Test person operated 2-Alternative Forced Choice Audiometry compared to traditional audiometry

Jesper Hvass Schmidt[1,2], Christian Brandt[3], Jakob Christensen-Dalsgaard[3], Ture Andersen[1], Jesper Bælum[1], and Torben Poulsen[4]

1 Dept. of Audiology Odense University Hospital, University of Southern Denmark, DK-5000 Odense C, Denmark

2 Dept. of Occupational Health and Environmental Medicine, Odense University Hospital, University of Southern Denmark, DK-5000 Odense C, Denmark

3 Institute of Biology, Centre for Sound Communication, University of Southern Denmark, DK-5230 Odense M, Denmark

4 Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Lyngby, Denmark

A new constructed user operated audiometry system was evaluated and compared to traditional audiometry. User operated audiometry was based on the 2 Alternative Forced Choice (2AFC) paradigm in combination with the method of maximum-likelihood (MML), which was used dertemine user-operated hearing thresholds after fitting of the most probable psychometric function. Combination of the 2AFC paradigm and the MML gave reliable hearing thresholds. User operated audiometry was validated by comparison to traditional audiometry. 30 persons (60 ears) performed traditional audiometry as well as user-operated 2AFC-audiometry. Test subjects were normal and moderately hearing impaired. User operated audiometry was reliable compared to traditional audiometry. User-operated audiometry thresholds were 1-2 dB lower compared to traditional audiometry. Standard deviations between the two test methods were below 4.5 dB for frequencies from (250-4000 Hz) and up to 6.7 dB for frequencies above 4000 Hz. Test-retest studies of user-operated audiometry were comparable to traditional audiometry. User-operated 2AFC audiometry can be a reliable alternative to traditional audiometry especially under certain circumstances, where it can be difficult to get skilled technical assistance to conduct the audiometry.

INTRODUCTION

Measurement of correct and reliable hearing thresholds is dependent on correct measurement techniques and patient compliance. In the most optimal clinical settings standard deviations will be 3-4 dB for the frequencies up to 4000 Hz, and even larger for higher frequencies. To keep test-retest variability within this scale it will require otological normal and well motivated subjects. The test-retest standard deviations in industrial audiometry range from 6-10 dB (Dobie, 1983). Patients have different response criterions on thresholds to what they regard, as the faintest sounds they can hear. These response criterions can depend on the method used for obtaining hearing

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

thresholds, and the instructions given to the patient (Marshall and Jesteadt, 1986; Marshall, 1991; Marvit *et al.*, 2003).

Different paradigms in psychoacoustic research have been used to obtain hearing thresholds. Two of the known paradigms have been described as the Yes-No paradigm and 2 Alternative Forced Choice (2AFC). Different methods have been used in combination with these paradigms in order to obtain hearing thresholds. The method of maximum-likelihood is an adaptive method, based on the most probable psychometric function taken from a set of possible candidate psychometric functions. In general the method of maximum-likelihood is combined with the Yes-No paradigm. The method of maximum-likelihood can be very efficient and a reliable threshold can often be obtained with only 15 trials (Green, 1993; Gu and Green, 1994; Leek *et al.*, 2000; Marvit *et al.*, 2003). A disadvantage of this procedure for clinical use is that it requires a stable response criterion from the patient, which is not always the case. The 2AFC paradigm is more insensitive to a change in the patient's response criterion. 2AFC paradigms are faster than 3AFC paradigms but still these paradigms can consume a lot of time (Marshall and Jesteadt, 1986; Marshall *et al.*, 1996

Another method with widespread use in the literature is the transformed up-down method. One of the big advantages of this method is a high reliability. The up-down method is often combined with the 2AFC paradigm, but this requires a large number of trials to measure accurate thresholds (Levitt, 1971, Leek *et al.*, 1992; Marvit *et al.*, 2003). Systematic bias is likely to occur as the number of trials is reduced in a 2AFC paradigm where the probability of success is near 1/2 just by chance for every trial. Furthermore subjects also have a certain risk of making false alarms, which can influence threshold estimates (O'Regan and Humbert, 1989, Marvit *et al.*, 2003).

The 2 up – 1 down adaptive method used with the 3AFC paradigm will target the 70,7% point of correct responses. Thresholds are in general lower with the 3AFC paradigm and up-down methods compared to traditionally obtained thresholds (Gatehouse and Davis, 1992). This study only examined the thresholds obtained for the 2000 Hz frequency.

The goal of the present study is to obtain hearing thresholds comparable to traditional manual audiometry thresholds with a limited influence of patient related response bias and false alarms. This is important, as the patient themselves conducts the audiometry without any assistance, and the potential influence on hearing thresholds from systematic bias and false alarms with the 2AFC paradigm must be handled.

METHODS

Subjects and groups

13 males and 17 females in total participated in the study. 16 subjects (9 females and 7 males) were recruited from ordinary patient examination in the department of Audiology, Odense University Hospital. The reasons for referral to audiologic examination were various causes (tinnitus, hearing loss, ototoxicity control). The

remaining subjects were without any known hearing loss. The subjects are naive listeners and the majority had no records of previous hearing tests. Subjects were aged 20-69 years and divided into 3 age groups: 9 below 30 years, 12 from 30-50 years and 9 above 50 years. Subjects were otologically examined to disclose obstructing cerumen prior to hearing testing. Only patients without substantial asymmetric hearing loss of any frequency >30 dB and with known hearing thresholds <70 dB of any frequency (0.25 – 8 kHz) were included in the study.

All subjects took 2 audiological tests on the same day separated by a short break of typically 10 minutes. The two tests are a traditional audiometry and the new constructed 2AFC audiometry. Half of the subjects started out with the traditional audiometry, and the other half took the 2AFC audiometry as the primary test.

13 other subjects (9 females and 4 males) were included in a test-retest study with the 2 AFC-audiometry alone. The two tests were separated in time from an hour up to several days.

Measurement procedures

Traditional audiometry was conducted within a soundtreated booth according to standards described in ISO 8253-1(International Organization for Standardization, 1989). The equipment used was MADSEN Midimate 622 audiometers with TDH-39 Telephonics headphones. The audiometric test used a modified Hughson-Westlake technique. The frequencies (250, 500, 1000, 2000, 3000, 4000, 6000, 8000 Hz) were tested as air conductions. The right ear was tested before the left ear.

Automatic 2AFC audiometry is conducted with a computer (Compaq nx6310) coupled to a transportable mobile device – Mobile Processor RM2, Tucker Davis Technologies (TDT) through an USB-connection. A Senneheiser HDA200 headphone is connected to the mobile device. In order to measure low hearing thresholds, and possibly below 0 dB HL, an extra attenuator of 400 ohm is connected between the mobile processor and the headphones. All 2AFC audiometry tests have been conducted outside a soundtreated booth in a quiet room, with the attenuation from the HDA200 headphones as the only primary sound attenuation.

The RM2-TDT mobile device is controlled by computer software. The test tones are presented randomly in 1 of 2 intervals as a 2AFC paradigm without feedback. The presence of a test tone is indicated by a coloured box on the computer screen. Interval 1 is marked by a red box and interval 2 is marked by a blue box. The subject's task is to choose the interval containing the signal. The test tones are played as 3 tones each length of 200 ms with rise/fall times of 15 ms separated by intervals of 300 ms. The frequencies (250, 500, 1000, 2000, 3000, 4000, 6000, 8000 Hz) are tested as air conductions. Left ear is tested first.

The measurement strategy can be described as a set of three rules that govern the process of a psychometric procedure (Marvit *et al.*, 2003).

Starting rule

For the majority of subjects a starting level corresponding to 40 dB HL was chosen, but in cases of known hearing loss this starting point was chosen to 60 dB HL or 70 dB HL.

Progression rule

Correct answers lowers the test tone with 10 dB (see Fig. 1). The hearing threshold is crossed after a few trials, and the subjects will no longer hear the signal. The subjects must guess on the most probable interval containing the signal. Test tones are allocated randomly to one of two intervals and false alarms can occur by chance. The first error results in an intensity increase of 30 dB (see Fig. 1). This rule serves to correct false alarms and increase the familiarization with the procedure. Correct responses again results in 10 dB decrease in intensity. The steps are now controlled by the maximum-likelihood psychometric function, as described by Gu and Green (1994) based on the previous answers. The 95% point of correct responses of the most probable psychometric function based upon the previous responses is calculated. Intensive testing occurs 5 dB above the expected threshold point corresponding to 95% correct responses on the psychometric function. This level is called the upper limit. In between testing at this upper limit, test trials 5 dB below the expected threshold is presented. This level is called the lower limit. The psychometric function of pure tone thresholds has been shown to span approximately 8 dB (Watson *et al.*, 1972). Thus if the expected psychometric function is placed at a correct level, the upper limit will be close to 100% and the lower limit will be close to 50% correct responses in an 2AFC paradigm. The lowering of the intensity progress until a level of 5 dB above the calculated threshold is reached. After the second fault the intensity is increased by 5 dB (see Fig. 1). Intensity raise after the second fault only will minimise the effects of a false alarms from the patients in cases of false negative faults. The patients can correct obviously wrong answers, but they do not notice in all cases.

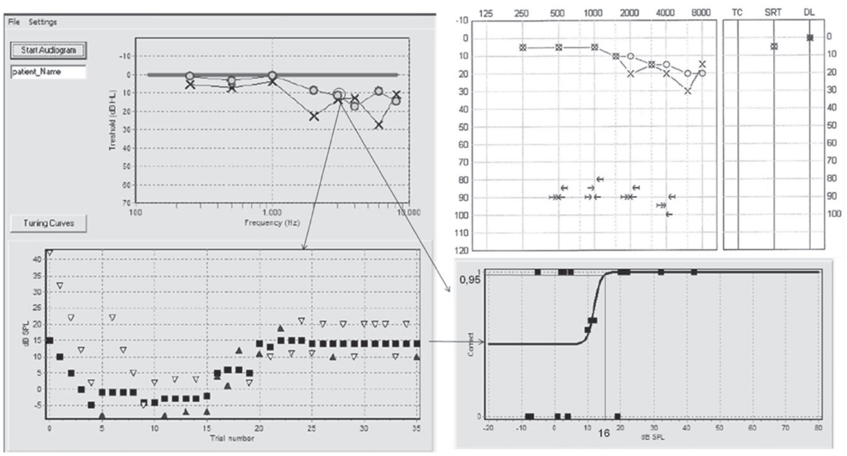

Fig. 1: Principle of the 2AFC test procedure used and a comparison of audiograms. 2AFC audiogram is shown to the left and a traditional audiogram to the right. White triangles in the left lower panel indicates correct responses, black triangles indicate wrong responses. Black squares are the calculated thresholds based on the most likely psychometric function (bottom-right) based on previous answers.

Stopping rule

Testing proceeds until at least 6 consecutive correct responses have been made at the upper limit level. If at least 6 consecutive correct responses can be made it is assumed that these values on the upper limit are very close to 1 on the psychometric function. 6 consecutive correct responses significantly limit the risk of false positive responses due to a correct guess by chance. To stop the test it will also require a minimum of 2 incorrect responses at the lower limit (see Fig. 1).

If all answers at the lower limit are correct, the testing procedure will search for lower thresholds as the most probable psychometric function shifts to lower levels. The procedure continuously searches for 2 faults at the lower limit. If the subject makes too many false alarms at earlier stages of the test, this will affect the placement of the most probable psychometric function. If no faults at the lower limit occur, the procedure will return to earlier stages in the test and search again for a probable threshold. The procedure runs until a least 30 trials have been completed for each frequency. If the stopping rule can be fulfilled, the test proceeds to the next frequency. If the stopping rule can not be fulfilled, the test proceeds until the stopping rule can be fulfilled. The number of trials required to determine the hearing threshold varies.

Datum definition

The datum definition is set arbitrarily to 95% correct responses. The threshold was estimated at the signal level corresponding to the 95% point on the most likely psychometric function after at least 30 trials.

Calibration

Calibration of headphones is done according to the standards described in ISO 389-8 by using a coupler with artificial ear type 4153 from Brüel and Kjær as specified in IEC 60318-3 and specified for the Sennheiser HDA-200 headphone (International Organization for Standardization, 2004).

Statistical analyses

Test procedures are compared by using the method of limits of agreement as described by Bland and Altman (1986).

RESULTS

The two test-methods are compared in the Bland Altman plot (Fig. 2). This figure shows the 8 tested frequencies in 8 different diagrams.

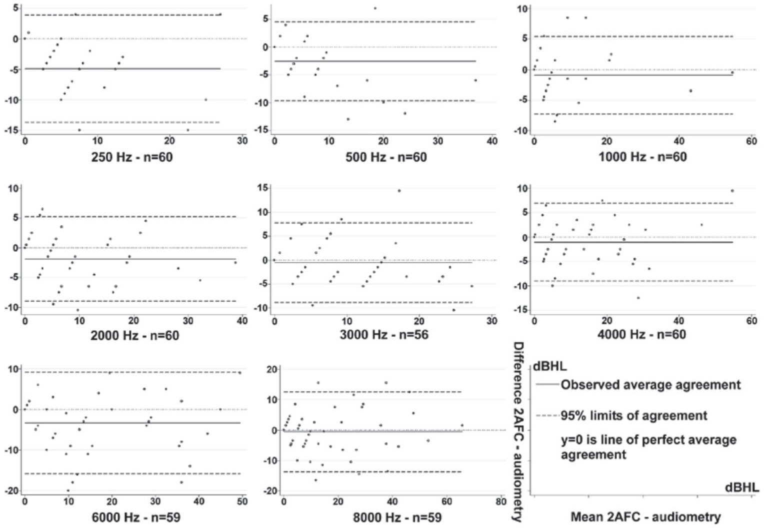

Fig.2: Bland and Altman plot. Dashed line: 95% confidence interval. Solid line: Observed mean difference. Dotted line: If no difference was observed.

The thresholds obtained with 2AFC audiometry are slightly lower, or they closely equal thresholds obtained with traditional audiometry for all tested frequencies. The observed standard deviations and corresponding confidence intervals from Fig. 2 is listed in Table 1.

One potential outlier corresponding to one ear at 3000 Hz and one at 8000 Hz has been excluded from dataanalysis (data not shown). The observed standard deviations from 250 Hz-4000 Hz are from 3.2-4.5 dB. The standard deviations for 6 and 8 kHz are 6.4-6.7 dB. At least for the outlier representing 3000 Hz this measurement was repeated again for the 2AFC procedure and the 2 measurements deviated only 4 dB.

It is assumed that an error in the traditional performed audiometry was made at this specific point (data not shown). The corresponding confidence intervals tells us that 2 measurements obtained with traditional audiometry and 2AFC audiometry with a 95% probability deviates less than +/- 8.8 dB for the frequencies 250-4000 Hz and +/- 13.1 dB for 6-8 kHz.

The 2 AFC audiometry test did allow patients to test below 0 dB HL to the absolute minimum they possibly could hear. This will probably add a larger uncertainty, especially when the measurement is below 0 dB HL. Standard deviations from the repeated measurements were calculated, as well as the average standard deviations for the different frequencies. The results are shown in table 2.

Frequency [Hz]	Difference [dB]	Std. Dev. Avg. dB]	95% limits of agreement [dB] (Bland and Altman, 1986)	
250	-4.9	4.5	-13.7	3.9
500	-2.6	3.6	-9.7	4.6
1000	-0.9	3.2	-7.3	5.4
2000	-1.9	3.6	-9.0	5.2
3000	-0.5	4.3	-8.9	7.8
4000	-1.1	4.1	-9.0	6.9
6000	-3.4	6.4	-15.8	9.1
8000	-0.6	6.7	-13.7	12.6

Table 1: Summary statistics of Fig. 2.

Frequency [Hz]	No. of ears	Mean Std. Dev. [dB HL]
250	26	2.4
500	26	3.0
1000	26	2.3
2000	26	2.5
3000	10	1.4
4000	26	1.9
6000	10	2.5
8000	26	3.0

Table 2: Mean standard deviations observed from test-retest with the automatic 2 AFC procedure.

DISCUSSION

The validity of the automatic 2AFC audiometry has been tested by comparing the test to the known standard procedure of traditional audiometry. The two test methods differ in many important aspects, but the objective of both methods is to estimate reproducible hearing thresholds as low as possible. Previous studies comparing thresholds obtained from a procedure based on 2 Interval Forced Choice (2IFC) with traditional obtained thresholds find mean differences of 6.5 dB, with the lowest thresholds observed with the 2IFC method. In this study the 71% correct point was estimated on the psychometric function (Marshall and Jesteadt, 1986). In another study 2AFC thresholds were found to be 2.9 dB lower, when the 79% correct response point was bracketed (Marshall *et al.*, 1996). Others have also found increased differences between a three-interval-forced choice procedure and clinical thresholds related to hearing loss and age (Gatehouse and Davis, 1992). These different datum definitions contribute to the differences seen between different psychometric methods. In the present study a datum definition of 95% of correct responses is used. The objective of the present study was to create a reliable self testing audiometry system, which could minimise subject related response bias, expected systematic bias and effects from false alarm. This method differs from previous methods with the maximum likelihood technique. The present method is also adaptive, but the method defines a possible upper limit and a lower limit corresponding to the most probable psychometric function. These limits function as control elements which will limit influence from systematic bias and false alarms. The 2AFC paradigm itself reduces subject related response bias. The validation shows that thresholds obtained with 2AFC audiometry are not poorer than thresholds obtained with traditional audiometry. This is true for all measured thresholds, as no mean differences were larger than 0 dB. Furthermore the standard deviations of the differences between the traditional audiometry and the 2AFC Audiometry is below 4.5 dB for 250-4000 Hz, as it is evident from Table 1. This corresponds well with the known uncertainties of the traditional audiometry. The known uncertainty for the frequencies 250-4000 Hz has been given in terms of standard deviation to 4.9 dB in a draft version of uncertainties related to the revision of ISO8253-1 (International Organization for Standardization, 1989). For frequencies above 4000 Hz the uncertainty is even larger. The two methods do not differ significantly in terms of repeatability, and it should be possible to use both methods interchangeably. However, it should be taken into account that 2AFC audiometry thresholds are slightly lower than traditional clinical obtained thresholds. The 2AFC paradigm can reduce response bias from the subject and this is suggested to play a role in the lower thresholds observed. Subjects tend to respond differently and some patients want to be surer, when they answer “yes” the tone was heard (Marshall and Jesteadt, 1986; Marshall, 1991). In a 2AFC paradigm it is “acceptable” to guess and patients can pick the right interval, even in situations where they are in doubt. This will introduce a risk of systematic bias in the 2AFC paradigm, and it is therefore important to minimise this risk as much as possible in the procedure. Furthermore subject related false alarms can also lead to wrong threshold estimates. Response criterions and false alarms may vary across the procedure and this is one of the main explanations why simulations often fail to reproduce results from empirical experiments (Marvit *et al.*, 2003).

A larger difference between the two methods is seen at low frequencies (250 Hz and 500 Hz). One main reason for the differences seen at 250 Hz is of technical character. Two different headphones were used, and the one used for traditional audiometry (TDH-39) is known to be more dependent than the circumaural headphones like HDA-200 on correct placement, especially when measuring low frequencies (Shaw, 1966; Riedner, 1980).

The clinical potential of an automatic 2AFC audiometry system will be to lower biases related to the used methods under certain circumstances. This computer and subject operated procedure tends to lower biases related to the operator as well as the subject. These biases are known contributors of uncertainties in traditional audiometry (International Organization for Standardization, 1989). Thus the method will be valid to use under circumstances where it is difficult to keep up to required standards. This could be as a part of an occupational hearing conservation programme. The audiometries from these programs are often with much larger uncertainty, than the uncertainty observed in the present study (Dobie, 1983). Furthermore if there is a lack of qualified operating personal, this method of audiometry can be used to give reliable audiometries comparable to clinical obtained audiometries. It should be noted, that the subjects selected for this study can not be considered as otologically normal as hearing impaired subjects are included. The group would be more comparable to i.e. a group of industrial workers, where hearing loss is likely to occur.

CONCLUSION

Automatic 2AFC audiometry is a valid alternative to traditional audiometry. When using 2AFC audiometry, it is important to notice that 2AFC audiometry gives a lower threshold of typically 1-2 dB for most of the frequencies. In general this little difference will only have minor clinical consequences, when thresholds obtained with two different methods are compared. Furthermore the reliability of Automatic 2AFC audiometry is comparable to traditional audiometry.

REFERENCES

Bland, J. M., and Altman, D. G. (**1986**). "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet **1**, 307-310.

Dobie, R. A. (**1983**). "Reliability and validity of industrial audiometry: implications for hearing conservation program design," Laryngoscope **93**, 906-927.

Gatehouse, S., and Davis, A. (**1992**). "Clinical pure-tone versus three-interval forced-choice thresholds: effects of hearing level and age," Audiology **31**, 31-44.

Green, D. M. (**1993**). "A maximum-likelihood method for estimating thresholds in a yes-no task," J. Acoust. Soc. Am. **93**, 2096-2105.

Gu, X., Green, D. M. (**1994**). "Further studies of a maximum-likelihood yes-no procedure," J. Acoust. Soc. Am. **96**, 93-101.

International Organization for Standardization (**1989**). “Acoustics - Audiometric test methods - Part 1: Basic pure tone air and bone conduction threshold audiometry,” *ISO8253-1* (Geneva: ISO). Including a draft version of the revision of this standard.

International Organization for Standardization (**2004**). “Acoustics - Reference zero for the calibration of audiometric equipment - Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones,” *ISO 389-8* (Geneva: ISO).

Leek, M. R., Dubno, J. R., He, N., and Ahlstrom, J. B. (**2000**). “Experience with a yes-no single-interval maximum-likelihood procedure,” J. Acoust. Soc. Am. **107**, 2674-2684.

Leek, M. R., Hanna, T. E., and Marshall, L. (**1992**). “Estimation of psychometric functions from adaptive tracking procedures,” Percept Psychophys **51**, 247-256.

Levitt, H. (**1971**). “Transformed up-down methods in psychoacoustics,” J. Acoust. Soc. Am. **49**, Suppl.

Marshall, L. (**1991**). “Decision criteria for pure-tone detection used by two age groups of normal-hearing and hearing-impaired listeners,” J. Gerontol. **46**, 67-70.

Marshall, L., Hanna, T. E., and Wilson, R. H. (**1996**). “Effect of step size on clinical and adaptive 2IFC procedures in quiet and in a noise background,” J Speech Hear Res **39**, 687-696.

Marshall, L., and Jesteadt, W. (**1986**). “Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures,” J Speech Hear Res **29**, 82-91.

Marvit, P., Florentine, M., and Buus, S. (**2003**). “A comparison of psychophysical procedures for level-discrimination thresholds,” J. Acoust. Soc. Am. **113**, 3348-3361.

O’Regan, J. K., and Humbert, R. (**1989**). “Estimating psychometric functions in forced-choice situations: significant biases found in threshold and slope estimations when small samples are used,” Percept Psychophys **46**, 434-442.

Riedner, E. D. (**1980**). “Collapsing ears and the use of circumaural ear cushions at 3000 Hz,” Ear Hear **1**, 117-118.

Shaw, E. A. (**1966**). “Earcanal pressure generated by circumaural and supraaural earphones 2,” J. Acoust. Soc. Am. **39**, 471-479.

Watson, C. S., Franks, J. R., and Hood, D. C. (**1972**). “Detection of Tones in Absence of External Masking Noise .1. Effects of Signal Intensity and Signal Frequency,” J. Acoust. Soc. Am. **52**, 633-643.