometer equipped with a Radioear B71 ossi-vibrator. Calibration was performed for both devices by a Natus specialist technician, in accordance with EN ISO 389-8:2004³⁰ and IEC 60318-1:2009,³¹ using a Brüel & Kjær 4153 coupler, a Brüel & Kjær 0843 adaptor, a Brüel & Kjær 0304 cone, and a Brüel & Kjær Artificial Mastoid Type 4930. The sound pressure level was measured with a Brüel & Kjær 2250

sound level meter. The sound pressure levels measured at the coupler were converted into dB HL using EN ISO 389-1:2017³² and 389-8:2004,³⁰ specifying the reference thresholds for the specific headphones used.

ML-based audiometry system and interface

The improved ML audiometer is a computer-based software (iAudiogram) intended to be used by hearing professionals to obtain diagnostic-quality air-conduction pure-tone audiometric data. Audiometry can be performed both in a manual and in a fully automated manner. Adult listeners of both genders may be tested, except for subjects who are unable to cooperate due to age or other conditions such as attentional disorders, or inability to follow instructions.

For all audiometry tests, the experimenter places the subject in a test booth and instructs the subject to click on the remote-control button when they hear a pure-tone sound in the test ear as shown on the subject's monitor. Subjects are also told that noise may be presented to the nontest ear, but they should focus on the pure tones and respond when they hear the pure tones in the test ear. Subjects are also informed that pure tones may be presented in a random manner with silent intervals in between and are reminded that their task is to simply click the response button to let the experimenter know when they hear the pure tones. Following these oral instructions, the experimenter equips the subject with the headphones and launches the automated test before leaving the test booth. The experimenter can view the subject's responses on a second monitor placed outside the test booth and may pause the test or restart it if the subject has questions.

The ML-audiometer software was developed in Python and includes an experimenter interface and a subject interface in French language (see Supporting Information S1: Figure 1). On the experimenter interface, ML-based audiograms can be displayed as continuous curves or transformed into discrete data points. For comparison purposes for the current study, the thresholds of all ML-based audiograms were discretized to 11 conventional audiometry frequencies.

Manual audiometry procedure

Manual pure-tone audiometry test systematically begins by testing the better ear declared by the subject, and if no better ear is declared, the right ear is always tested first. The manual audiometry procedure tests audiometry frequencies of 1, 1.5, 2, 3, 4, 6, 8, 0.75, 0.5, 0.25, and 0.125 kHz in the given order as recommended in French audiometry guidelines.³³ The intensity level varies in 5 (up) and 10 dB steps (down), also referred to as an asymmetric up-down procedure.³⁴ The experimenter adjusts the frequency and level directly from the conventional audiometer or the software until the threshold is obtained. The subject's response when he/she presses the button to indicate that he/she hears appears directly on the experimenter interface. The experimenter saves the audiometric thresholds directly on the interface. The duration of the stimuli for each trial is defined

by the experimenter and varies according to general audiometry recommendations.^{18,33}

Manual masking rules applied

The manual pure-tone audiometry test was performed following a set of contralateral masking rules implemented in France and described extensively in French audiometry guidelines.¹⁹

Hearing status of subjects

Following the conventional manual audiometry with the Reference audiometer, a Pure-Tone Average (PTA) was computed for each ear by averaging audiometric threshold measures at the following frequencies: 500, 1000, 2000, and 4000 Hz (in line with the French guidelines³⁵). The PTA of each ear was next linked to a specific hearing status as shown in Table 1. The distribution of thresholds at individual frequencies for all subjects is detailed in Supporting Information S1: Table 1.

Automated ML audiometry procedure

For the automated ML-based audiometry, a total of eight pure pulsed tones of the same level and frequency were presented for each trial. The use of multiple pulses was chosen to promote pulse detection in subjects with tinnitus that might interfere with pure-tone detection. The duration of each pulse was 250 ms, including a 20 ms sinusoidal ramp at the start and stop of each signal. The pulsed tones were terminated as soon as the subject responded by pressing the response button on the remote control. Silence intervals of 250 ms separated each pulsed tone (inter-pulse interval). The duration between two distinct test tones (interstimulus interval) was between 2 and 5.5 s with a jitter to avoid predictability effects.

These values differed from Schlittenlacher et al.²⁸ (the Yes/No ML method), in which the authors presented three tone pulses of 250 ms duration, with a 100 ms interval. Here, a higher number of tone pulses and longer inter-pulse intervals made the task easier, especially for elderly subjects (see also IEC 60645-1:2017³⁶ on interval lengths).

Initialization phase

Similar to the manual audiometry procedure, the automated pure-tone audiometry test systematically begins by testing the better ear declared by the subject, and if no better ear is declared, the right ear is always tested first.

In line with Schlittenlacher et al.,²⁸ more precisely, the Yes/No ML method, subjects are asked to press a response button only when they hear a tone. Subjects are provided with a 4 s time window to provide their answer. No answer within the 4 s window is counted as a negative response, that is, the subject did not perceive the test sound. Subjects are previously informed that tones are sent stochastically and that some passages of several seconds may not contain any sound. Next, similar to the initial test phase described in Schlittenlacher et al.,²⁸ an Initialization phase is performed to

approximate the hearing thresholds of the subject for the following audiometric frequencies: 1, 1.5, 2, 3, 4, 6, 8, 0.75, 0.5, 0.25, and 0.125 kHz.

The first test sound is presented at a frequency of 1 kHz and an intensity level of 60 dB HL. When the stimulus is heard, the intensity level is decreased in 20 dB steps until it is no longer audible. If it is not heard, the intensity level is increased by 20 dB until the sound is audible. This procedure is repeated until both a positive and a negative response are obtained. For intensity levels over 70 and 80 dB, the step size is reduced to 10 and 5 dB, respectively. The next test tone is 1.5 kHz and is presented at a level of -20 dB below the last intensity tested if the last tested sound was heard. If not, a level of +20 dB was used. Overall, for all 11 audiometric frequencies tested, both positive and negative responses were obtained. Hence, a minimum of 22 test stimuli is used.

For the audiometric frequencies at the extreme ends of the audiogram (i.e., 0.125 and 8 kHz), the "increase" step size was reduced to 10 dB instead of 20 dB. This allowed a finer estimate of those thresholds, which in turn lowered the number of test points during the subsequent Testing phase (see section below).

As described in Schlittenlacher et al.,²⁸ the positive and negative responses provide a first approximation of the audiometry for estimating the audiogram using a Gaussian process (GP) classifier.³⁷ This classifier provides a Gaussian distribution of response probabilities for all intensity (range: -20 to 90 dB, unit: 1 dB) and frequency combinations (range: 0.125–8 kHz, unit: 0.1 octave). The frequency/intensity combination for which the average probability according to the GP classifier is closest to 0.5 is considered the current audiogram estimate. The GP incorporates prior knowledge within its covariance function (kernel) – covariance between different audiometric points. For instance, frequency-wise, a squared exponential kernel with a length scale of 0.5 octaves captures the fact that thresholds at adjacent frequencies are correlated.^{28,38} On the intensity side, the GP linear kernel captures the fact that the probability of a sound being heard increases with increasing intensity. This parameterization of the GP function, along with the experimentally obtained responses, are used to generate a latent function, processed via a likelihood function. The GP linear kernel initially has a zero mean, which is equivalent to agreeing on a Bayesian prior threshold level of 0 dB HL before any measurement.

In line with Schlittenlacher et al.,²⁸ the likelihood function of the GP takes the form of a Gaussian cumulative density function, with a set standard deviation of 3 dB, considered to approximate the slope of the psychometric function. The likelihood function was scaled to cover probabilities between 0.01 and 0.99. Subjects are assumed to make "errors" in their responses (misses or false alarms); for example, due to accidentally pressing or not the answer button on average in 1 out of 100 trials. Since our stimuli are relatively long (eight pulsed tones, in comparison to three tones used in Schlittenlacher et al.,²⁸ the error rate is expected to be lower than what has been previously observed (miss rate measured in ²⁸ at 1.2% and false alarm rate measured at about 4.1%).

Testing phase

In line with Schlittenlacher et al.,²⁸ the Bayesian active learning mechanism uses the probabilities given by the GP function to select the next tone intensity and frequency to maximize the mutual information between the expected response and the GP estimate.³⁹

$I(y^*; \theta|x^*) = H(y^*|x^*, D) - E_{\theta \sim p(\theta|D)}(H[y^*|x^*, \theta])$ with the first term on the right being the expected response entropy, the second term is the expected conditional response entropy given the GP function estimate, H is the Shannon entropy,⁴⁰ D are the answers already obtained, x^* represents the frequency and signal intensity level for the next test, y^* represents the expected answer, and θ represents the GP function.

Choosing the intensity/frequency pair in such a way that minimizes uncertainty^{23,28} also allows the tested frequencies to vary widely from one another (i.e., back-to-back frequencies can be far apart), thus avoiding predictability issues – for instance, with non-cooperative subjects. We should note that the use of a long series of pure tones with eight pulsed tones limits any attentional issue that elderly subjects might experience.

Stopping criteria

Two stopping criterion values were implemented. The first criterion was fixed at a minimum of 50 trials and a maximum of 70 trials following the end of the Initialization phase. The second criterion was based on the uncertainty of the threshold prediction, which is estimated as ± 1 SD around the threshold estimated by the model. The criterion is reached when the biggest uncertainty is <6 dB. These stopping criterion values were optimized to ensure: (1) the highest possible reliability, i.e., stopping the tests when additional information becomes only weakly informative,^{26,28} and (2) compatibility with daily clinical practice, and to avoid the impact of high-level factors. In fact, this maximum number of trials was fixed to limit the duration of the tests for subjects with relatively high response variability. Moreover, data from Schlittenlacher et al.²⁸ showed that the root mean square difference (RMSD) fell below 4 dB as early as 30 trials following Initialization, and below 2 dB after 60 trials. Finally, no statistically significant difference was obtained beyond 70 trials independently of the testing method used (see Figure 6 of Schlittenlacher et al.²⁸). In the current study, although we did not measure the exact number of trials for each subject, the overall minimum number of trials was 72 (22 trials minimum for the Initialization phase and 50 trials minimum for the Testing phase).

General safety limits for automated audiometry procedure

As the main aim of this improved ML method is to test a wide range of NH and HI listeners in a fully automated manner, we first established some safety limits in terms of the maximum stimulus level to be presented to all subjects. Hence, the maximum stimulation level was set at 90 dB HL (the chosen equipment allowed this intensity level to be reached, even at 125 and 250 Hz) and the minimum stimulation level was set at -20 dB HL. If one of these intensity limits is reached during the Initialization phase, the protocol is to move to the next test frequency.

Next, to test subjects with severe hearing loss with maximum safety in terms of high-intensity sound exposure, the variation intensity step is reduced to 5 dB during the Initialization phase when the stimulus intensity level exceeds 80 dB HL, to guarantee that HI subjects are not presented with excessively high sound levels. This limits auditory overstimulation and uncomfortable levels that could occur due to loudness recruitment (i.e., abnormally rapid growth of the sensation of sound force in the presence of hearing loss). Similarly, during the Testing phase following Initialization, when the stimulation intensity to be presented is above 80 dB HL, a safety limit is imposed such that the stimulation intensity cannot be higher than 5 dB with respect to the last point tested within the concerned octave.

These limits are different from those used in Schlittenlacher et al.²⁸ (maximum level of 77 dB HL and minimum level of -10 dB HL), but it allows us to test listeners with severe hearing loss. While Song et al.²⁴ used -20 to 100 dB HL limits, the transducers used here do not allow for such a high stimulation level, especially at low frequencies, considering the maximum distortion rate set forth in IEC 60645-1:2017³⁶ for type 2 audiometers.

Addition of positive and negative transient responses during initialization phase when necessary

As our aim is to test a wide range of listeners as in daily clinical settings, including those with cochlear dead zones and severe hearing loss, previously unaddressed in published ML-based audiometry approaches, we implemented a second key change. More precisely, during the Initialization phase, if the subject provides no response for a subset of frequencies due to the presence of either severe hearing loss or an unresponsive cochlear region, then, for the subsequent Testing phase, the audiogram and uncertainty estimates are only assessed for that particular subset of frequencies that the subject can still hear. In fact, transient negative responses are added automatically below the measurement range (i.e., below -20 dB HL) when the test sound is perceived at -20 dB HL. Similarly, transient positive responses are added automatically above the maximum testable intensity (i.e., above 90 dB HL), to limit additional test stimuli to be presented beyond this level during the Testing phase. These transient responses constrain the threshold search interval zones during the Testing phase. In fact, if no response is obtained on more than two consecutive audiometric frequencies following the Initialization phase, these frequencies are not tested during the Testing phase.

Those transient positive and negative responses (when present) are not taken into account for the final audiogram calculation and those points are highlighted as transient responses in the audiogram display. No such implementation was available in previous ML-based audiometry approaches.

Automated contralateral masking during automated audiometry procedure

Importantly, for automated audiometry tests to be fully autonomous, we have also implemented automated contralateral masking in contrast to the ML-based audiometry approach described in

Schlittenlacher et al.²⁸ (however see Heisey et al.²⁹). Contralateral masking prevents cross-hearing and should be applied in cases of asymmetric hearing loss, or conductive, or unilateral hearing loss. Several assumptions were made here for the calculation of the masking noise levels, in line with audiometry standards and guidelines,^{19,41} as well as the literature.

The standard protocol implemented on the current improved ML audiometer is to always apply a roving contralateral masking using narrow-band noise when the test stimulus to be presented is over 35 dB HL.⁴² The maximum masking intensity is set at 80 dB HL. Masking noise began randomly in the 1–3 s interval before the onset of the pure-tone sequence. This ensures that inadvertent response button presses at the onset of the masking noise are not considered as the subject actually hearing the test stimuli. The masker remained present for a total of 5–8 s. The noise ramped on for 70 ms at the beginning of the intersequence interval and ramped off during the final 70 ms. In Heisey et al.,²⁹ an automated masking procedure was implemented and masking noise began randomly in the 250–1500 ms before the onset of the test sound. Pilot data (not shown here) suggested that this was too short, especially for elderly subjects, and led to a high number of false alarms.

Like for manual audiometry, for each test sound, an efficacy criterion (i.e., the minimum masking intensity necessary to effectively mask the contribution of the contralateral nontested ear¹⁹) and a no-overmasking criterion (i.e., the maximum masking intensity applicable to the contralateral or nontest ear above which the masking noise could be heard by the test ear and may impact the detection of the test stimuli¹⁹) are calculated to determine the masking intensity needed for the nontest ear.

The efficacy criterion (M_{eff} , in dB) is calculated for air-conduction audiometry as follows:

$$M_{\text{eff}} = \text{PL} - \text{IA} - \text{Masking Min} + \text{ABG of NTE.}$$

The no-overmasking criterion (M_{nov} , in dB) is calculated as follows:

$$M_{\text{nov}} = \text{PL} - \text{ABG of TE} + \text{IA} - \text{Masking Max.}$$

(Key: TE, test ear; NTE, nontest ear; ABG, air bone gap, PL, presentation level; IA, interaural attenuation).

The final Masking value is determined as follows:

1. If $M_{\text{eff}} > M_{\text{nov}}$, the masking value used is M_{eff} .
In this case, a warning message appears on the software interface at the end of the test stating that “the thresholds obtained may be adversely affected by contralateral masking. Ipsilateral Rainville masking⁴³ is recommended.”
2. If $M_{\text{eff}} = M_{\text{nov}}$, the masking value used is M_{eff} .
3. If $M_{\text{eff}} < M_{\text{nov}}$, the masking value used is the arithmetic mean of M_{eff} and M_{nov} .

The Masking Max is set at 0 dB and refers to the maximum signal-to-noise ratio at the level of the (inner) test ear that ensures detection of the test stimuli despite the presence of masking noise in the nontest ear. In fact, pure-tone detection is still possible at a signal-to-noise ratio of around -5 to -10 dB in the presence of a narrow-band masker.⁴⁴

The Masking Min is set at -20 dB and refers to the minimum signal-to-noise ratio observed at the level of the (inner) nontest ear to ensure full masking of the nontest ear.

In fact, the Masking Max currently at 0 dB could be set lower (at -5 or -10 dB) and the Masking Min currently at -20 dB could be set higher (at -10 or -15 dB). Together, this provides an overall -10 – 20 dB range allowing for adequate masking considering the presence of unknown parameters.

The value of the Interaural Attenuation (IA, in dB) used to calculate masking is determined as a function of the specific on-ear transducer (TDH 39 tested here; see Methods section "Material and Calibration") used to perform the audiometry test, and is also dependent on the test frequency. IA values available from the literature have been implemented,^{42,45,46} and when these IA values are unknown for a given on-ear transducer, the default value applied is 50 dB, which corresponds approximately to the IA value observed for the different on-ear transducer types, averaged over all frequencies.

The ABG of the nontest ear refers to one of the following:

1. When no previous audiometry results are available (as in the current study for the first ear tested per subject), a 40 dB ABG assumption is adopted.
2. When only the air-conduction threshold is available (as in the current study for the second ear tested per subject), ABG is defined as being equal to the estimated air-conduction threshold at the frequency tested divided by a coefficient (see example cases in Supporting Information S1: Note 2), depending on the suspected hearing loss etiology from the result of the Lateralization test. If there is no evidence to suggest a conductive hearing loss, the ABG of the NTE is estimated from the air conduction pure-tone threshold divided by a factor $[3 + ((\text{Threshold of NTE}/120) * 2)]$. When a conductive hearing loss is suspected, the division coefficient is $[1 + (\text{Threshold of NTE}/120)]$.

The ABG of the test ear is unknown in the absence of bone-conduction thresholds. Here we adopted the worst-case assumption for the calculation of masking, that is, cases whereby the risk of conductive hearing loss is maximal. Clinically, this maximum audiometric ABG for any given frequency is considered to be 60 dB HL.^{47–49} Hence, this maximum value was systematically used for the current study.

Supporting Information S1: Note 2 provides three example cases to illustrate the respect of both the efficacy and no-overmasking criteria. In addition, for all three example cases provided, we compare masking calculations as applied in the current study, as well as the masking calculation used by Heisey et al.,²⁹ whereby the masking value is always computed as 40 dB below stimulus presentation level. Overall, this latter method seems inadequate for sufficient masking.

Measurement of thresholds

Although the two manual audiograms (Reference and manual improved ML audiometer) were obtained using the same manual

audiometry procedure, they differed with regard to the following: (1) the headphones used despite the calibration of both types of headphones and (2) the audiometric booths used for testing, although both were double-walled, sound-proof audiometric booths.

On the other hand, the automated improved ML audiometer differed from the two manual audiograms as it assessed continuous threshold estimates in terms of frequency and provided confidence interval estimates. For comparison purposes, the thresholds of the automated improved ML audiometer were discretized to the conventional audiometric frequencies. Importantly, the manual measures differed from the automated measures with regard to the threshold definition. The two manual audiometry approaches measured the subject's threshold using the asymmetric up-down procedure. In contrast, the automated improved ML method defined the threshold as the predicted 50% audible contour, in line with Schlittenlacher et al.²⁸

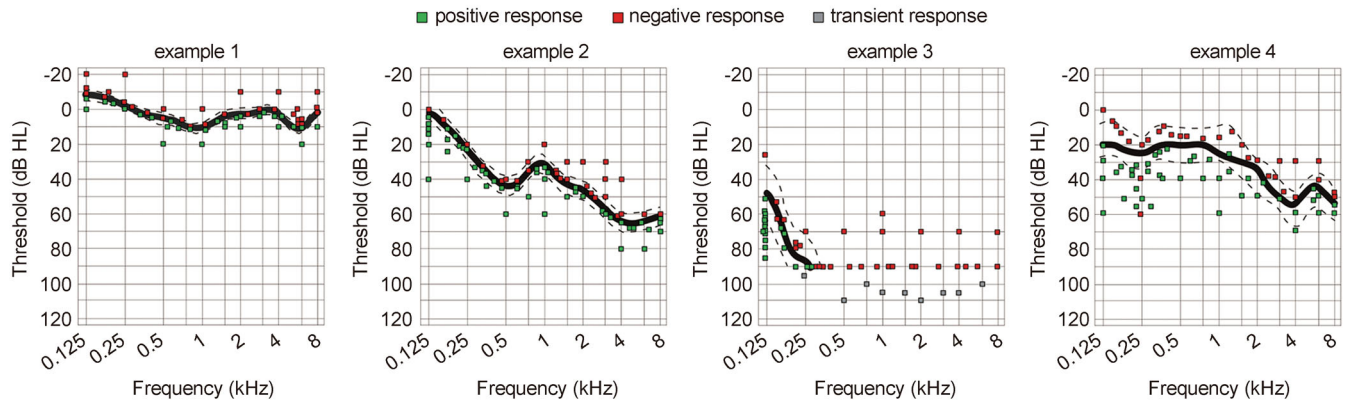
Statistical tests

All group-level statistical tests and effect size calculations were performed using JMP Pro 14.0 on a Mac platform. The Shapiro–Wilk test of normality was performed for all data sets. Non-normally distributed data was examined using nonparametric tests. Pairwise comparisons were carried out using the Steel–Dwass method for nonparametric comparisons. To compare more than two groups, one-way analysis of variance rank tests (Kruskal–Wallis H test) were used. To assess agreement between testing methods and test-retest measures, intra-class correlation (ICC) measures were obtained using a one-way random-effects model. For posthoc multiple comparisons analyses, α values were Holm–Bonferroni-corrected.

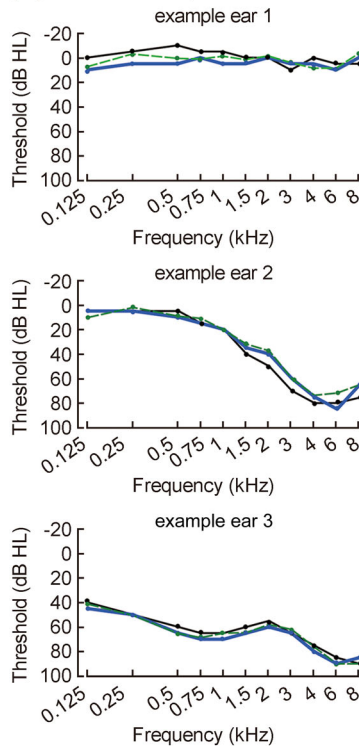
Results

Figure 1A shows examples of audiograms measured for four example ears using the automated ML method (the black line indicates the 50% contour and the dashed line indicates the uncertainty estimate). Figure 1B shows examples of audiograms measured for three distinct representative ears using the three methods (i.e., manual Reference audiogram: black line; manual improved ML method: blue line; automated improved ML method: green line): example ear 1 with NH thresholds (Figure 1B; top), example ear 2 with sloping high-frequency hearing loss (Figure 1B; middle), and example ear 3 with hearing loss at most frequencies (Figure 1B; bottom). All thresholds measured using the manual Reference method are shown in Figure 1C (thin lines indicate individual ears, thick line indicates mean, shaded area indicates SD). Similarly, thresholds using the manual and automated improved ML methods are shown in Figure 1D,E, respectively. For comparison, all mean ± 1 SD thresholds using the three methods are shown in Figure 1F.

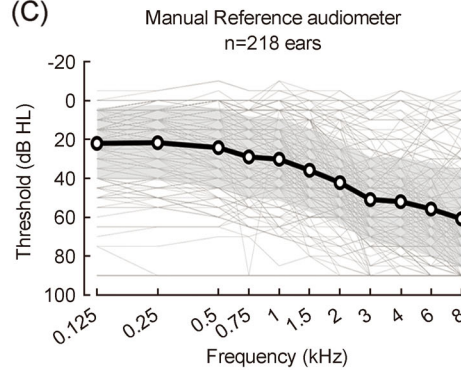
(A) Fully automated improved ML-based audiogram



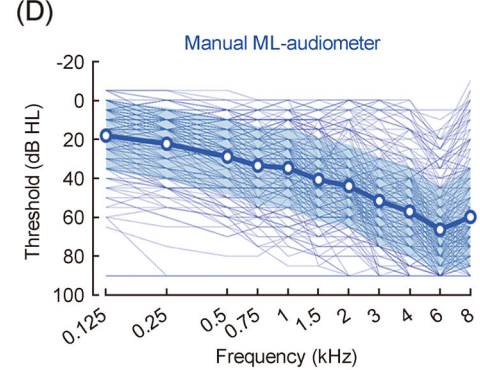
(B) Threshold comparisons



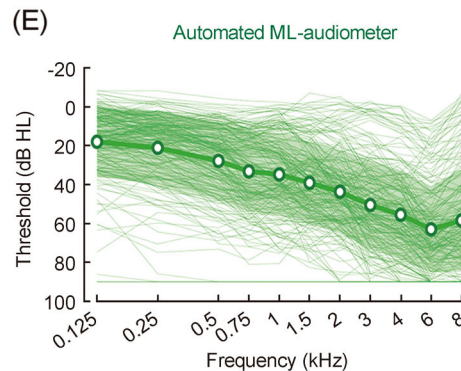
(C)



(D)



(E)



(F)

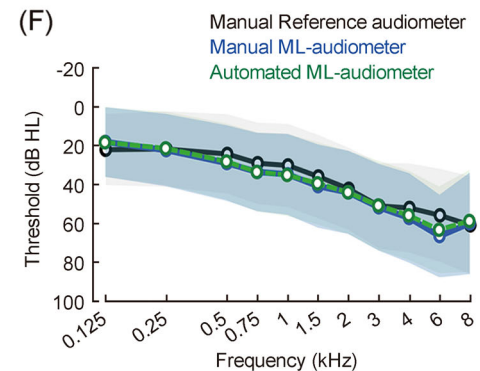


FIGURE 1 Comparison of air-conduction audiometry measures using three different methods. (A) Four examples of fully automated ML-based air-conduction pure-tone audiograms. The red and green squares indicate each subject's negative and positive responses, respectively. The thick black lines correspond to the continuous threshold estimate and the dashed thin lines surround the area within which 50% is within one standard deviation of the highest probability, according to the GP classifier. Transient responses (gray squares) are not taken into account for the 50% audible contour. (B) Air-conduction pure-tone audiograms for three representative subjects tested using the three methods as follows: (1) manual reference audiometer (black), (2) manual improved ML audiometer (blue), (3) automated improved ML audiometer (green). The continuous automated audiometry thresholds were discretized to the 11 frequencies manually tested for comparison. (C) Mean and SD of audiograms for the 218 ears tested using the manual Reference audiometer (thin lines correspond to individual ears, thick line corresponds to mean, shaded area indicates 1 SD). (D) Mean and SD of audiograms for the 218 ears tested using the manual improved ML audiometer. (E) Mean and SD of audiograms for the 218 ears tested using the automated improved ML audiometer. (F) Comparison of all means and SDRR with the three methods.

Comparison of threshold values

The mean raw signed differences for all 109 subjects and all 218 ears are shown in Table 2 (for individual frequencies, see Supporting Information S1: Table 2). These values compare favorably to two

audiometry measurements performed under identical conditions (see Margolis et al.²¹; SD of 5.5 dB using the same audiometer, same location, and same experimenter; see also Mahomed et al.,²⁰ Gosztanyi et al.,⁵⁰ Schmuziger et al.,⁵¹ and Ishaq et al.).⁵² The mean absolute differences are also shown in Table 2 (for individual

TABLE 2 Overall performance differences computed with the first procedure minus the second one listed ($n = 218$ ears).

A pair of comparisons for all frequencies tested	Raw differences (mean \pm SD, dB HL)	Absolute differences (mean \pm SD, dB HL)
Manual ML audiometer versus Reference	2.97 \pm 7.93	5.65 \pm 6.30
Automated ML audiometer versus Reference	1.89 \pm 7.87	5.64 \pm 5.81
Manual versus automated ML audiometer	-1.08 \pm 4.42	3.18 \pm 3.25

Abbreviation: ML, machine learning.

TABLE 3 Comparison of accuracy with previous ML-based audiometry (signed difference computed with the first procedure minus the second one listed).

	Automated ML-based audiometry versus Reference			Current study ($n = 218$ ears)
	Song et al. ²⁴ ($n = 42$ ears)	Barbour et al. ²⁶ ($n = 42$ ears)	Schlittenla- cher et al. ²⁸ ($n = 40$ ears)	
Signed difference	-0.011 \pm 5.61 dB	-0.969 \pm 6.02 dB		Manual ML audiometer versus Reference = 2.97 \pm 7.93 dB Automated ML audiometer versus Reference = 1.89 \pm 7.87 dB Manual versus automated ML audiometer = -1.08 \pm 4.42 dB
Absolute difference	4.16 \pm 3.76 dB	3.24 \pm 5.15 dB		Manual ML audiometer versus Reference = 4.19 \pm 5.21 Automated ML audiometer versus Reference = 4.53 \pm 5.01 Manual versus automated ML audiometer = 3.18 \pm 3.25
RMSD		5.58 dB	NH: 4.9 dB HI: 7.2 dB	Manual ML audiometer versus Reference = 6.71 dB \pm 14.93 Automated ML-audiometer versus Reference = 6.76 dB \pm 14.41 Manual versus automated ML-audiometer = 4.28 dB \pm 8.50

Abbreviations: HI, hearing impaired; ML, machine learning; NH, normal hearing; RMSD, root mean square difference.

frequencies, see Supporting Information S1: Table 3). The fully automated ML method differed from the Reference audiometry thresholds by ~ 6 dB on average, and the manual and automated ML methods differed by only ~ 3 dB.

Performance comparisons

Next, performance or accuracy comparison between the three different methods was carried out by computing the RMSDs in dB HL from raw threshold differences. The overall RMSD for all subjects tested here when comparing the automated ML-audiometer with the Reference audiometer = 6.76 dB (NH = 5.53; HI = 6.70; see Supporting Information S1: Table 4). When comparing the manual and automated ML audiometer, the overall RMSD = 4.28 dB (NH = 4.01; HI = 4.31). The two manual approaches (Reference and manual ML audiometer) had an overall RMSD = 6.71 dB (NH = 4.88; HI = 6.66). In Table 3, all differences and RMSD values are compared with previous ML-based audiometry approaches developed.^{24,26,28}

Statistical comparisons

First, we checked if the threshold distribution at the individual tested frequencies using the three different methods was normally distributed using the Shapiro-Wilk test of normality. As the data sets were

not sufficiently Gaussian (Shapiro-Wilk test of normality), we chose to perform nonparametric statistical tests.

Next, to assess the presence of statistically significant differences, we compared the raw thresholds from the three audiogram measures (manual Reference, and manual and automated ML audiometer). No significant difference was found between thresholds from the three methods (Kruskal-Wallis H test, $\chi^2(2) = 3.95$, $p = 0.139$; for posthoc comparison for individual frequencies, see Supporting Information S1: Table 5). Moreover, no significant difference was found when comparing: (1) the manual ML audiometer and Reference (Steel-Dwass, $p = 0.132$), (2) the automated ML audiometer and Reference (Steel-Dwass, $p = 0.361$), and (3) the manual and automated ML audiometer (Steel-Dwass, $p = 0.824$).

We also used a two-way random-effects (inter-method reliability) and expected mean squares to look at the reliability between the three methods (manual Reference, and manual and automated ML audiometer). The main effect of the methods confirmed good agreement between the three methods (F ratio = 2.97; $p = 0.0522$).

PTA comparisons

The air-conduction pure-tone audiogram is often summarized for each ear with a pure-tone average (PTA) of thresholds, measured for

TABLE 4 Hearing status from PTA distribution.

Classification of hearing status	Reference	Manual ML audiometer	Automated ML audiometer
NH hearing PTA \leq 20 dB HL	$n = 32$ ears	$n = 26$ ears	$n = 27$ ears
Hearing loss PTA $>$ 20 dB HL	$n = 186$ ears	$n = 192$ ears	$n = 191$ ears

Abbreviations: NH, normal hearing; PTA, pure-tone average.

a subset of frequencies (0.5, 1, 2, 4 kHz; see guidelines³⁵). The distribution of hearing status computed from the PTA is shown in Table 4 for the three methods. Detailed analysis of the ears not classified similarly by the 3 methods revealed that: (1) for the manual ML audiometer, PTA differences varied between 1.25 and 2.5 dB HL with respect to the Reference audiometer values, and (2) for the automated ML audiometer, PTA differences varied between 0.5 and 2.2 dB HL with respect to the Reference audiometer values.

Moreover, the intra-class correlation (ICC) was used to assess the agreement between the PTA values obtained using the three methods (manual Reference, and manual and automated ML audiometer). We found a high ICC value (i.e., close to 1), indicating high agreement between the three methods (ICC value obtained using a one-way random-effects model = 0.977).

Sensitivity and specificity analysis

Sensitivity and specificity values are often calculated for new testing procedures and provide interesting metrics for hearing screening purposes. In fact, sensitivity is calculated based on how many people have the disease in comparison to the Reference audiometer results.⁵³ It is also referred to as the true positive rate. With a fixed decision criteria to conclude hearing loss when PTA $>$ 20 dB, the sensitivity of both manual and automated ML audiometer = 100%. Importantly, these results demonstrate that the pathology is never missed by both measures of the ML audiometer (False negatives = 0). Specificity is calculated based on how many people do not have the disease in comparison to the Reference audiometer results. With the same criteria as above, the specificity of manual ML audiometer = 81.25% and automated ML audiometer = 84.38%. These results demonstrate that the pathology is attributed to certain subjects who are not classified as hearing loss by the gold standard, that is, the Reference audiometer (False positives = 6 or 5 ears over a total of 218 ears for manual and automated ML audiometer, respectively; ~2%–3%). Finally, the positive predictive value of the manual ML audiometer = 96.88%, and the automated ML audiometer = 97.38%. Although the above results are from the comparison of the ML audiometer with the Reference audiometer (i.e., comparison of different equipment), the Sensitivity and Specificity when comparing audiometry from exactly the same equipment (i.e., manual vs. automated ML audiometer) correspond to 99.48% and 100%, respectively.

Duration of tests

For all subjects tested, the automated ML audiometry took on average 13.73 ± 2.80 min (mean \pm SD; min–max = 9.10–19.90) to obtain continuous audiometry thresholds with 1-dB precision, as well as uncertainty estimates. The manual Reference audiometry and manual ML audiometry took an average of 7.67 ± 1.89 min to obtain discrete thresholds at 11 frequencies with 5-dB precision. Even though the manual audiometry tests are faster, the fully automated ML-based audiometry does not require the presence of the hearing specialist and can therefore be carried out before seeing the hearing specialist. In addition, the automated method provides continuous audiometry thresholds instead of 11 discrete thresholds for the two manual methods. The maximum test time measured here at ~20 min for the automated ML audiometry indicates compatibility with daily clinical practice while avoiding the impact of high-level factors, such as listener fatigue.

STUDY 2: PERFORMANCE OF A FAST ML-BASED AUDIOMETRY

For screening purposes, we developed a fast version of the automated improved ML audiometry (detailed in Study I) that can be used in both medical and nonmedical settings; for example, for hearing screening in schools or hearing follow-ups during the course of certain medical treatments. In the case of screening, our reasoning was that the number of trials during the Testing phase could be reduced for a less precise audiogram to reduce the testing time from the regular ML-based method described in Study I (average duration estimated: 13.73 ± 2.80 min).

Methods

Subjects

A subset of subjects from Study I also participated in Study II ($n = 43$, 22 women), which was run on the same day following all tests of Study I. The mean age of subjects tested was 55.3 ± 12.03 (min–max age: 18–88 years) and the hearing status of the tested ears was distributed as follows: 34 NH ears and 52 HI ears (HI when PTA $>$ 20 dB HL from manual Reference audiometry measures).

Manual audiometry procedure

See Methods section "Manual audiometry procedure" of Study I.

Fast automated ML audiometry procedure

The procedure is the same as the one described in Study I (see Methods section "Automated ML audiometry procedure" in Study I), except for the first stopping criterion used. More precisely, the same

Initialization phase was used as for study I. During the Testing phase, to decrease the test duration, the first stopping criterion was fixed at a minimum of 25 trials and a maximum of 45 trials. The second stopping criterion used during the Testing phase was similar to the one described in Study I.

Results

Comparison of threshold values for fast automated ML audiometry

The mean raw and absolute differences between the fast automated ML audiometry measures and the manual Reference measures were -4.17 ± 7.02 dB and 6.46 ± 4.99 dB HL, respectively (see Supporting Information S1: Table 6). While those differences are higher than the ones obtained for the regular automated ML audiometer from Study I (raw difference: 1.89 ± 7.87 and absolute difference: 5.64 ± 5.81), they remain within 15 dB of the manual Reference measures. Moreover, the raw audiometric thresholds obtained using the two methods (i.e., fast automated ML vs. Reference) did not differ significantly for all frequencies (Kruskal–Wallis H test, $\chi^2(2) = 3.57$, $p = 0.059$). The ICC was used to assess the agreement between the PTA values obtained using the two methods (manual Reference and fast automated ML). We found a high ICC value (i.e., close to 1) indicating high agreement between the two methods (ICC value obtained using a one-way random-effects model = 0.991).

Duration of tests

For all subjects tested here, the duration of the fast automated ML audiometry test was 9.85 ± 2.36 min (mean \pm SD; min–max = 4.33–14.90) to obtain continuous audiometry thresholds with 1-dB precision, as well as uncertainty estimates. Reducing the testing phase to a maximum of 45 trials instead of 70 trials (as in Study I) led to a gain of ~4 min. While the duration of the fast automated ML audiometry test is still higher than what has been reported for manual and conventional audiometry tests (5–10 min to obtain discrete thresholds,¹² as well as for other ML-based methods with a fixed masking procedure²⁹ (5–6.9 min), such a fast ML-based automated test can be run in the absence of a hearing specialist or health professional, and with a wide range of HI listeners.

STUDY 3: TEST-RETEST RELIABILITY OF ML-BASED AUDIOMETRY

Methods

Subjects

To examine the test–retest reliability of the regular, automated ML-based audiometry method described in Study I, a subset of subjects

from Study I also participated in Study III ($n = 50$ subjects, 22 women). Unlike Study II, which was run on the same day as Study I, Study III was performed on a different day (spaced by a maximum of 3 months from Study I). No notable otological history could be identified between Studies I and III for all subjects. The mean age of subjects tested was 60.1 ± 12.7 years (min–max age: 18–95 years) and the hearing status of the tested ears was distributed as follows: 16 NH ears and 84 HI ears (HI when PTA > 20 dB HL from Reference audiometer).

In parallel, to examine the cross-clinical agreement between manual Reference audiometry, we examined the audiograms of a separate group of 134 subjects (i.e., those subjects did not participate in Study I; 77 women). More precisely, two different hearing practitioners in two different clinics in France measured the hearing thresholds of the same group of subjects, using the same general testing protocol (i.e., the modified Hughson–Westlake procedure), but with different testing materials, although they have all been previously calibrated for audiometry tests. The mean age of subjects tested was 62.17 ± 17.04 (min–max: 18–95 years) and the hearing status of the tested ears was distributed as follows: six NH ears and 262 HI ears (HI when PTA > 20 dB HL from practitioner 1 manual Reference audiometry measures).

Assessing test–retest reliability of automated ML audiometry

The automated ML audiometry (described in Study I) was run twice for a subset of 50 subjects in the same conditions (i.e., same transducer, same test booth).

Assessing cross-clinical agreement with Reference audiometry

To examine the agreement between audiograms performed in clinical conditions using different Reference-type audiometers and transducers, we examined the audiograms of 134 subjects tested by two different hearing practitioners during routine clinical assessment and management protocols. The first hearing practitioner used a calibrated MADSEN Itera II audiometer with TDH 39 or ME-70 headphones. The second hearing practitioner used either a calibrated Interacoustics AD528 audiometer with DD45 headphones; or a calibrated Siemens Unity 3 with Sennheiser HDA 300 headphones. The order of tests performed by the two hearing practitioners was randomized. All manual audiometry procedures were carried out with appropriate masking (in line with Favier et al.¹⁹).

Results

Test–retest reliability of automated ML audiometry measures

For all 50 subjects tested, Figure 2A shows the mean of the test–retest thresholds measured with the automated ML audiometer

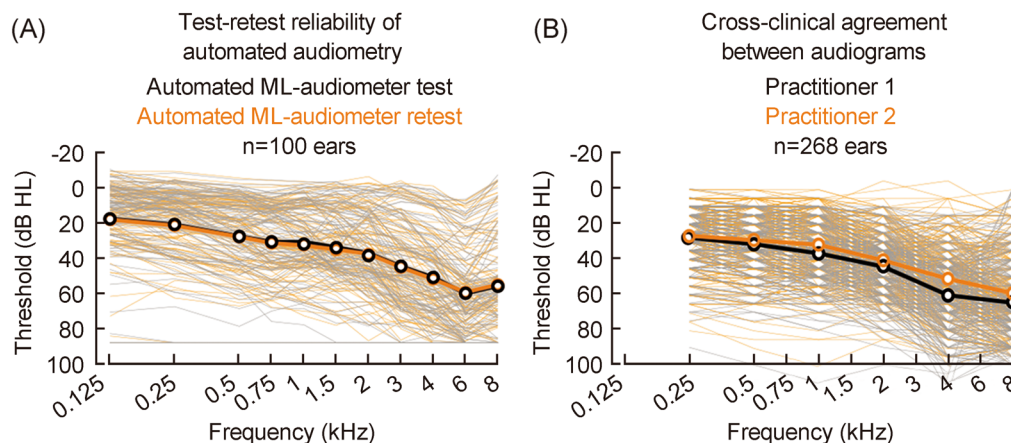


FIGURE 2 Test-retest comparisons of automated ML audiometry measures and cross-clinical agreement of Reference audiometry measures. (A) Mean of test-retest audiograms measured with the automated ML audiometer from Study I (thin lines correspond to individual ears, thick line corresponds to mean; test audiograms are shown in black, and retest audiograms are shown in orange). (B) Mean of the two Reference audiograms measured manually by two different practitioners (thin lines correspond to individual ears, thick line corresponds to mean; Practitioner 1 data are shown in black and Practitioner 2 data are shown in orange). ML, machine learning.

TABLE 5 Overall test-retest reliability of automated ML audiometry and cross-clinical agreement in manual Reference audiometry measures (different practitioners and different equipment).

For all frequencies tested	Raw differences (Mean \pm SD, dB HL)	Absolute differences (Mean \pm SD, dB HL)	RMSD (Mean \pm SD, dB HL)
Test-retest reliability of automated ML audiometry, $n = 100$ ears	-0.31 ± 4.38	2.92 ± 3.27	4.39 ± 7.99
Cross-clinical agreement with different materials and practitioners, $n = 268$ ears	-4.87 ± 10.33	8.13 ± 8.02	11.41 ± 21.34

Abbreviations: ML, machine learning; RMSD, root mean square differences.

described in Study I (individual lines show individual ears, test thresholds are in black, and retest thresholds are in orange). The mean signed difference between the two automated ML audiometry measures was 0.31 ± 4.38 dB and the mean absolute difference was 2.92 ± 3.27 dB HL (Table 5, for individual frequency comparisons, see Supporting Information S1: Table 7). Moreover, no significant difference was observed between the test-retest audiometry measures (Kruskal-Wallis H test, $\chi^2(1) = 0.025$, $p = 0.874$; all posthoc differences for individual frequencies > 0.05 , Supporting Information S1: Table 8). We also used a two-way mixed effects to look at the test-retest reliability for the automated ML-based audiometry measures. The main effect confirmed good agreement between the test-retest measures (F ratio = 0.003; $p = 0.954$). Moreover, the ICC was used to assess the agreement between the test-retest PTA values. We found a high ICC value (i.e., close to 1) indicating high agreement between the test-retest PTA measures (ICC value obtained using a one-way random-effects model = 0.995).

The overall RMSD for the test-retest automated ML audiometry measures was equal to 4.39 dB (Table 6; NH: 3.61; HI: 4.50; see also Supporting Information S1: Table 9). Table 6 shows the test-retest differences and RMSD values measured here in comparison with previous ML-based audiometry approaches.^{24,26,28}

Cross-clinical agreement of Reference audiometry measures

For all 134 subjects tested, Figure 2B shows the mean of the thresholds measured by two hearing practitioners (individual lines show individual ears, practitioner 1 thresholds are in black, practitioner 2 thresholds are in orange). The mean signed difference between the two Reference audiometry measures for all subjects was -4.87 ± 10.33 dB and the mean absolute difference between the two Reference audiometry measures was 8.13 ± 8.02 dB HL (Table 5, for individual frequency comparisons, see Supporting Information S1: Table 10). Furthermore, the two measures differed significantly, suggesting weak cross-clinical agreement (Kruskal-Wallis H test, $\chi^2(1) = 17.65$, $p < 0.0001$; significant posthoc differences at 500, 1000, 4000, and 8000 Hz; see Supporting Information S1: Table 11). The overall RMSD for the two manual Reference audiometry measures was equal to 11.41 dB (Table 6; NH: 8.98; HI: 11.46; see also Supporting Information S1: Table 12). Nevertheless, the ICC used to assess the cross-clinical agreement revealed a relatively high ICC value of 0.916 (obtained using a one-way random-effects model) and suggests relatively high agreement between the cross-clinical PTA measures.

TABLE 6 Comparison of test–retest reliability of automated ML audiometry method described here with previous ML-based audiometry, as well as the cross-clinical agreement of manual Reference audiometry.

	Song et al. ²⁴ (n = 42 ears)	Barbour et al. ²⁶ (n = 42 ears)	Schlittenlacher et al. ²⁸ (n = 40 ears)	Current study (n for automated ML audiometer = 100 ears; n for Reference audiometer = 268 ears)
Signed difference	0.75 ± 6.29 dB	−0.486 ± 7.15 dB		Test–retest of automated ML audiometry = −0.31 ± 4.38 dB Cross-clinical agreement with manual Reference audiometry = −4.87 ± 10.33 dB
Absolute difference	4.51 ± 4.45 dB	2.85 ± 6.57 dB		Test–retest of automated ML audiometry = 2.92 ± 3.27 dB Cross-clinical agreement with manual Reference audiometry = 8.13 ± 8.02 dB
RMSD		6.32 dB	6.9 dB	Test–retest of automated ML audiometry = 4.39 ± 7.99 dB Cross-clinical agreement with manual Reference audiometry = 11.41 ± 21.34 dB

Abbreviations: ML, machine learning; RMSD, root mean square difference.

DISCUSSION

In the last decade, fully automated ML-based audiometry approaches have been increasingly proposed in the literature^{12,24–26,28,54} in an effort to streamline the clinical examination process. Here, we extend the approach developed and tested in Schlittenlacher et al.²⁸: a non-parametric approach to estimating the audiogram in a frequency-continuous manner using Bayesian estimation and ML classification. Moreover, for fully automated tests to be used in clinical settings without the need for a hearing practitioner, we applied three major changes: (1) general safety limits during all automated audiometry testing in an effort to extend the method developed by Schlittenlacher et al.²⁸ for safe testing of a wide range of NH and HI subjects, including those with severe hearing loss; (2) addition of transient positive and negative responses to constrain the audiometry Testing phase, to cater for subjects with cochlear dead zones (or unresponsive regions) or when the threshold is beyond the intensity limits of the transducers, and (3) automated contralateral masking rules for improved threshold measures extending beyond the method described in Heisey et al.²⁹ In the latter, the authors apply an automated masking method to their ML-based audiometry through the addition of masking noise presented at a default masking value of 40 dB below the test stimulus level. However, as detailed in Supporting Information S1: Note 2, this approach remains limited and is, in some cases, insufficient to mask responses of the nontest ear (see Example Cases #1–3). In contrast, the masking rules described here aim to cater to a wide range of HI listeners. The main limit is that, in some cases, the efficacy criterion, that is, the intensity level at which contralateral masking becomes effective, is higher than the overmasking criterion, that is, the intensity level at which masking is too loud to ensure correct detection of the test stimulus. This configuration only occurs when using transducers with low transcranial transfer for listeners with substantial ABGs. However, this issue can be easily addressed by using inserts for air-conduction stimulation that limits the transcranial transfer (see Supporting Information S1: Note 2, Example Case #3).

In Study I, we showed that automated measures of the improved ML audiometer provide accurate hearing thresholds for air-

conduction pure-tone audiometry, not statistically different from those obtained from a well-established and conventional audiometer (Table 2). Importantly, in Study I, we tested a large number of subjects with a wide range of hearing status, including those with severe hearing loss and asymmetric losses. In fact, the mean absolute difference between the automated ML-based method and the conventional Reference one was ~6 dB. Such raw and mean absolute differences are in line with previously published ML approaches,^{24,28} as described in Table 3, and with other automated audiometry approaches in general.^{12,20} Finally, the high ICC also suggests very high agreement in PTA values obtained using the three methods. Future studies should aim to evaluate such automated ML-based methods on hard-to-test populations, including testing a larger number of subjects with severe hearing loss.

In Study II, we showed that a fast automated ML-based method with fewer test trials may be used for screening purposes as the difference between the fast automated ML-based method and the conventional Reference one was <15 dB (mean absolute difference ~7 dB). While the duration of ML-based audiometry tests is not faster for subjects as compared to manual conventional approaches, it is important to note that test automation allows a key time save for hearing professionals. Overall, such medical time save granted by test automation should, in principle, allow hearing practitioners to address a larger number of patients and decrease appointment waiting delays. Thus, an audiologist may set up a first patient in a test booth, including, provide all test instructions orally, answer any question from the patient, and equip the patient with the transducers before launching the automated test. During the entire testing time, the audiologist does not need to be in the test booth of the first patient and may use this time to set up a second patient in a different test booth. All individual responses of the subject or patient are displayed on the experimenter interface along with the 50% audiometric contour and the uncertainty range, which allows the audiologist to ensure that the subject responded consistently.

Finally, in Study III, we showed that the test–retest reliability of the automated improved ML approach was high, with no statistical difference between the two measures and high ICC (Table 5). In fact,

the mean absolute difference was ~3 dB. In comparison, the general agreement between two audiograms measured manually in different clinics using different conventional audiometers was particularly low (~8 dB). It should be noted that this value does not reflect a true “test-retest reliability” measure for the Reference audiogram as different hearing practitioners measured the hearing thresholds of the same group of subjects, using the same general testing protocol (i.e., the modified Hughson–Westlake procedure), but with different materials that are nevertheless all calibrated to provide similar results. This rather poor cross-clinical agreement may be due to various factors, including human variability (e.g., transducers used, calibration performed, expertise of clinician, positioning of material, and general testing method). Together, the results from all three studies confirm the performance and reliability of an automated improved ML audiometer within a heterogeneous hearing-loss population.

A remaining challenge with ML-based audiometry technologies is to convince clinicians to use them to directly monitor their time gain, and subsequently, assess the added benefits of higher audiometry precision (1-dB precision, uncertainty estimates, and continuous thresholds along the frequency axis) and higher audiometry repeatability as automated procedure implies fewer human errors and variability in testing procedure. Estimating thresholds to 1-dB precision along the full frequency axis may eventually lead to improved and targeted patient follow-up and improved hearing-aid adjustment parameters. Ultimately, such automated ML-based audiometry may facilitate close monitoring of hearing function in different settings, including hospitals, workplaces, and assisted living facilities, and at all stages of life, particularly during vulnerable periods (e.g., during development, during the course of ototoxic treatments). Long-term monitoring of such ML audiometry approaches is key to prevent and reduce as much as possible systematic errors by different users.

In Europe, automated techniques such as the AMTAS²¹ are available to clinicians for use with adult patients but remain scarcely used. One of the reasons may be due to the limited benefits of such techniques, for example, they may not provide continuous estimates of thresholds in frequency and predictable stimuli may not work for uncooperative subjects.⁵⁵ While ML-based methods address this disadvantage, such techniques do not allow the administration of other audiometry tests, for example, speech audiometry tests with automated annotation of patients’ responses. Thus, the time save provided by such automated techniques may appear limited. In the future, we aim to expand the current framework by including additional tests, such as bone-conduction audiometry⁵⁶ and speech audiometry.^{9,10} In addition, most automated methods, including the one described here, are unsuited for use with children.⁵⁷ Hence, ML-based audiometry needs to be expanded to include child-friendly approaches. Furthermore, to optimize the overall clinical procedures, future versions of ML-based methods should take into account listener parameters, such as age and sex, medical history, and otological examination results.⁵⁸ Finally, ML may offer additional advantages, such as automated evaluation of tympanic membrane images before audiometry testing. Such a combination of methods and tests within one medical device may offer a more unified and reliable diagnosis

for improved and personalized patient care, in addition to providing large data sets for future hearing research.

AUTHOR CONTRIBUTIONS

Nicolas Wallaert, Benoit Godey, Gwenaelle Creff, and Nihaad Paraouty designed research. Nicolas Wallaert secured funding. Nicolas Wallaert, Antoine Perry, and Hadrien Jean developed software. Nicolas Wallaert, Sandra Quarino, Benoit Godey, and Gwenaelle Creff collected behavioral data. Nihaad Paraouty analyzed data. Nicolas Wallaert and Nihaad Paraouty wrote and revised the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare the following competing interests: Nicolas Wallaert has a patent pending on technology described in the manuscript. Nicolas Wallaert has equity ownership in My Medical Assistant SAS. Antoine Perry, Hadrien Jean, and Nihaad Paraouty receive salaries from My Medical Assistant SAS. Sandra Quarino, Benoit Godey, and Gwenaelle Creff declare no competing interests.

ACKNOWLEDGMENTS

The work was supported by the French National i-Nov Grant (Nicolas Wallaert, DOS0127610/00) and Region Frand Est Deeptech-BPI France Grant (Nicolas Wallaert).

DATA AVAILABILITY STATEMENT

Data can be shared upon email request to the last author.

ETHICS STATEMENT

The study was approved by the French Regional Ethics Committee and the Comité de Protection des Personnes Est III (SI number: 22.03364.000107). All subjects were fully informed about the goal of the study and provided written consent before their participation.

ORCID

Nihaad Paraouty  <http://orcid.org/0000-0002-7530-5891>

REFERENCES

- Whitton JP, Polley DB. Evaluating the perceptual and pathophysiological consequences of auditory deprivation in early postnatal life: a comparison of basic and clinical studies. *J Assoc Res Otolaryngol*. 2011; 12:535-547.
- Marques T, Marques FD, Miguéis A. Age-related hearing loss, depression and auditory amplification: a randomized clinical trial. *Eur Arch Otrhinolaryngol*. 2022;279:1317-1321.
- Huang AR, Reed NS, Deal JA, et al. Depression and Health-Related quality of life among older adults with hearing loss in the ACHIEVE study. *J Appl Gerontol*. 2023;43:550-561.
- Lesica NA, Mehta N, Manjaly JG, Deng L, Wilson BS, Zeng FG. Harnessing the power of artificial intelligence to transform hearing healthcare and research. *Nat Mach Intell*. 2021;3:840-849.
- Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. *Nat Med*. 2018;24:539-540.
- Kim KM, Heo TY, Kim A, et al. Development of a fundus image-based deep learning diagnostic tool for various retinal diseases. *J Pers Med*. 2021;11:321.

7. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. *Deep Learn Med Image Anal.* 2020;3:21.
8. Shen Y, Shamout FE, Oliver JR, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat Commun.* 2021;12:5645.
9. Ooster J, Huber R, Kollmeier B, Meyer BT. Evaluation of an automated speech-controlled listening test with spontaneous and read responses. *Speech Commun.* 2018;98:85-94.
10. Ooster J, Tuschen L, Meyer BT. Self-conducted speech audiometry using automatic speech recognition: simulation results for listeners with hearing loss. *Comp Speech Language.* 2023;78:101447.
11. Crowson MG. Artificial intelligence to support hearing loss diagnostics. *Hear J.* 2020;73:8-9.
12. Wasmann JW, Pragt L, Eikelboom R, Swanepoel DW. Digital approaches to automated and machine learning assessments of hearing: scoping review. *J Med Internet Res.* 2022;24:e32581.
13. Carl AC, Hohman MH, Cornejo J. *Audiology Pure Tone Evaluation.* StatPearls Publishing; 2023.
14. Hughson WA, Westlake H. Manual for program outline for rehabilitation of aural casualties both military and civilian. *Trans Am Acad Ophthalmol Otolaryngol.* 1944;48(suppl):1-15.
15. ISO. *EN ISO 8253-1. Acoustics-Audiometric Test Methods-Part 1: Pure-tone Air and Bone Conduction Audiometry,* 2010.
16. American National Standards Institute. *Methods for Manual Pure-Tone Threshold Audiometry.* Vol S3, 2004:21.
17. American Speech-Language-Hearing Association. *Guidelines for Manual Pure-Tone Threshold Audiometry,* 2005.
18. British Society of Audiology (BSA). *Pure-Tone Air-conduction and Bone-conduction Threshold Audiometry With and Without Masking,* 2018.
19. Favier V, Vincent C, Bizaguet É, et al. French Society of ENT (SFORL) guidelines (short version): audiometry in adults and children. *Eur Ann Otorhinolaryngol Head Neck Dis.* 2018;135:341-347.
20. Mahomed F, Swanepoel DW, Eikelboom RH, Soer M. Validity of automated threshold audiometry: a systematic review and meta-analysis. *Ear Hear.* 2013;34:745-752.
21. Margolis RH, Glasberg BR, Creeke S, Moore BCJ. AMTAS[®]: automated method for testing auditory sensitivity: validation studies. *Int J Audiol.* 2010;49:185-194.
22. Eikelboom RH, Swanepoel DW, Motakef S, Upson GS. Clinical validation of the AMTAS automated audiometer. *Int J Audiol.* 2013;52:342-349.
23. Gardner JR, Song X, Weinberger KQ, Barbour DL, Cunningham JP. Psychophysical detection testing with Bayesian active learning. *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, of UAI'15.* AUAI Press; 2015: 286-297.
24. Song XD, Wallace BM, Gardner JR, Ledbetter NM, Weinberger KQ, Barbour DL. Fast, continuous audiogram estimation using machine learning. *Ear Hear.* 2015;36:e326-e335.
25. Song XD, Garnett R, Barbour DL. Psychometric function estimation by probabilistic classification. *J Acoust Soc Am.* 2017;141: 2513-2525.
26. Barbour DL, Howard RT, Song XD, et al. Online machine learning audiometry. *Ear Hear.* 2019a;40:918.
27. Barbour DL, DiLorenzo JC, Sukesan KA, et al. Conjoint psychometric field estimation for bilateral audiometry. *Behav Res Methods.* 2019b;51:1271-1285.
28. Schlittenlacher J, Turner RE, Moore BCJ. Audiogram estimation using Bayesian active learning. *J Acoust Soc Am.* 2018;144:421-430.
29. Heisey KL, Walker AM, Xie K, Abrams JM, Barbour DL. Dynamically masked audiograms with machine learning audiometry. *Ear Hear.* 2020;41:1692-1702.
30. ISO. *EN ISO 389-8. Acoustics-Reference Zero for the Calibration of Audiometric Equipment-Part 8: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Circumaural Earphones,* 2004.
31. International Electrotechnical Commission. *IEC 60318-1. Electroacoustics-Electroacoustics-Simulators of Human Head and Ear-Part 1: Ear Simulator for the Measurement of Supra-aural and Circumaural Earphones,* 2009.
32. ISO. *EN ISO 389-1. Acoustics-Reference Zero for the Calibration of Audiometric Equipment-Part 1: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Supra-aural Earphones,* 2017.
33. Société Française d'Audiologie. *Guide des bonnes pratiques en audiométrie de l'adulte.* Société Française d'Audiologie; 2006.
34. Kaernbach C. Simple adaptive testing with the weighted up-down method. *Percept Psychophys.* 1991;49:227-229.
35. Bureau International d'AudioPhonologie (BIAP). *02/1 Bis: Audiometric Classification of Hearing Impairments,* 1996.
36. International Electrotechnical Commission. *EN IEC 60645-1. Electroacoustics-Audiometric Equipment-Part 1: Equipment for Pure-tone and Speech Audiometry,* 2017.
37. Rasmussen CE, Williams CK. *Gaussian Processes for Machine Learning.* Vol 2. MIT Press; 2006:3-4.
38. Bisgaard N, Vlamming MSMG, Dahlquist M. Standard audiograms for the IEC 60118-15 measurement procedure. *Trends Amplif.* 2010; 14:113-120.
39. Houlsby N, Huszár F, Ghahramani Z, Lengyel M. 2011. Bayesian active learning for classification and preference learning. <https://arxiv.org/abs/1112.5745>
40. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J.* 1948;27:379-423.
41. ISO. *EN ISO 389-4. Acoustics-Reference Zero for the Calibration of Audiometric Equipment-Part 4: Reference Levels for Narrow-band Masking Noise,* 1994.
42. Munro KJ, Agnew N. A comparison of inter-aural attenuation with the etymotic ER-3A insert earphone and the telephonics TDH-39 supra-aural earphone. *Br J Audiol.* 1999;33:259-262.
43. Rainville MJ. L'Epreuve D'Assourdissement Ipsilateral Par Conduction Osseuse. *Int Audiol.* 1962;1:171-173.
44. Moore BCJ, Alcántara JL, Dau T. Masking patterns for sinusoidal and narrow-band noise maskers. *J Acoust Soc Am.* 1998;104:1023-1038.
45. Denes P, Naunton RF. Masking in pure-tone audiometry. *Proc R Soc Med.* 1952;45:790-794.
46. Hall III, JW, Mueller III, HG. *Audiologist's Desk Reference.* Vol 1. Singular Publishing Group; 1997:914.
47. Reger SN, Glasser O, ed. *Audiometers and Hearing Aids.* Vol 1. Medical Physics; 1944:9.
48. Zernotti ME, Arauz SL, Di Gregorio MF, Arauz SA, Tabernero P, Romero MC. Vibrant soundbridge in congenital osseous atresia: multicenter study of 12 patients with osseous atresia. *Acta Otolaryngol.* 2013;133:569-573.
49. Wolf M, Agterberg M, Snik A, Mylanus E, Hol M, Hempel J. Vibrant soundbridge and Bonebridge: bilateral application in a child with bilateral congenital ear canal atresia. *Br J Med Med Res.* 2015;5: 705-710.
50. Gosztonyi Jr., RE, Vassallo LA, Sataloff J. Audiometric reliability in industry. *Arch Environ Health Int J.* 1971;22:113-118.
51. Schmuziger N, Probst R, Smurzynski J. Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and etymotic research ER-2 earphones. *Ear Hear.* 2004;25: 127-132.
52. Ishak WS, Zhao F, Stephens D, Culling J, Bai Z, Meyer-Bisch C. Test-retest reliability and validity of Audioscan and Békésy compared with pure tone audiometry. *Audiol Med.* 2011;9:40-46.

53. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*. 2008;56:45.
54. Cox M, de Vries B. 2015. A Bayesian binary classification approach to pure tone audiometry. <https://doi.org/10.48550/arXiv.1511.08670>
55. Margolis RH, Morgan DE. Automated pure-tone audiometry: an analysis of capacity, need, and benefit. *Am J Audiol*. 2008;17:109-113.
56. Margolis RH, Moore BCJ. AMTAS[®]: automated method for testing auditory sensitivity: III. Sensorineural hearing loss and air-bone gaps. *Int J Audiol*. 2011;50:440-447.
57. Govender SM, Mars M. Validity of automated threshold audiometry in school aged children. *Int J Pediatr Otorhinolaryngol*. 2018;105:97-102.
58. Cox M, De Vries B. Bayesian pure-tone audiometry through active learning under informed priors. *Front Digit Health*. 2021;3:723348.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wallaert N, Perry A, Quarino S, et al. Performance and reliability evaluation of an improved machine learning-based pure-tone audiometry with automated masking. *World J Otorhinolaryngol Head Neck Surg*. 2024;1-16. doi:10.1002/wjo2.208