Buenos Aires – 5 to 9 September, 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Psychological and Physiological Acoustics (others): Paper ICA2016-114

Evaluation of Apple iOS-based automated audiometry

Yuan Xing, Zhen Fu, Xihong Wu, Jing Chen

Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China
xing_yuan@pku.edu.c, fuzhenfz@126.com, wxh@cis.pku.edu.cn, chenj@cis.pku.edu.cn

Abstract

Audiometry has been widely used for assessing hearing situation in clinic. With the development of smart phones, several applications based on Apple iOS can be used to test audiometric thresholds automatically. However, the reliability of these tests was rarely studied in previous works. In this work, an iPhone-based automated puretone testing application was made, in which a standard procedure (STA) referring to ISO 8253-1 and a relatively quicktesting procedure (QT) were both implemented. A human behavior experiment was designed and conducted to evaluate the accuracy and efficiency of the audiometry on iPhone. Two factors were manipulated for iPhone: test environment (sound booth vs. normally quiet room), test procedure (STA vs. QT). And a conventional audiometry on audiometric equipment (Madsen, AURICAL) was also conducted as a control condition. Additionally, the test-retest reliability was also measured by repeating the 5 tests (2×2+1) on a different day. Eight young university students took part in this experiment, and they were all tested on both ears. The test orders were balanced across participants. The experimental results show 1) hearing thresholds tested in the sound booth are significantly lower than in the normally quiet room by about 3 dB HL; 2) there is no significant difference between STA and QT for hearing thresholds, but the test duration is significantly less for QT (mean 349 seconds) than SAT (mean 568 seconds); 3) there is no significant difference between the tests and retests both for hearing thresholds and test durations; 4) comparing to the results on audiometer, hearing thresholds are significantly lower and test durations are significantly less for the iPhone application. The differences observed are further analyzed and discussed.

Keywords: Audiometry, Hearing Threshold, Apple iOS, iPhone

Evaluation of Apple iOS-based automated audiometry

1 Introduction

Hearing loss is a serious public health problem. It’s reported that there are more than 20 million people suffering from hearing loss in China, and the population with hearing loss keeps growing with the aging of society [1]. Pure tone audiometry (PTA) is one of the most popular method of hearing assessment, determining the minimum sound level can be heard at each test frequency by listeners. PTA uses both air and bone conduction audiometry, depending on the equipment used. Only air conduction was considered in this paper.

Manual pure-tone audiometry is the earliest and most popular audiometry, in which the audiologist operates the audiometer to determine listener’s hearing threshold according to listener’s feedback. The ascending method defined in ISO 8253-1 is usually used for PTA [2]. The procedure of audiometry is as follows. The ear with better hearing is tested first followed by the contralateral ear. The test order of frequency is from 1 kHz upwards to 8 kHz, and then from 250 Hz to 1 kHz in an ascending order, indicating the 1 kHz is tested twice. The hearing threshold at 1 kHz need to be re-tested if the difference between the two measures is above 5 dB. Two procedures are specified for the presentation order of sound level: an ascending method and a bracketing method. It has been reported that the test results are similar for these two methods but test duration is less for the ascending than for the bracketing method [3]. For the ascending method, the sound level is increased by 5-dB step from a weak sound. Once listeners report a sound is heard, the sound level is decreased by 10-dB step until it’s reported that no sound is heard. And then the second ascending circulation begins. This procedure is repeated until the same-level sound is heard three times or five circulations are finished. The level of the same sound which has been heard most frequently is defined as the hearing threshold level. ISO 8253-1 also defined a shortened version of the ascending method. In the shortened version, the hearing threshold is determined when the same-level sound is heard twice or three circulations are finished. The shortened version is less time-consuming than the original version, and has been shown to yield nearly equivalent results.

As audiologists dominate the test progress of the manually PTA, they probably introduce subjective errors for results. With the development of technologies, automatic audiometry is invented, which is controlled by computer programs. Listeners input feedback in a certain form into the program and the parameters of stimuli would be adjusted automatically according to the feedbacks. Some programs adopted the ascending method consistent with the manual audiometry, but others used different methods [4]. Previous studies have shown that accuracy and reliability of automatic audiometry are approximate to manual audiometry.

Recently, there is a growing interest in implementing automated audiometry on smartphones. The widespread usage of smartphones provides automated audiometry to people who lack the resources of manual audiometry. Several works on smartphone-based audiometry have been reported. An iOS App named uHear (Unitron, Canada) provides audiometry using the ascending method defined in ISO 8253-1. Szudek studied the performance of uHear [5]. The result showed that uHear could rule out general degree of hearing loss, though uHear tended to overestimated hearing thresholds. Van Tasell and Folkeard implemented two audiometry methods on an iPad [6]. One method was a software-controlled adaptive method, the other was a user-controlled method in which users adjusted test tones to threshold. The reliability and accuracy of methods were evaluated, and it was found that the differences of hearing threshold between the manual audiometry and both methods were within a reasonable range. Allen developed an audiometry App called EarTrumpet, which used the ascending method in ISO 8253-1 [7]. The standardized output of sound intensity from different Apple devices was compared, and the feasibility of the EarTrumpet was evaluated. The result showed that the differences of sound intensity across various Apple iPhone, iPod touch and iPad devices were within 4 dB, and 94% of the differences between threshold measured by EarTrumpet and audiometer were within 10dB. Kam implemented a self-administered audiometry method on iPhone 3GS, and evaluated its accuracy [8]. The result showed no significant difference between thresholds obtained by iPhone and audiometer.

Smartphone-based audiometry brings a lot of benefits to people. However, it also faces several important problems. One of them is about sound calibration. Unlike audiometer, the output sound of smartphones is not calibrated. There is no guarantee that pure tone’s frequency and sound intensity range meet the requirement of audiometry. There may also be inconsistency between the output sound intensity of different devices. Another problem is accuracy. The hardware of smartphones and audiometers is different. Smartphone-based audiometry may not yield accurate result even after calibration. The test environment of smartphone-based audiometry usually contains louder background noise than in a chamber, which may lead to the raise of thresholds. Although several studies on smartphone-based audiometry have been reported, the systematic evaluation on the validation of the method was rare. Some of them showed inconsistency between thresholds obtained by smartphones and audiometers but the factors governing this inconsistency were rarely analyzed.

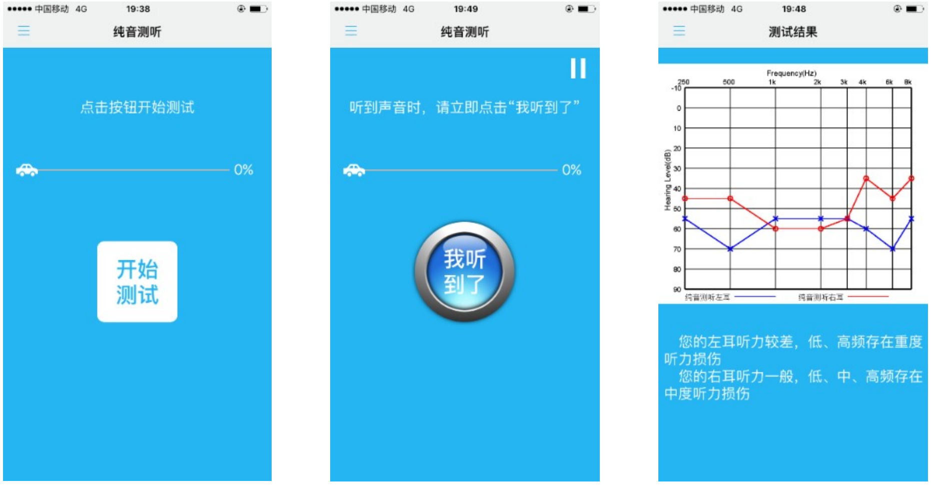
The aim of this study is to evaluate the feasibility and accuracy of an iOS-based automated audiometry. The operating system of iOS is chosen because of the consistency across different Apple devices [8]. Firstly, the audiometry was implemented and its feasibility was verified by preliminary calibration. Calibration was done on an iPhone and on an iPad Air, respectively, by a sound level meter. After the preliminary calibration, a further calibration was done to confirm that the devices were calibrated. Secondly, a human behavior experiment was conducted to evaluate the accuracy of the method. Two main factors were manipulated in the experiment: test procedures and test environment. Both hearing thresholds and test durations were recorded for each test condition and for each subject. The test-retest reliability was also verified by repeating the whole test on another day.

2 Method

2.1 Implementation of iOS-based PTA

Two versions of the ascending method described in ISO 8253-1 were implemented based on iOS: a standard version (STD) and a quick test version (QT). The standard test has been introduced early as the original version of the ascending method, and the quick version corresponds to the shortened version.

The test interface based on iOS was shown in Figure 1. At first, subject pressed the start button to start the test. Then the App began to play pure tone stimuli. The listener was instructed to press the ‘I hear’ button as soon as (s)he heard the stimuli. A progressing bar showed the percent of the test frequencies been finished. When the whole test was finished, the result was displayed in an audiogram, and a suggestion of the hearing situation was also showed with text.



点击按钮开始测试

听到声音时，请立即点击“我听到了”

您的左耳听力较差，低、高频存在重度听力损伤

您的右耳听力一般，低、中、高频存在中度听力损伤

Figure 1: Test interface based on iOS. The left panel indicates the start screen, the middle panel indicates the test screen. Right panel indicates the result screen.

2.2 Preliminary calibration

2.2.1 Equipment & Environment

A sound level meter (Larson Davis, Model 824) and the coupler (Larson Davis, Model AEC201-A) was used to measure the absolute sound pressure level for all equipment. Tested equipment were iPhone 6 and iPad Air 2 both based on iOS 8.3 (Apple Inc, Cupertino, CA) with two Apple earphones EarPods MD827ZM/A and MD827FE/A. The Calibration was conducting in an anechoic room, and the sound level of the background is 15 dBA SPL.

2.2.2 Procedure

Three sets of equipment were used: iPhone 6 with EarPods MD827ZM/A (set-1), iPad 2 with EarPods MD827ZM/A (set-2), and iPad 2 with EarPods MD827FE/A (set-3). Test frequencies consist of 0.25, 0.5, 1, 2, 3, 4, 6 and 8 kHz. The highest sound level of each set was calibrated firstly, and then the sound level was decreased by 10-dB step till to the background sound level.

2.2.3 Results

The output sound levels of the three sets were approximately consistent: when amplitude of digital signal was set as equal, the maximum difference across the three sets was 2.71 dB. Since the precision of pure-tone audiometry was 5 dB here, this difference was not sufficient to affect the final results, and only the calibration data of set-1 was adopted. The highest output levels for each frequency of set-1 are shown in Table 1.

Table 1: Highest output levels for each frequency of iPhone 6 with EarPods MD827ZM/A (set-1)

f(Hz)	250	500	1000	2000	3000	4000	6000	8000
Highest level (dB HL)	75.3	86.7	98.5	115.1	111.4	102.6	88.7	92.4

The highest output level was bigger than or close to 90 dB HL for all test frequencies except 250 and 500 Hz, indicating the equipment can be used for PTA due to the appropriate dynamic range of the output. For each test frequency, the amplitude of digital signal at each sound level interleaved by 10 dB was obtained after the sound level calibration, and denoted as a data point, representing the signal amplitude as a function of sound level in dB HL. The data points were fitted by linear regression to calculate the amplitude for any given sound level.

2.3 Further calibration

As there was no standard coupler for Apple earphones when the sound calibration was conducted, the output level of set-1 might be different from a standard audiometer for a given sound, which was thought as a systematic error. To calibrate this error, a pilot human behavior experiment was designed. Participants were required to do PTA both on a standard audiometer, and on equipment of set-1 and set-3. The averaged performance differences across equipment were treated as the systemic errors.

2.3.1 Participants

Five young university students with audiograms in the normal range (< 25 dB HL at test frequencies 125, 250, 500, 1000, 2000, 4000, and 8000 Hz) and with less than a 15-dB difference in thresholds between the two ears at all test frequencies) participated in this study. They were paid for their participation.

2.3.2 Equipment & Environment

The standard audiometer was of Madsen, AURICAL, and it was calibrated once a year. The experiment was conducted in a sound-proof room.

2.3.3 Procedure

Firstly, every subject took PTA test by the audiometer, the Apple equipment set-1, and set-3, respectively. The PTA test was conducted ear by ear for each subject, and hearing threshold was averaged between two ears. The system errors were evaluated by calculating the differences of hearing threshold across equipment at each test frequency. These errors were adjusted by compensating the differences on the equipment set-1, and set-3, respectively. And then the participants were required to do PTA on all equipment (set-1 and set-3) at the second time, to testify the validity of the systemic error calibration.

2.3.4 Results

The differences of the hearing threshold of each subject according to the second PTA tests are shown in Figure 2. The mean differences across subjects between set-1 and set-3 (left panel) ranged from 0 to 5 dB HL and the standard errors were relatively small, indicating PTA on set-1 is nearly equal to set-3 and the differences probably were caused by the inherent fluctuation of behavior tests. The mean difference between the audiometer and set-1 (right panel of Figure 2) ranged from 2 to 5 dB HL, but the standard errors were relatively big, indicating the PTA on the audiometer was considerably close to the set-1 but some systematic errors probably still existed.

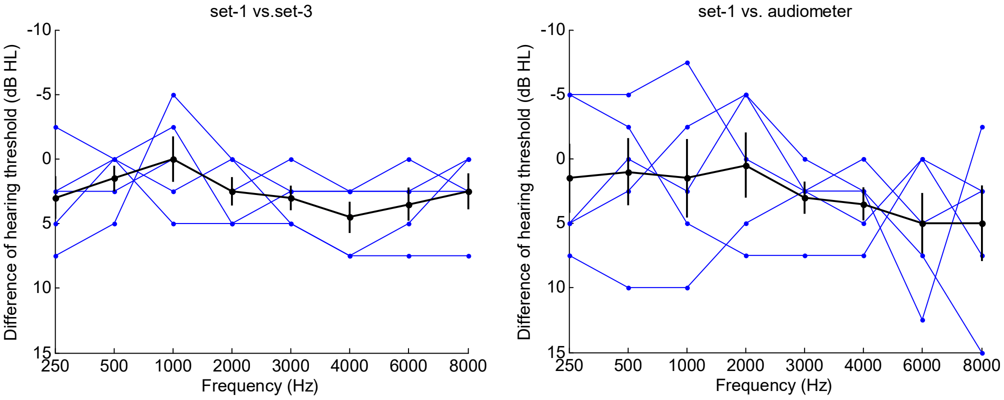


Figure 2: Left panel represents the differences of hearing threshold obtained between iPhone 6 and iPad Air 2. Right panel represents the differences between iPhone 6 and the audiometer. Black lines represent the mean values across subjects and error bars represent standard errors.

3 Experiment

3.1 Participants

Eight normal-hearing (NH) university students participated in the formal experiment. The criterion of normal hearing was the same as the previous test for calibration (see 2.3.1). Four of them took part in the previous test and the other four joined as new. All of them were paid for their participation.

3.2 Equipment & Environment

Testing equipment consisted of the standard audiometer and the set-1. Test environment consisted of a sound-proof room (SR) and a small underground library (UL) representing commonly quiet environment in daily life.

3.3 Procedure

Two factors were manipulated in the formal experiment, test procedures (STD vs. QT) and test environment (SR vs. UL), resulting in 4 (2×2) test conditions. Test order of the four conditions was balanced among subjects by a Latin-Square design. Additionally, the PTA on the audiometer in the sound-proof booth was also conducted before of the formal test (for 4 subjects), and after of the formal test (for the other 4 subjects). For each subject, eight frequencies of pure tone were tested, including 0.25, 0.5, 1, 2, 3, 4, 6 and 8 kHz, and the left ear was tested first followed by the right ear.

Totally, there were 5 (4+1) test conditions. Subjects were required to repeat the 5 tests on a different day for evaluating the re-test reliability of the automated PTA proposed in this work. The test duration in each condition was recorded automatically by the Apple app, and it was useful to evaluate the effectiveness of PTA on the equipment.

4 Results

A 2 (test procedures) × 2 (test environments) within-subject ANOVA on the hearing thresholds indicated the main effect of test procedures was not significant (F(1,7) = 0.63, $p = 0.43$), but the main effect of test environments was significant (F(1,7) = 41.62, $p < 0.001$). The interaction effect of the two factors was not significant (F(1,7) = 0.00, $p = 0.98$). The average hearing threshold of test done in SR and UL were separately 4.34 and 7.28 dB HL. The average hearing threshold of STD test and QT test were separately 5.99 and 5.62 dB HL. These results suggest the hearing thresholds measured by automated PTA on the iPhone were mainly affected by test environment rather than test procedures. Hearing threshold measured in UL was higher than that measured in SR.

A similar analysis on the test duration indicated the main effect of test procedures was significant (F(1,7) = 468.47, $p < 0.001$), but the main effect of test environments was not significant (F(1,7) = 0.10, $p = 0.76$). The interaction effect of the two factors was not significant (F(1,7) = 0.47, $p = 0.50$). The average test duration of STD and QT were 569.22 seconds and 349.25 seconds, respectively. The average test duration in SR and UL were 460.81 seconds and 457.66 seconds, respectively. These results suggest the PTA by QT procedure could save test time without significant precision loss.

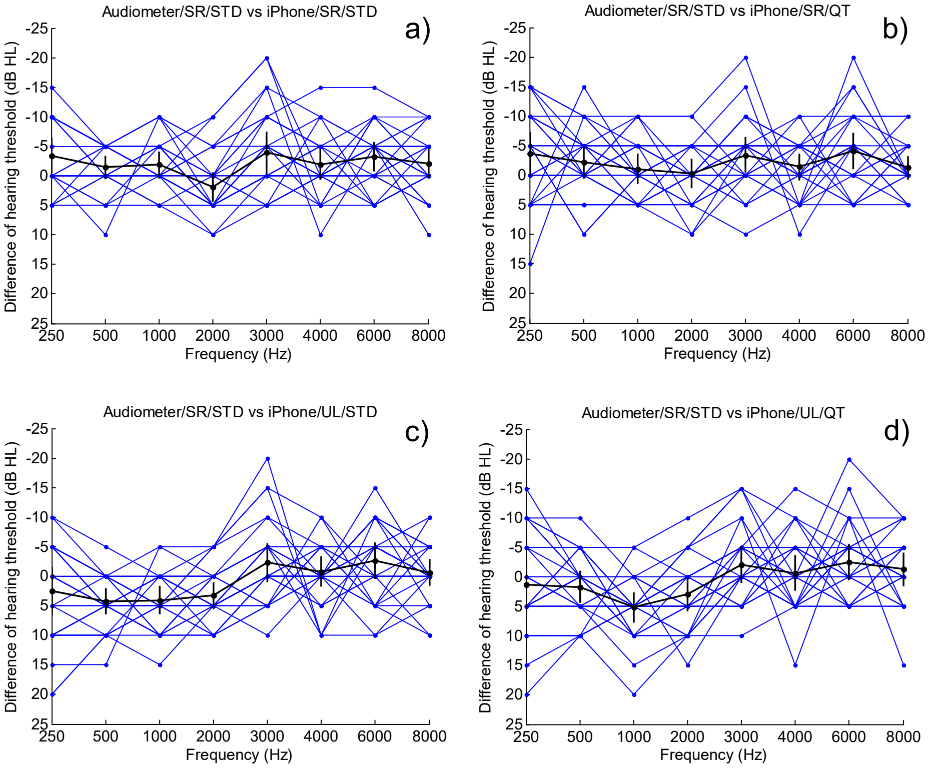


Figure 3 - Differences of hearing threshold obtained by audiometer and in each of the four test conditions: a) Audiometer vs. iPhone in SR with STD; b) Audiometer vs. iPhone in SR with QT; c) Audiometer vs. iPhone in UL with STD; d) Audiometer vs. iPhone in UL with QT. Black lines represent average across subjects and error bars represent standard errors.

The differences of hearing threshold measured by the standard audiometer and measured by the iPhone are shown in Figure 3. The mean value and standard errors of hearing threshold differences among all frequency of all four conditions are within 5 dB. The shape of four curves is similar. The curves of upper row (SR) and lower row (UL) show a deviation, indicating that test environment may have an effect on test result. Paired t-test on hearing thresholds measured by audiometer and measured by iPhone show there is significant difference between the audiometer and iPhone in all conditions (p < 0.001) except for the condition of iPhone in UL with QT (p = 0.15). The reason why there isn't a significant difference between these two conditions may be the counteract of three factors (on average, hearing threshold measured by Audiometer is 1.99 dB higher than that measured by iPhone, hearing threshold measured in SR is 2.94 dB lower than that measured in UL, hearing threshold measured with STD is 0.37 dB higher than that measured with QT). Pearson correlations analysis on hearing thresholds show the r > 0.60 for all four conditions and p < 0.001, indicating that test results of iPhone is significantly correlated to that of audiometer.

The reliability of test-retest was evaluated by Paired t-test and Person correlations on the hearing thresholds between the first test and the repeated test for each of the five test conditions, respectively. The results are shown in Table 2.

Table 2: Reliability of retest for the five testing conditions

Test conditions	**Paired t-test**	**Pearson correlations**	
	p	*r*	*p*
Audiometer/SR/SDT	0.07	0.84	< 0.001
iPhone/SR/SDT	0.19	0.83	< 0.001
iPhone/SR/QT	0.19	0.76	< 0.001
iPhone/UL/SDT	0.67	0.84	< 0.001
iPhone/UL/QT	0.59	0.70	< 0.001

As shown in Table 2, there was no sigificant difference between the test and re-test for all conditions (p > 0.05). For Pearson correlations analysis, there is significant correlations between the test and re-test for all conditions ($r \geq 0.70$, $p < 0.001$). These results suggest the iOS-based automated audiometry proposed in this work could be reliablely repeated for the subjects. Please notice that as all subjects in this experiment are normal hearing, it remains unclear whether the reliability shown in this study could be consistent for listeners with hearing impaired.

5 Discussion and Conclusion

A PTA on Apple iOS was implemented, and the feasibility and accuracy were evaluated in this work. According to the preliminary calibration, it was found the highest output level of iPhone 6 was bigger than 90 dB HL for all test frequencies except 250 and 500 Hz. Since most hearing loss occurs at high frequencies, the frequency and sound level range generally met the requirement of PTA. The differences between output sound level of iPhone 6 and iPad Air 2 were within 2.71 dB, which showed a consistency between these two Apple devices, and it was similar as the previous conclusion by Allen [7], although the devices were iPhone 3G, iPhone 4 and various iPod Touch and iPad devices. The result of preliminary calibration shows that Apple devices are suitable for PTA.

The further calibration was conducted to compensate the possible error caused by the coupler for sound level calibration, as there was no standard coupler for Apple earphones. With this manipulation, the mean differences of hearing threshold measured by Apple devices and audiometer were relatively small (≤ 5 dB for the mean values), but the standard errors were relatively big. Additionally, the hearing thresholds measured by iPhone were significantly lower 3-5 dB than that measured by audiometer in the formal experiment. These results suggest the sound calibration play an important role for PTA audiometry, and it may be not easy to make the output of iPhones exactly the same as the standard audiometry, but the error could be controlled in a reasonable range.

The results of the experiment showed that tested hearing thresholds were mainly affected by test environments. The noise in the normally quiet room lifted hearing threshold by approximately 3 dB, which was smaller than the precision of PTA (5 dB in this work). The test procedure mainly affects test duration rather than the hearing threshold. The test duration of QT was only 60% of the duration of STD. These results indicate that QT can shorten the test duration without losing accuracy, and QT can be a replacement of STD on iPhones. All five test conditions have high test-retest reliability, which shows that differences observed for test equipment, test environments and test procedures don't affect test-retest reliability.

In summary, the iOS-based PTA implemented in this study could work well for normal hearing listeners, but it remains unclear whether conclusions above could be available for people with hearing loss. In future work, subjects with hearing loss should be included to further test the accuracy of iOS-based PTA. We will also work on the measurement of noise on Apple devices, and compensate for the accuracy lose caused by test environment.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No. 61473008 and No. 11590773); a 985 grant by Peking University, and a Newton alumni funding by the Royal Society, UK.

References

[1] China Disabled Persons' Federation. Communique on major statistics of the Second China National Sample Survey on Disability: Leading group of the Second China National Sample Survey on Disability & National Bureau of Statistics of the People's Republic of China, China, 2008.

[2] International Organization for Standardization, ISO 8253-1: Acoustics-Audiometric test methods-Part 1: Basic pure tone air and bone conduction threshold audiometry, Geneva, 1989.

[3] Arlinger S D. Comparison of ascending and bracketing methods in pure tone audiometry. A multi-laboratory study, *Scandinavian Audiology*, Vol 8 (4), 1979, pp 247-251.

[4] Wood T J.; Wittich W W.; Mahaffey R B. Computerized pure-tone audiometric procedures, Journal of Speech, Language, and Hearing Research, Vol 16 (4), 1973, pp 676-684.

[5] Szudek J.; Ostevik A.; Dziegielewski P.; et al. Can Uhear me now? Validation of an iPod-based hearing loss screening test, *Journal of Otolaryngology--Head & Neck Surgery*, 2012, p 41.

[6] Van Tasell D J.; Folkeard P. Reliability and accuracy of a method of adjustment for self-measurement of auditory thresholds, *Otology & Neurotology*, Vol 34 (1), 2013, pp 9-15.

[7] Allen F.; Peggy B.; Hamid D. Automated audiometry using Apple iOS-based application technology, *Otolaryngology--Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, Vol 149 (5), 2013, pp 700-706.

[8] Kam A C S.; Sung J K K.; Lee T.; et al. Clinical evaluation of a computerized self-administered hearing test, *International Journal of Audiology*, Vol 51 (8), 2012, pp 606-610.