

Utilizing True Wireless Stereo Earbuds in Automated Pure-Tone Audiometry

Trends in Hearing
Volume 25: 1–13
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165211057367
journals.sagepub.com/home/tia



Zhenyu Guo^{1,2} , Guangzheng Yu², Huali Zhou²,
Xianren Wang³ , Yigang Lu², and Qinglin Meng²

Abstract

True wireless stereo (TWS) earbuds have become popular and widespread in recent years, and numerous automated pure-tone audiometer applications have been developed for portable devices. However, most of these applications require specifically designed earphones to which the public may not have access. Therefore, the present study investigates the accuracy of automated pure-tone audiometry based on TWS earbuds (Honor FlyPods). The procedure for developing an automated pure-tone audiometer is reported. Calibration of the TWS earbuds was accomplished by electroacoustic measurements and establishing corrected reference equivalent threshold sound pressure levels. The developed audiometer was then compared with a clinical audiometer using 20 hearing-impaired participants. The average signed and absolute deviations between hearing thresholds measured using the two audiometers were 3.1 dB and 6.7 dB, respectively. The overall accuracy rate in determining the presence/absence of hearing loss was 81%. The results show that the proposed procedure for an automated air-conduction audiometer based on TWS earbuds is feasible, and the system gives accurate hearing level estimation using the reported calibration framework.

Keywords

hearing loss, automated audiometry, calibration

Received 24 April 2021; Revised received 13 October 2021; accepted 14 October 2021

Introduction

The prevalence of hearing loss has increased in the past decade, and this is now a serious burden on the global healthcare system (Wilson et al., 2017). A serious problem is that people with mild-to-moderate hearing loss often ignore it, although hearing loss is associated not only with decreased speech communication but also with cognitive decline (Manheim et al., 2018; Wayne & Johnsrude, 2015). To examine hearing function and provide suitable remedies, several hearing tests are widely used in clinical practice (Kanji et al., 2018), for example, pure-tone audiometry (PTA), otoacoustic emissions, tympanometry, and auditory brainstem responses. Of these, PTA is regarded as the gold standard for hearing screening for people older than four years (Bright & Pallawela, 2016).

PTA is usually performed by an audiologist using a manual audiometer connected to calibrated air- or bone-conduction transducers in a soundproof booth. Although air-conduction PTA usually takes only about 10 min, both the equipment and audiologists for PTA screening may be in

short supply (Margolis & Morgan, 2008). In some underdeveloped and developing areas in particular, the extreme scarcity of hearing healthcare services means that the gap between supply and demand is large (Swanepoel et al., 2010). Unfortunately, the COVID-19 pandemic may exacerbate this gap due to the infection risk posed by physical contact (Centers for Disease Prevention & Control, 2020).

Automated PTA was developed to reduce audiologists' workload (Margolis & Morgan, 2008), and both its validity

¹School of Architecture, South China University of Technology, Guangzhou, China

²School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China

³Department of Otorhinolaryngology, the First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Corresponding author:

Qinglin Meng, School of Physics and Optoelectronics, South China University of Technology, No. 381, Wushan Rd, Guangzhou 510641, China.
Email: mengqinglin@scut.edu.cn



and effectiveness have been evaluated comprehensively (Mahomed et al., 2013). With the rapid development of consumer electronics, many applications have been developed to provide self-administered hearing screening via mobile phones (Paglialonga et al., 2015), tablets (Thompson et al., 2015), and personal computers (Choi et al., 2007). The automated PTA applications are implemented either locally on devices or remotely through the Internet (Honeth et al., 2010). With low-cost and accessible features, these applications pave the way for measuring hearing levels (HLs) of people with potential hearing loss (Louw et al., 2017). Some of the applications had impressive reliability (Margolis et al., 2018) and consistency with manual audiometry (Margolis et al., 2016; Thompson et al., 2015), especially when researchers have imposed strict controls over the conditions (e.g., earphones, calibration, and test environment). Besides, with the increasing interest in self-fitting hearing aids (Keidser & Convery, 2016) in recent years, the requirement for accurate audiograms measured without supervision has increased.

Most automated pure-tone audiometers require specifically designed hardware and standard transducers, which the general public cannot afford, thereby limiting their use. The practicality of these automated PTA applications is restricted if only standard earphones are used, although audiometer applications with some commercially available earphones can yield results consistent with those from clinical PTA (Chu et al., 2019; Foulad et al., 2013; Swanepoel et al., 2014).

In recent years, true wireless stereo (TWS) earbuds, especially those fitting in the cavum conchae (e.g., Apple AirPods, Huawei FreeBuds, Honor FlyPods), have become extremely popular globally. It would make sense to use TWS earbuds for automated hearing screening, given their popularity and portability. Moreover, since the digital audio signal is transmitted through Bluetooth (Bluetooth SIG, 2015) and since the analog audio output processing circuits are packaged in the bodies of the earphones (Floros et al., 2002), wireless earphones can give consistent output sound levels across platforms (e.g., different mobile phones, tablets, or computers). In contrast, different platforms equipped with different audio interfaces (e.g., sound cards) may achieve inconsistent output sound levels with wired earphones (Foulad et al., 2013). The consistent output across platforms means that once a pair of wireless earphones is calibrated, consistent PTA test accuracy across multiple platforms can be obtained, if used with a well-designed PTA application. In summary, the popularity and calibration advantages of TWS earbuds make them a good option for automated PTA applications, which may help numerous people via self-administered hearing screening at home and prescreening in clinics.

The purpose of the present study was to develop an automated PTA procedure using a pair of calibrated TWS earbuds (first-generation Honor FlyPods) and to assess whether it

gives HLs consistent with those for a standard manual air-conduction audiometer and earphones.

Automated PTA Procedure and Calibration

Various factors influence the accuracy of PTA, including the measurement procedure, device calibration (Schmidt et al., 2014; Van Tasell & Folkeard, 2013), cooperativeness of participants, environmental noise (Na et al., 2014), and earphone position (Paquier et al., 2016). In the following, we focus mainly on the first three issues. To avoid disturbance by environmental noise, all experiments with participants were conducted in a soundproof booth in the First Affiliated Hospital of Sun Yat-sen University (FAH-SYU; background noise was ~21 dBA). The earphone position was checked before testing to ensure that the earbuds were tightly in the cavum conchae of the participant. All human tests were approved in advance by the research ethics review committee of FAH-SYU.

Calibration of TWS Earbuds

A critical issue in obtaining accurate HLs lies in calibrating the electroacoustic system, especially for nonstandard earphones (Corry et al., 2017). A standard calibration procedure is (i) to use a standard ear simulator or acoustic coupler combined with a sound level meter to calibrate the output sound level of a specific earphone in units of dB SPL and then (ii) to convert the sound level (in units of dB SPL) to hearing level (in units of dB HL) using:

$$\text{HearingLevel} = \text{SoundLevel} - \text{RETSPL}, \quad (1)$$

where RETSPL (reference equivalent threshold sound pressure level) is the average SPL corresponding to the audiometric zero level (i.e., 0 dB HL) of a group of otologically normal subjects. RETSPL values can be obtained either from existing standards (ANSI/ASA, S3.6-2018, 2018; ISO 389-1: 2007, 2007; ISO 389-8: 2004, 2004) or by measuring the equivalent threshold sound pressure level (ETSPL, i.e., the hearing threshold in units of dB SPL measured by an electroacoustic system calibrated with specific calibrators) of a sufficient number of otologically normal subjects (Larson et al., 1988).

The choice of calibrator and earphone can affect the values of RETSPL (Wilber et al., 1988). Moreover, the enclosure of the earbuds used in the present study differs from that of any standard earphones used for audiological applications, so the existing RETSPLs of those standard earphones cannot be adopted directly. Besides, there is no standard ear simulator for earbuds. As an alternative approach, Ho et al. (2017) and Kam et al. (2012) used a KEMAR manikin (Burkhard & Sachs, 1975) to calibrate the output SPL.

After the electroacoustic calibration, the average hearing threshold of a group of otologically normal people needs to be measured. However, it has been shown that audiometric zero levels differ across regions (Borchgrevink et al., 2005; Wang et al., 2018). The actual average HLs of otologically normal young people were often below 0 dB HL when measured using standard earphones and audiometers in these large-sample surveys. Thus, the values of the RETSPLs could vary with demographics. To obtain HLs consistent with those achieved with a clinical audiometer using standard earphones, one workable strategy is to compensate the RETSPL of earbuds with average HLs measured using standard earphones, as follows (Ho et al., 2017; Wilber et al., 1988):

$$\text{RETSPL}_c = \text{RETSPL} - \text{HearingLevel}_{ave}, \quad (2)$$

where RETSPL_c and RETSPL are the corrected and acoustically measured RETSPL of the earbuds, respectively. $\text{HearingLevel}_{ave}$ is the average HL (in units of dB HL) of an otologically normal subject group measured using a clinical audiometer with calibrated standard earphones.

In this study, the aforementioned calibration procedure was implemented with a pair of FlyPods TWS earbuds. (i) First, the output level of the earbuds was calibrated using a head and torso simulator (HATS) (Brüel & Kjær type 5128), which had anatomical pinnae, ear canals, and ear simulators (Brüel & Kjær type 4620). Then the SPL of a pure tone (in dB) was mapped to digital audio magnitude (in root-mean-square) (Elberling & Crone Esmann, 2017). The conversion factor from digital magnitude to SPL at each frequency was:

$$\text{factor} = \frac{\text{ref} \times 10^{\text{SPL}/20}}{\text{RMS}(\text{pure tone})}, \text{ref} = 2.0 \times 10^{-5}. \quad (3)$$

(ii) After electroacoustic calibration, the HLs of 25 otologically normal participants (aged from 18 to 25 years, mean = 20.6 years) were measured twice via the automated audiometer with FlyPods. Provisional RETSPLs were obtained via Equation (1). (iii) 16 of the participants took hearing tests via a conventional audiometer (Otometrics Madsen Astera 2) with a pair of TDH 39 supra-aural earphones. As Table 1 shows, the average hearing thresholds of these

normal-hearing participants did not reach 0 dB HL. (iv) After that, the average HLs of participants using TDH 39 headphones were used to correct the RETSPL values via Equation (2), and then the HL of a specific participant was calculated as:

$$\text{HearingLevel} = \text{SoundLevel} - \text{RETSPL}_c. \quad (4)$$

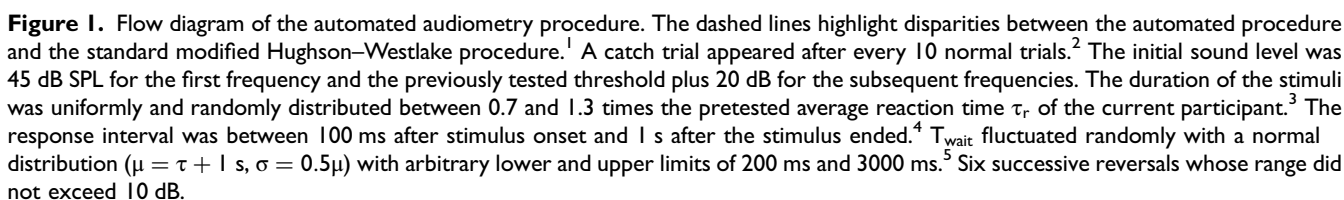
Procedure for Measuring the Hearing Threshold

Two main kinds of methods have been used for PTA: (i) the modified Hughson–Westlake method (Carhart & Jerger, 1959) and (ii) the Békésy method (von Békésy, 1947). The former is often used for manual audiometry in clinical practice. The sound levels of pure tone stimuli are changed according to the participants' responses, and positive responses in ascending sequences determine the hearing thresholds. The Békésy method, which can be used for mass screening (Jerger, 1962), requires the participant to hold down or release a response button when a sound is heard or not, respectively. When collecting responses from participants in automated PTA with adaptive methods, either a *yes-no* task (Swanepoel et al., 2014) or a two-alternative forced-choice (2AFC) task (Schmidt et al., 2014) can be used.

In this study, a specifically designed automated PTA procedure (illustrated later in Figure 1) was developed and the *yes-no* task was used. To simplify the task and reduce the testing time, only a 'heard' button was available to the subjects (Derin et al., 2016) rather than both 'heard' and 'not heard' (Louw et al., 2017; Margolis et al., 2010). The system deemed that a participant had not perceived the pure-tone stimulus (i.e., negative response) unless a response was given within a certain time (i.e., positive response). Since participants were required to respond within a limited time, the latency of the wireless earphones and the automatic procedure were taken into account. Different people may have different reaction speeds, and the variability could be increased by the latency of the audio transmission associated with wireless earphones. In a pilot test (see Appendix), different people with different earphones showed vastly disparate reaction speeds. Therefore, a pre-test phase was used to estimate the participants' "average reaction time" (denoted by τ_r and defined as the time from stimulus onset to response) before the formal HL test. In the pre-test phase, each participant underwent six trials with a 50 dB SPL 1-kHz pure tone (the level can be adjusted to suit the participant). The average reaction time across the last five trials determined the value of τ_r for that participant. Note that the reaction time included the times for both audio transmission latency and participants' judgment. The stimulus duration and the inter-trial silence duration were adjusted according to the value of τ_r , as described below.

Table 1. Average and Standard Deviation (SD) of Hearing Thresholds of 16 Normal-Hearing Participants Measured With TDH 39 Headphones.

		Frequency (Hz)							
		125	250	500	1k	2k	3k	4k	8k
Left	Ave.	10.6	12.5	8.8	8.8	15.6	15.0	11.9	10.6
	SD	6.2	3.7	3.5	6.4	3.2	2.6	4.6	6.2
Right	Ave.	8.8	11.3	8.8	3.8	17.5	16.3	11.9	15.0
	SD	4.4	3.5	3.5	5.8	4.6	5.2	5.3	5.3



In the main test, each ear was tested at eight frequencies: 0.125, 0.25, 0.5, 1, 2, 3, 4, and 8 kHz. The test frequency ascended from 1 to 8 kHz and then descended from 1 to 0.125 kHz. Note that 1 kHz was tested twice. The first run at 1 kHz was used for training, and only the threshold measured at the second time was taken as the formal result.

The flow diagram of measurement for each frequency (denoted a “block”) is shown in Figure 1. In each block, pure-tone stimuli were presented in sequence. The participant was required to respond using the button displayed on the screen. A stimulus and a corresponding response (either positive or negative) are called as a “trial”. The stimulus duration fluctuated randomly around the estimated reaction time of the participant (uniformly distributed between $0.7 \tau_r$ and $1.3 \tau_r$).

The duration was varied to prevent the participant from anticipating the next tone. The initial SPL for the first trial for the first test frequency (i.e., the first 1 kHz) was 45 dB. For the following frequencies except the second 1 kHz tone, the initial level was 20 dB above the HL for the preceding frequency. For the second 1 kHz block, the initial level was 20 dB above the HL for the first 1 kHz block.

The sound level was changed according to the subject’s response on preceding trials. If the participant failed to respond in time (between 100 ms after stimulus onset and 1000 ms after stimulus end), then the sound level of the next stimulus was increased by 5 dB; otherwise, the sound level was decreased by 10 or 5 dB (10 dB until 2 reversals had occurred). A ‘reversal’ is a change from a positive

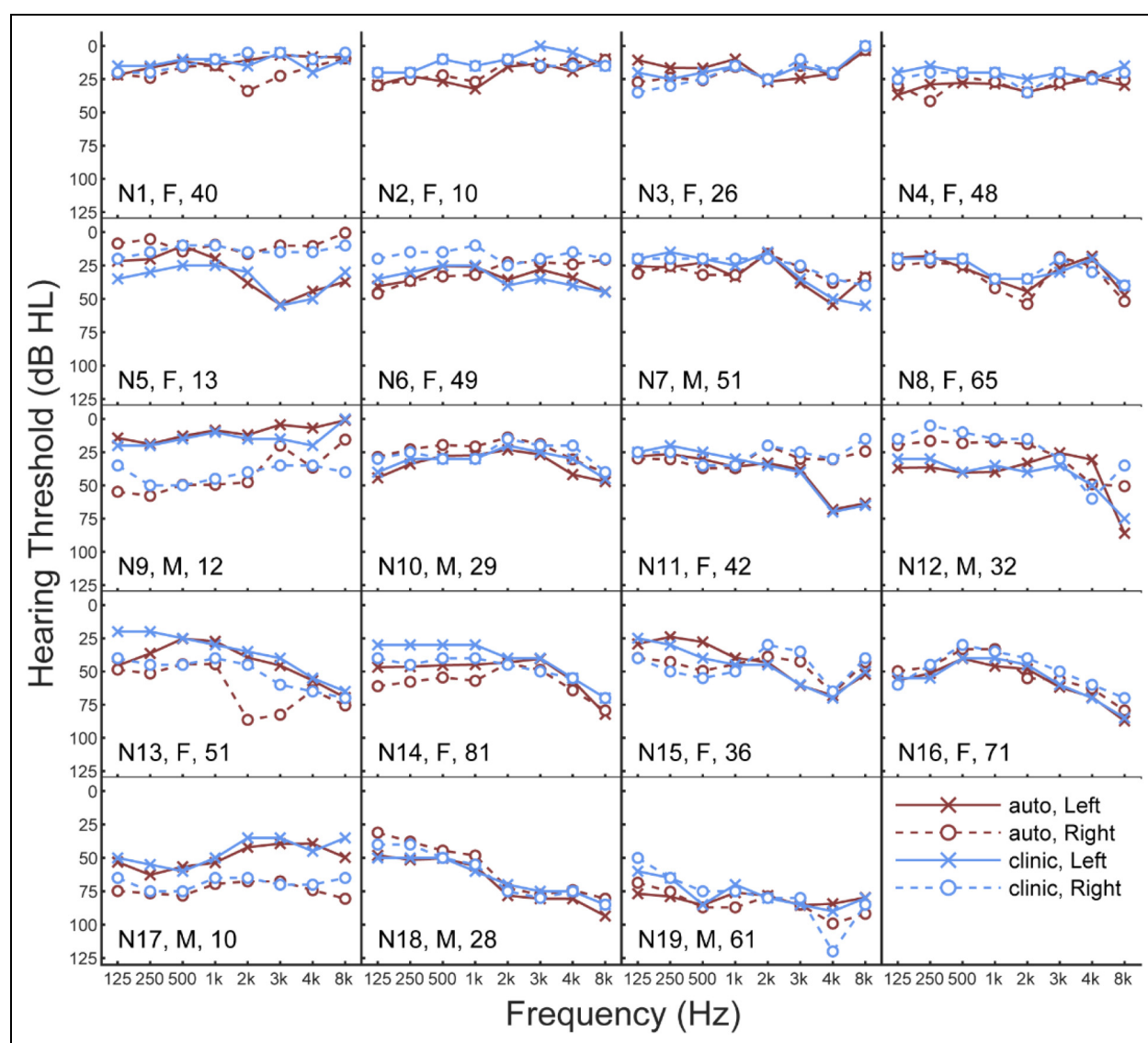


Figure 2. The HLs of the 19 participants. The brown and blue lines illustrate the thresholds obtained via the automated audiometer with FlyPods (auto) and the clinical manual audiometer with TDH 39 headphones (clinic). ‘F’ and ‘M’ denote female and male, respectively. Participant are sorted by the degree of hearing loss. The number after ‘F’ and ‘M’ indicates age.

response to a negative response, and vice versa. Using a larger step in the early trials tends to accelerate convergence.

When the stimulus was playing, it was terminated immediately when the participant responded. The next trial began after a silence with a random duration (T_{wait}). The fluctuating value of T_{wait} prevented the participant from predicting the occurrence of the stimulus; T_{wait} was calculated using a normal distribution (mean $\mu = \tau_r + 1$ s, standard deviation $\sigma = 0.5\mu$), and its upper and lower limits were set arbitrarily as 3000 ms and 200 ms, respectively, to avoid extreme durations.

The procedure ran until six consecutive reversals occurred within a 10 dB range. If the number of trials reached 30, the test block was also terminated. The arithmetic mean of the sound levels at the last four reversals was taken as the HL for that frequency.

There are several differences between our method and the modified Hughson–Westlake method (Figure 1). (i) A catch trial (Schlittenlacher et al., 2018) was inserted after every 10 normal trials. If the participant responded during the 2 s catch trial, a voice prompt (combined with a text prompt) instructing the participant to focus was presented. (ii) If the participant responded during the waiting time, the sound level was increased by 10 dB. The purpose of these two schemes was to decrease the false alarm rates, since a high false alarm rate could lead to an incorrect low threshold (McNicol, 2005). (iii) The estimated reaction time was measured in the pre-test phase and the durations of stimulus, maximum response time restriction, and inter-trial silence were adjusted accordingly.

Experimental Verification of Automated PTA

An experiment was conducted to evaluate the accuracy of the automated PTA with a pair of TWS earbuds compared to conventional PTA with a set of TDH 39 headphones. The clinical audiometer and earphones were the same as those used for calibration.

Participants

Twenty participants (aged from 10 to 81 years, mean = 40.6 years) with mild-to-moderate hearing loss were recruited from FAH-SYU. Each participant took ~20 min to complete the entire experiment including both types of PTA. One participant failed to finish the tests because of severe hearing loss in one ear, and so was excluded from the statistical analysis.

Experimental Procedure

PTA was conducted first using the conventional ascending method and then using automated PTA. PTA was performed

for the left ear first. The earbuds were connected to a Surface Pro 6 laptop running a MATLAB program for the automated PTA. Most of the participants were able to finish the hearing screening without additional guidance. An assistant monitored the whole experimental procedure and offered support if participants encountered any difficulties.

Results of Verification Experiment

The audiograms of all the participants (denoted N1 to N19) are shown in Figure 2. In most cases, there was a high degree of consistency between HLs measured via the two methods (53% of the threshold differences were below 5 dB). Nevertheless, there were large differences in some cases. For instance, large differences were observed for the right ear at 2 and 3 kHz for N1 and N13, whereas HLs for the other frequencies were highly consistent.

The distribution of the HLs is shown in Figure 3. Note that the orthogonal regression line is slightly above the symmetric line, which indicates that the participants tended to have higher HLs (i.e., a higher degree of hearing loss) with the automated PTA than with the standard PTA.

A repeated-measures analysis of variance (ANOVA) was performed on the HLs with three within-subject factors: PTA method, frequency, and ear. No significant effect of PTA method was observed ($F[1, 17] = 0.63$, $p = 0.44$), whereas ear and frequency had significant effects ($F[1, 17] = 10.76$, $p = 0.004$ for ear and $F[1, 17] = 7.92$, $p = 0.01$ for frequency). There was a frequency $\times$ ear interaction ($F[1, 17] = 7.28$, $p = 0.015$). There was no significant three-way interaction ($F[1, 17] = 0.21$, $p = 0.651$). Paired-sample

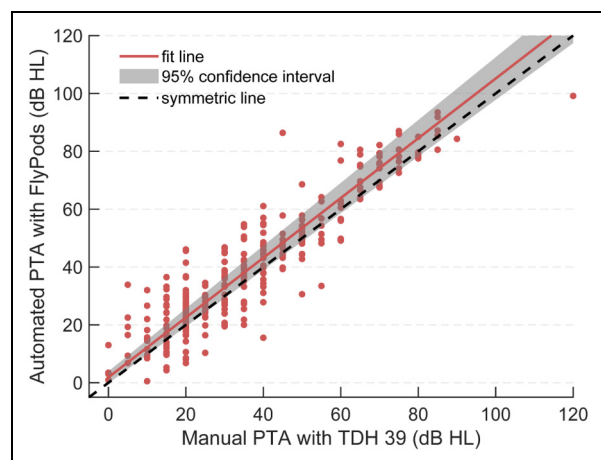


Figure 3. HLs for the automated PTA with flyPods versus those for manual PTA with TDH 39 headphones. Each sample point indicates thresholds measured using the two methods at one frequency from one ear. The solid line and the shaded region are the orthogonal regression line of the samples and its 95% confidence interval. The dashed symmetric line represents an ideal outcome that thresholds measured with the two methods are the same.

Table 2. Average HLs Across Participants Measured With the Two Methods for the Eight Frequencies and for Each Ear.

Ear		Frequency (Hz)							
		125	250	500	1k	2k	3k	4k	8k
L	TDH 39	31.1	30.3	31.1	31.6	34.5	37.1	45.3	45.5
	FlyPods	36.1	34.3	32.1	34.5	36.6	38.4	43.5	48.7
	Absolute Diff.	8.3*	7.4*	5.1	5.7	4.7	4.7	6.0	6.4
	SD	6.6	5.1	5.6	4.2	2.8	4.0	5.5	6.0
	Effect size	0.3	0.2	0.1	0.2	0.1	0.1	-0.1	0.1
R	TDH 39	32.6	33.2	33.2	32.6	33.7	35.0	43.2	40.3
	FlyPods	38.2	37.9	37.2	37.3	39.9	37.7	43.3	43.0
	Absolute Diff.	9.7*	7.6*	6.4*	7.1*	7.5*	5.9	5.1	7.8
	SD	7.1	5.9	5.3	5.7	11.2	6.1	5.1	5.9
	Effect size	0.3	0.2	0.2	0.2	0.3	0.1	0.0	0.1

“Absolute Diff.” and SD denote the average absolute difference between the two PTA methods and corresponding standard deviations. Effect size was calculated with Cohen’s *d*.

*Differences at a significance level of 0.05 under paired-sample *t*-tests.

t-tests were conducted to further compare HLs measured for individual frequencies and individual ears with the two methods. The average HLs measured with each method, absolute threshold deviation between the two method,

standard deviation, and effect size are given in Table 2. The automated PTA HLs were significantly higher than the clinical manual PTA HLs at 125 Hz ($t[18] = 2.21$, $p = 0.04$), 250 Hz ($t[18] = 2.43$, $p = 0.025$), 500 Hz ($t[18] =$

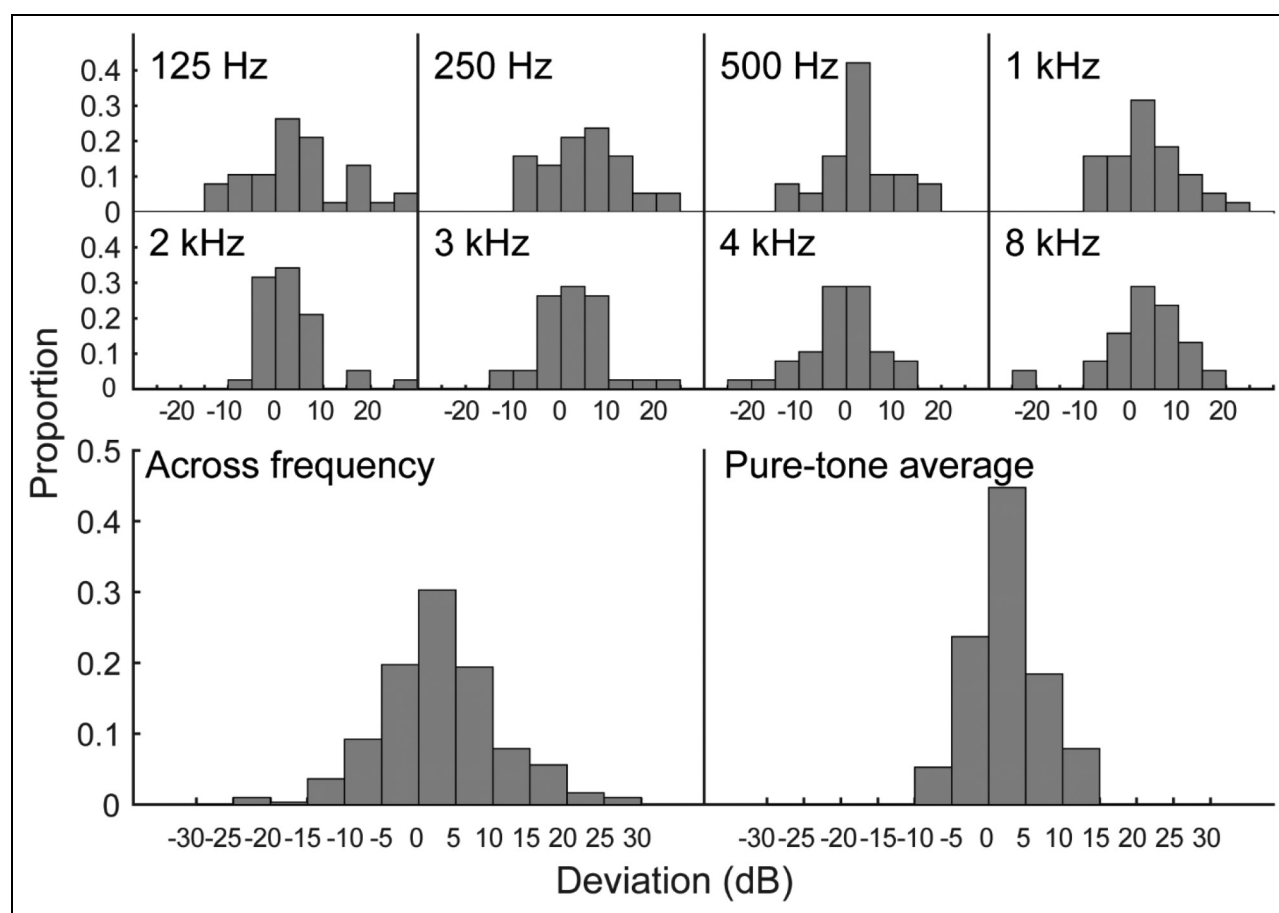


Figure 4. Distribution of the deviations between HLs for two methods. Deviations are the HLs for FlyPods minus those for TDH 39. The eight top subfigures present the deviations for each frequency. The two bottom subfigures present deviations for all HLs across frequencies and deviations of the pure-tone average HL for the frequencies 0.5, 1, 2, and 4 kHz.

Table 3. Accuracy, Sensitivity, and Specificity of the Automated Audiometry With FlyPods.

	Frequency (Hz)								Overall
	125	250	500	1k	2k	3k	4k	8k	
Accuracy	0.68	0.68	0.74	0.76	0.89	0.84	0.97	0.92	0.81
Sensitivity	0.95	0.84	0.94	0.95	1.00	0.95	1.00	0.96	0.95
Specificity	0.42	0.53	0.55	0.50	0.75	0.69	0.92	0.83	0.63

For the 304 thresholds measured with standard PTA, there were 131 *normal hearing* (≤ 25 dB HL) values and 173 *impaired hearing* (>25 dB HL).

2.39, $p = 0.028$), 1 kHz ($t[18] = 2.52$, $p = 0.021$), and 2 kHz ($t[18] = 2.27$, $p = 0.036$) for the right ear, and at 125 Hz ($t[18] = 2.32$, $p = 0.032$) and 250 Hz ($t[18] = 2.21$, $p = 0.042$) for the left ear.

To give a more intuitive view of how the results differ, the distributions of the automated–manual deviations are presented in Figure 4. The average deviation (Masalski et al., 2018) across frequencies and ears was 3.1 dB with a standard deviation (SD) of 8.4 dB, whereas the average absolute threshold deviation was 6.7 dB with an SD of 6.0 dB (see Table 2). The proportion of absolute deviations below 5 and 10 dB for all thresholds was 53% and 78%, respectively. There was a relatively large difference between the two methods at 125 and 250 Hz, the proportions of deviations below 5 dB at these two frequencies being 37% and 34%, respectively. For the pure-tone average, which is typically used as an evaluation metric for hearing (the average of thresholds at 0.5 and 1, 2, and 4 kHz, Ho et al., 2009), 63% and 92% of the absolute deviations were below 5 and 10 dB, respectively.

In present study, having a HL above 25 dB HL was considered as *impaired hearing* (World Health Organization, 1991). With this criterion, sensitivity, specificity, and accuracy were calculated (see Table 3) with reference to HLs obtained with manual PTA. Sensitivity is defined as the proportion of correct classification as *impaired hearing*, and specificity is defined as the proportion of correct classification as *normal hearing*. Accuracy is the overall proportion of correct classification both as *impaired hearing* and as *normal hearing*. The overall accuracy across frequencies was 81% and the results were similar to those from the HL difference analysis, that is, accuracy was higher at high frequencies.

Time Cost of Automated PTA

The time cost for automated PTA mainly depends on the response time for each trial and the number of trials in each test block. The response time depends on the estimated average reaction time obtained in the pre-test phase. The average response time across all participants and trials was 1.86 s (SD = 0.60). Figure 5 shows the distribution of the number of trials across all frequencies. The participants required 11.2 trials per block on average (median = 11, SD = 4.4). Overall, 91% of the blocks needed no more than 15

trials. Only 2% of the blocks reached the predefined ceiling of 30 trials. Responses were made on 13% of catch trials.

The number of trials was not normally distributed according to the Anderson–Darling test. Therefore, a Friedman test was performed to examine whether the number of trials depended on the frequency. There was no significant difference in the number of trials across frequencies when excluding the training block. There was a significant difference between the training block and the 125 Hz block ($p = 0.046$). The medians for these two blocks were 12 and 9, respectively.

Discussion

In this study, automated PTA was developed using commercially available TWS earbuds (Honor FlyPods). Wireless earphones of this type have achieved considerable popularity in recent years, but few studies have evaluated their feasibility for PTA.

In the procedure for estimating HLs, a user-dependent maximum effective response time, which is evaluated in a pre-test phase, is used to allow for different audio latencies

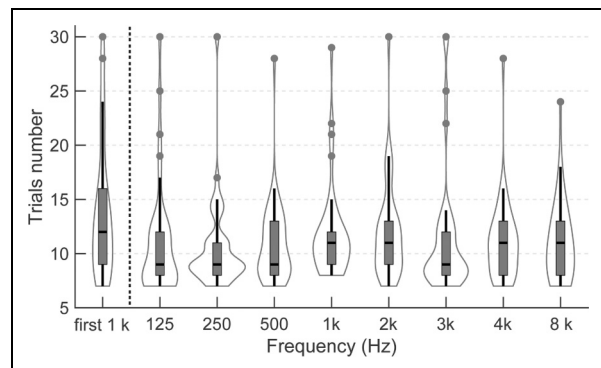


Figure 5. Box-and-whisker plots and violin plots of the number of trials for each frequency. The upper, median, and lower horizontal lines of each box indicate the 75 percentile, the median, and the 25 percentile, respectively. The whiskers attached to the boxes represent the most extreme data that are below 1.5 times the interquartile range (IQR). The widths of the violin plots depict distribution density of the trial number. Symbols indicate extreme cases far from the data centers. The first box on the left is for the training block, which was not taken into account in the threshold evaluation.

and reaction speeds of the participants. A short maximum effective response time could make the test difficult for the participants, leading to an overestimated proportion of negative responses. Participants can hardly keep focused for a long time, and thus a long (Barbour et al., 2019) or even no time restriction (Margolis et al., 2010) may be unacceptable to them. Moreover, due to the additional latency of wireless earphones, the automated program may misjudge if participants do not respond in time. Thus, it is reasonable to adjust the maximum effective response time for each person and each set of earphones before testing.

Automated PTA is often conducted without supervision from audiologists, and so the cooperativeness of the participants can influence the results, especially for children and older people (Yeung et al., 2013). However, a widely accepted supervision scheme is yet to be established for self-screening automated PTA. The proposed procedure included catch trials (Margolis et al., 2010) and responses were detected during a waiting period. In the verification experiment, responses were made on 13% (22 of 175) of the catch trials. This indicates that the participants may respond even when no tone is presented.

To calibrate the FlyPods earbuds, a Brüel & Kjær type 5128 HATS was used and the corresponding provisional RETSPLs were determined from 25 participants with normal hearing. It has been reported that there are differences in the average normal HLs of different populations (Borchgrevink et al., 2005; Wang et al., 2018), and thus an audiometric zero level bias could exist between our participants and those using standard earphones. Of the otologically normal young participants, 16 out of 25 them underwent PTA using an Otometrics Madsen Astera 2 audiometer with TDH 39 headphones. Their average HL was used to compensate for the bias. It could be argued that it is unnecessary to perform an electroacoustic calibration because the RETSPLs were corrected to the levels of a standard earphone. A simpler calibration method would be to adjust the output level of non-standard earbuds (in units of dB SPL) to the output level of standard earphones driven by clinical audiometers (in units of dB HL) via calibrators directly rather than by obtaining a set of corrected RETSPLs (Kam et al., 2013). The purpose of the preliminary electroacoustic calibration is that other earbuds with similar enclosures could be calibrated through electroacoustic measurement and RETSPLs easily, without requiring additional behavioral tests.

In the present study, the average absolute deviation between HLs measured using automated and standard PTA was 6.6 dB across all frequencies and both ears. The proportions of absolute deviations below 5 and 10 dB were 53% and 78%, respectively. The threshold deviations between the automated PTA with FlyPods and manual PTA with the standard device increased at 125 and 250 Hz. Masalski et al. (2018) reported a similar phenomenon. Moreover, the values of RETSPL at low frequencies vary dramatically for

different devices (Vencovsky et al., 2018). Given that the output level was adjusted to match the level for a conventional audiometer, the lower accuracy at 125 and 250 Hz was unexpected. The error at low frequencies may be due to the earphone leakage (Welti, 2015), since the state of contact between earbuds and ears varies dramatically between participants compared with TDH 39 headphone, which have soft cushions. In addition, automated PTA tended to give slightly higher HLs (see Figure 3). Compared with participants in the calibration procedure, a lower average capacity of attention of participants in the verification experiment may have caused this bias, since these participants' responses during the waiting intervals led to increments of the sound level.

In tests with participants, there are often differences between two repeated tests even with the same device and under the same conditions. Margolis et al. (2010) reported that the average absolute deviation between *test* and *retest* for clinical PTA was 4.1 dB. A. T. Ho et al. (2009) reported that *test-retest* absolute deviations below 5 and 10 dB for clinical PTA occurred 74% and 93% of the time, respectively. Swanepoel et al. (2010) found that absolute deviations below 5 and 10 dB occurred 45% and 88% of the time, respectively. In summary, the HLs obtained using automated PTA with FlyPods were slightly higher than those obtained using clinical PTA with TDH 39 headphone at low frequencies, but the errors were comparable to those for conventional manual PTA.

The sensitivity and specificity across frequencies for the automated PTA with FlyPods were 95% and 63%, respectively. The same two metrics were 82% and 83% in Louw et al. (2017) using hearScreenTM with Sennheiser HD 202 II, and 25% and 97% in Derin et al. (2016) using the EarTrump application with a Philips SHP 1900 headset. In Thompson et al. (2015), the sensitivity and specificity varied with frequency over the ranges 87%–100% and 90%–95%, respectively. They used TDH 39 headphones for both the clinical audiometer and the automatic audiometer application. In another comparative study, Yeung et al. (2013) also used TDH 39 headphones with automated PTA, and obtained 93% sensitivity and 95% specificity. Moreover, in a test using ER-3 earphones for both automated and clinical PTA, 89% of HL differences were below 5 dB (Mosley et al., 2019). It seems that standard earphones are more likely to give HLs consistent with those for clinical audiometry. Compared with these studies, especially those using standard earphones, the sensitivity of the automated PTA proposed here is high, but the specificity is lower.

A median of 11 trials was needed to complete a test block, and participants took an average of 1.86 s to respond to each stimulus. The first training block required more trials because the initial sound level was usually far from the threshold. The closer the initial level was to the HL of the participants, the fewer trials were required, so setting the initial level according to the preceding measured threshold could help to

accelerate the measurements (Carhart & Jerger, 1959). Thus, adopting methods based on statistical estimation or machine learning (Barbour et al., 2019; Schlittenlacher et al., 2018; Schmidt et al., 2014; Song et al., 2015) could further reduce the time cost.

The present study describes a method for using FlyPods TWS earbuds in automated PTA. The procedure could guide the development of an audiometer application and calibration. The results show that the system can provide accurate HL estimation without relying on an audiologist with appropriate design and calibration. Although the proposed system cannot replace conventional PTA entirely, people whose hearing is potentially impaired could take such a self-screening test regularly. Moreover, large clinical centers could benefit from automated audiometers by using them as a prescreening tool, which may release the audiologist from checking normal-hearing individuals. In particular, this form of contactless remote PTA could continue to provide hearing evaluations during the COVID-19 pandemic. Moreover, self-screened audiograms could be utilized in tuning self-fitting hearing aids.

Limitations and Outlook

The proposed automated PTA proposed here did not use a noise detector (Swanepoel et al., 2014), but in actual applications, a reliable noise detector is essential for ensuring that the automated PTA is performed in a suitable environment. Also, the automated PTA did not use contralateral masking when there was a large (≥ 30 dB) left–right threshold disparity (Margolis et al., 2010; Swanepoel et al., 2010). However, the hearing-impaired participants involved in this study had few asymmetric HLs (only 9 of 152 samples had an across-ear difference larger than 30 dB), so this limitation had little effect on the results.

The present study assessed the accuracy of automated PTA with FlyPods TWS earbuds through a subjective experiment. Nonetheless, complete electroacoustic output measurements (Van der Aerschot et al., 2016) of the FlyPods were not made, nor was the consistency across multiple pairs of the same type of earbuds evaluated. Variability of the output level could lead to additional errors in HL measurements. However, compared with the deviations between test and retest of PTA, modern earphones often have acceptably low output variability. Foulad et al. (2013) reported that the output-level differences among pairs of earphones were within 4 dB.

A HATS was used for the electroacoustic calibration since no suitable ear simulator or acoustic coupler was available for the earbuds used in this study. However, a HATS with anatomical pinnae is less robust and reliable than ear simulators, because the coupling is less reliable.

During the verification experiment, all participants were tested using manual PTA first and automated PTA later. This unbalanced test order may have led to biased results.

Another limitation is that the number of participants, 20, was relatively small.

Given the aforementioned limitations, future work should consider (i) monitoring the noise environment, (ii) contralateral masking for large left–right disparities, (iii) comprehensive electroacoustic measurements and checking consistency across multiple devices, (iv) balancing the order of manual and automated testing, and (v) recruiting more participants.

Conclusions

This study utilized FlyPods TWS earbuds in automated PTA and verified the feasibility and validity of the method through an experiment. The verification experiment showed that with reasonable calibration and PTA procedure design, the proposed self-screening PTA gave results consistent with those from clinical PTA. The average signed and absolute deviations between the two types of PTA were 3.1 dB and 6.7 dB, respectively, but there were large differences at 125 and 250 Hz. For hearing loss detection, the overall accuracy was 81%, but was lower at low frequencies. Although a further performance evaluation may be needed for low frequencies, the present experiment shows that it is feasible to conduct automated air-conduction PTA with affordable TWS earbuds.

Acknowledgments

We thank Z. Ellen Peng, and Changxin Zhang for suggestions on manuscript writing. We also express our gratitude to Brian C. J. Moore, Robert H. Margolis, and an anonymous reviewer for their kind revision and helpful comments. We express our gratitude to all the participants in this study.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the HUAWEI corporation, the Guangdong Basic and Applied Basic Research Foundation (grant number 2020A1515010386), and the Science and Technology Program of Guangzhou (grant number 202102020944).

ORCID iDs

Zhenyu Guo  <https://orcid.org/0000-0003-0946-3899>

Xianren Wang  <https://orcid.org/0000-0002-6760-3258>

References

- ANSI/ASA S3.6-2018 (2018). *Specification for Audiometers*. (American National Standards Institute/Acoustical Society America).
- Barbour D. L., Howard R. T., Song X. D., Metzger N., Sukesan K. A., DiLorenzo J. C., Snyder B. R. D., Chen J. Y., Degen E. A.,

- Buchbinder J. M., & Heisey K. L., (2019). Online machine learning audiometry. *Ear and Hearing*, 40(4), 918–926. <https://doi.org/10.1097/AUD.0000000000000669>
- Bluetooth SIG (2015). *Advance Audio Distribution Profile Specification (version 1.3)*. Retrieved from <https://www.bluetooth.com/specifications/specs/>
- Borchgrevink H., Tambs K., & Hoffman H. (2005). The Nord-Trøndelag Norway audiometric survey 1996–98: Unscreened thresholds and prevalence of hearing impairment for adults >20 years. *Noise and Health*, 7(28), 1. <https://doi.org/10.4103/1463-1741.31632>
- Bright T., & Pallawela D. (2016). Validated smartphone-based apps for ear and hearing assessments: A review. *JMIR Rehabilitation and Assistive Technologies*, 3(2), e13. <https://doi.org/10.2196/rehab.6074>
- Burkhard M. D., & Sachs R. M. (1975). Anthropometric manikin for acoustic research. *The Journal of the Acoustical Society of America*, 58(1), 214–222. <https://doi.org/10.1121/1.380648>
- Carhart R., & Jerger J. F. (1959). Preferred method for clinical determination of pure-tone thresholds. *Journal of Speech and Hearing Disorders*, 24(4), 330–345. <https://doi.org/10.1044/jshd.2404.330>
- Centers for Disease Prevention and Control (2020). *Interim US guidance for risk assessment and work restrictions for health-care personnel with potential exposure to COVID-19*.
- Choi J. M., Lee H. B., Park C. S., Oh S. H., & Park K. S. (2007). PC-based tele-audiometry. *Telemedicine Journal and E-Health*, 13(5), 501–508. <https://doi.org/10.1089/tmj.2007.0085>
- Chu Y. C., Cheng Y. F., Lai Y. H., Tsao Y., Tu T. Y., Young S. T., Chen T. S., Chung Y. F., Lai F., & Liao W. H., (2019). A mobile phone-based approach for hearing screening of school-age children: Cross-sectional validation study. *JMIR mHealth and UHealth*, 7(4), e12033. <https://doi.org/10.2196/12033>
- Corry M., Sanders M., & Searchfield G. D. (2017). The accuracy and reliability of an app-based audiometer using consumer headphones: Pure tone audiometry in a normal hearing group. *International Journal of Audiology*, 56(9), 706–710. <https://doi.org/10.1080/14992027.2017.1321791>
- Derin S., Cam O. H., Beydilli H., Acar E., Elicora S. S., & Sahan M. (2016). Initial assessment of hearing loss using a mobile application for audiological evaluation. *The Journal of Laryngology & Otolaryngology*, 130(3), 248–251. <https://doi.org/10.1017/S0022215116000062>
- Elberling C., & Crone Esmann L. (2017). Calibration of brief stimuli for the recording of evoked responses from the human auditory pathway. *The Journal of the Acoustical Society of America*, 141(1), 466. <https://doi.org/10.1121/1.4974145>
- Floros A., Koutroubas M., Tatlas N.-A., & Mourjopoulos J. (2002). A study of wireless compressed digital-audio transmission. In *Audio engineering society convention*. Audio Engineering Society. 112 (p. 5516).
- Foulad A., Bui P., & Djalilian H. (2013). Automated audiometry using apple iOS-based application technology. *Otolaryngology–Head and Neck Surgery*, 149(5), 700–706. <https://doi.org/10.1177/0194599813501461>
- Ho A. T., Hildreth A. J., & Lindsey L. (2009). Computer-assisted audiometry versus manual audiometry. *Otology & Neurotology*, 30(7), 876–883. <https://doi.org/10.1097/MAO.0b013e3181b120d0>
- Ho C.-Y., Li P. C., & Young S. T. (2017). Reference equivalent threshold sound pressure levels for Apple EarPods. *The Journal of the Acoustical Society of America*, 141(2), EL115. <https://doi.org/10.1121/1.4976110>
- Honeth L., Bexelius C., Eriksson M., Sandin S., Litton J. E., Rosenhall U., Nyren O., & Bagger-Sjoberg D., (2010). An internet-based hearing test for simple audiometry in nonclinical settings: Preliminary validation and proof of principle. *Otology & Neurotology*, 31(5), 708–714. <https://doi.org/10.1097/MAO.0b013e3181de467a>
- ISO 389-8: 2004 (2004). Acoustics: Reference zero for the calibration of audiometric equipment. Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones. (International Organization for Standardization).
- ISO 389-1: 2007 (2007). *Acoustics-Reference zero for the calibration of audiometric equipment: Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones*. (International Organization for Standardization).
- Jerger J. (1962). Békésy audiometry. *International Audiology*, 1(2), 160–164. <https://doi.org/10.3109/05384916209074031>
- Kam A. C. S., Gao H., Li L. K. C., Zhao H., Qiu S., & Tong M. C. F. (2013). Automated hearing screening for children: A pilot study in China. *International Journal of Audiology*, 52(12), 855–860. <https://doi.org/10.3109/14992027.2013.832419>
- Kam A. C., Sung J. K., Lee T., Wong T. K., & van Hasselt A. (2012). Clinical evaluation of a computerized self-administered hearing test. *International Journal of Audiology*, 51(8), 606–610. <https://doi.org/10.3109/14992027.2012.688144>
- Kanji A., Khoza-Shangase K., & Moroe N. (2018). Newborn hearing screening protocols and their outcomes: A systematic review. *International Journal of Pediatric Otorhinolaryngology*, 115, 104–109. <https://doi.org/10.1016/j.ijporl.2018.09.026>
- Keidser G., & Convery E. (2016). Self-Fitting hearing aids: Status Quo and future predictions. *Trends in Hearing*, 20, <https://doi.org/10.1177/2331216516643284>
- Larson V. D., Cooper W. A., Talbott R. E., Schwartz D. M., Ahlstrom C., & De Chicchis A. R. (1988). Reference threshold sound-pressure levels for the TDH-50 and ER-3A earphones. *The Journal of the Acoustical Society of America*, 84(1), 46–51. <https://doi.org/10.1121/1.396874>
- Louw C., Swanepoel W., Eikelboom R. H., & Myburgh H. C. (2017). Smartphone-Based hearing screening at primary health care clinics. *Ear and Hearing*, 38(2), e93–e100. <https://doi.org/10.1097/AUD.0000000000000378>
- Mahomed F., Swanepoel de W., Eikelboom R. H., & Soer M. (2013). Validity of automated threshold audiometry: A systematic review and meta-analysis. *Ear and Hearing*, 34(6), 745–752. <https://doi.org/10.1097/01.aud.0000436255.53747.a4>
- Manheim M., Lavie L., & Banai K. (2018). Age, hearing, and the perceptual learning of rapid speech. *Trends in Hearing*, 22. <https://doi.org/10.1177/2331216518778651>
- Margolis R. H., Bratt G., Feeney M. P., Killion M. C., & Saly G. L. (2018). Home hearing test™: Within subjects threshold variability. *Ear and Hearing*, 39(5), 906–909. <https://doi.org/10.1097/AUD.0000000000000551>
- Margolis R. H., Glasberg B. R., Creeke S., & Moore B. C. (2010). AMTAS: Automated method for testing auditory sensitivity: Validation studies. *International Journal of Audiology*, 49(3), 185–194. <https://doi.org/10.3109/14992020903092608>
- Margolis R. H., Killion M. C., Bratt G. W., & Saly G. L. (2016). Validation of the home hearing test. *Journal of the American*

- Academy of Audiology*, 27(5), 416–420. <https://doi.org/10.3766/jaaa.15102>
- Margolis R. H., & Morgan D. E. (2008). Automated pure-tone audiometry: An analysis of capacity, need, and benefit. *American Journal of Audiology*, 17(2), 109–113. [https://doi.org/10.1044/1059-0889\(2008/07-0047\)](https://doi.org/10.1044/1059-0889(2008/07-0047))
- Masalski M., Grysinski T., & Krecicki T. (2018). Hearing tests based on biologically calibrated Mobile devices: Comparison With pure-tone audiometry. *JMIR mHealth and UHealth*, 6(1), e10. <https://doi.org/10.2196/mhealth.7800>
- McNicol D. (2005). *A primer of signal detection theory*. Psychology Press.
- Mosley C. L., Langley L. M., Davis A., McMahon C. M., & Tremblay K. L. (2019). Reliability of the home hearing test: Implications for public health. *Journal of the American Academy of Audiology*, 30(3), 208–216. <https://doi.org/10.3766/jaaa.17092>
- Na Y., Joo H. S., Yang H., Kang S., Hong S. H., & Woo J. (2014). Smartphone-based hearing screening in noisy environments. *Sensors (Basel)*, 14(6), 10346–10360. <https://doi.org/10.3390/s140610346>
- Paglalionala A., Tognola G., & Pincioli F. (2015). Apps for hearing science and care. *American Journal of Audiology*, 24(3), 293–298. https://doi.org/10.1044/2015_AJA-14-0093
- Paquier M., Koehl V., & Jantzen B. (2016). Effect of headphone position on absolute threshold measurements. *Applied Acoustics*, 105, 179–185. <https://doi.org/10.1016/j.apacoust.2015.12.003>
- Schlittenlacher J., Turner R. E., & Moore B. C. J. (2018). Audiogram estimation using Bayesian active learning. *The Journal of the Acoustical Society of America*, 144(1), 421. <https://doi.org/10.1121/1.5047436>
- Schmidt J. H., Brandt C., Pedersen E. R., Christensen-Dalsgaard J., Andersen T., Poulsen T., & Baelum J. (2014). A user-operated audiometry method based on the maximum likelihood principle and the two-alternative forced-choice paradigm. *International Journal of Audiology*, 53(6), 383–391. <https://doi.org/10.3109/14992027.2013.879339>
- 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 and Hearing*, 36(6), e326–e335. <https://doi.org/10.1097/AUD.000000000000186>
- Swanepoel D. W., Mngemane S., Molemong S., Mkwanazi H., & Tutshini S. (2010). Hearing assessment-reliability, accuracy, and efficiency of automated audiometry. *Telemedicine Journal and E-Health*, 16(5), 557–563. <https://doi.org/10.1089/tmj.2009.0143>
- Swanepoel D. W., Myburgh H. C., Howe D. M., Mahomed F., & Eikelboom R. H. (2014). Smartphone hearing screening with integrated quality control and data management. *International Journal of Audiology*, 53(12), 841–849. <https://doi.org/10.3109/14992027.2014.920965>
- Thompson G. P., Sladen D. P., Borst B. J., & Still O. L. (2015). Accuracy of a tablet audiometer for measuring behavioral hearing thresholds in a clinical population. *Otolaryngology–Head and Neck Surgery*, 153(5), 838–842. <https://doi.org/10.1177/0194599815593737>
- Van der Aerschot M., Swanepoel de W., Mahomed-Asmail F., Myburgh H. C., & Eikelboom R. H. (2016). Affordable headphones for accessible screening audiometry: An evaluation of the Sennheiser HD202 II supra-aural headphone. *International Journal of Audiology*, 55(11), 616–622. <https://doi.org/10.1080/14992027.2016.1214756>
- Van Tasell D. J., & Folkeard P. (2013). Reliability and accuracy of a method of adjustment for self-measurement of auditory thresholds. *Otology & Neurotology*, 34(1), 9–15. <https://doi.org/10.1097/MAO.0b013e318278c05d>
- Vencovsky V., Rund F., & Slegl D. (2018). Reference equivalent threshold sound pressure levels for nonaudiometric headphones. *Journal of the Audio Engineering Society*, 66(3), 167–171. <https://doi.org/10.17743/jaes.2018.0009>
- von Békésy G. (1947). A new audiometer. *Acta Oto-Laryngologica*, 35(5-6), 411–422. <https://doi.org/10.3109/00016484709123756>
- Wang J., Qian X., Chen J., Yang Y., & Gao X. (2018). A cross-sectional study on the hearing threshold levels among people in Qinling, Qinghai, and Nanjing, China. *American Journal of Audiology*, 27(1), 147–155. https://doi.org/10.1044/2017_aja-16-0053
- Wayne R. V., & Johnsrude I. S. (2015). A review of causal mechanisms underlying the link between age-related hearing loss and cognitive decline. *Ageing Research Review*, 23(Pt B), 154–166. <https://doi.org/10.1016/j.arr.2015.06.002>
- Welti T. (2015). Improved measurement of leakage effects for circum-aural and supra-aural headphones. In *Audio engineering society convention*. Audio Engineering Society, 138.
- Wilber L. A., Kruger B., & Killion M. C. (1988). Reference thresholds for the ER-3A insert earphone. *The Journal of the Acoustical Society of America*, 83(2), 669–676. <https://doi.org/10.1121/1.396162>
- Wilson B. S., Tucci D. L., Merson M. H., & O'Donoghue G. M. (2017). Global hearing health care: New findings and perspectives. *The Lancet*, 390(10111), 2503–2515. [https://doi.org/10.1016/s0140-6736\(17\)31073-5](https://doi.org/10.1016/s0140-6736(17)31073-5)
- World Health Organization (1991). Report of the informal working group on prevention of deafness and hearing impairment programme planning. *Geneva: WHO*, 18–21.
- Yeung J., Javidnia H., Heley S., Beauregard Y., Champagne S., & Bromwich M. (2013). The new age of play audiometry: Prospective validation testing of an iPad-based play audiometer. *J Otolaryngology–Head and Neck Surgery*, 42, 21. <https://doi.org/10.1186/1916-0216-42-21>

Appendix

Variation of Reaction Times for Different Subjects and Devices

In conventional manual PTA, the audiologist can adjust the maximum response time according to participant's reaction time. For automated PTA, it is necessary to estimate the reaction times of participants wearing specific earphones. Consequently, it is meaningful to study the variation of reaction time. Before designing the automated audiometer, a preliminary experiment was carried out to assess the variation in reaction times for different participants and devices.

Experimental Procedure and Participants

Forty participants (aged from 11 to 57 years, mean = 27 years) were recruited remotely through the Internet, and the automated PTA software was sent to them via the Internet. Each was required to perform the hearing test three times using their own device and earphones in a quiet environment. The reaction times estimated from the pre-test phase were recorded.

Statistics for Reaction Times

Of the 40 participants, 26 submitted their test results. The final reaction time for each participant was determined as the average across the three repeated tests. Figure 6 shows the distribution of reaction times. The average reaction time was 968 ms with a standard deviation of 695 ms and a range from 450 to 3201 ms. Further, 42% of the reaction times were in the range 500–750 ms. The results show that different devices and participants may have very different reaction times. Therefore, using a fixed effective response time in automated PTA could lead to inaccuracies if the reaction time of a participant is larger

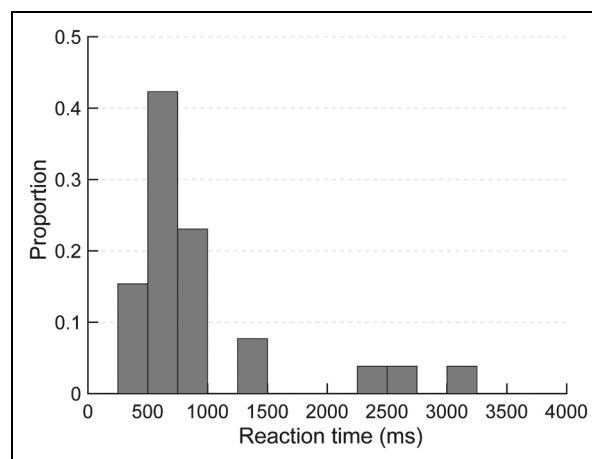


Figure 6. Reaction time measured in the pre-test phase for 26 participants. The x-axis shows the average time from the start of a stimulus to when the participant responded.

than the limit. Note that the reaction time may depend on the device, so the values reported here cannot be used as reference data.