Perceptual audio evaluation by hearing-impaired listeners – some considerations on task training

Renskje K. Hietkamp, Martin R. Andersen, and Thomas Lunner

Oticon A/S, Research Centre Eriksholm, Kongevejen 243, DK-3070 Snekkersten, Denmark

Use of perceptual audio evaluation is widespread in the audio and telecommunication areas and is also relevant to hearing-aid research, as it addresses artifacts of signal processing in hearing aids. However, in hearing-aid research, listeners are typically hearing impaired and this poses a challenge for the training material; the impaired hearing system differs markedly from the normal hearing system and shows large, individual, and highly unpredictable variation. This was demonstrated in a pilot study in the form of a listening experiment with music and speech stimuli, processed to generate different degrees of non-linear artifacts. Six subjects with mild-to-moderate, sloping hearing losses were tested after some initial training. Results were contradictory, but seemed to indicate that the subjects were not able to detect differences in sound quality. A stepwise training procedure was therefore developed for perceptual audio evaluation targeting hearing impaired listeners, which was inspired by Bech and Zacharov [*Perceptual Audio Evaluation - Theory, Method and Application* (2006)]. Key issues in the training procedure are priming to the artifacts under study and a test-flow that facilitates errorless learning. Using this training procedure in an experiment with the same type of stimuli as before, the results showed that differences in sound quality were detected in the 2-5 kHz region. These results hint at a need for careful designed training when hearing impaired listeners are to be used for perceptual audio evaluation.

INTRODUCTION

It can be assumed that future hearing aid algorithms or assistive listening systems may include signal data compression such as MPEG. This raises the question of whether artifacts will be audible to the hearing aid user. The impact of artifacts on perception cannot be easily predicted by objective measurements, as the auditory system is highly non-linear. Therefore listening experiments are needed to obtain perceptual audio evaluations. These are common in audio and tele-communication fields (Bech and Zacharov, 2006), using normal hearing subjects. The subjects are trained prior to testing, typically by use of a listening panel. After listening to different types of distortions, results are discussed within the panel and a common terminology is agreed upon (Hansen, 1987; ITU-R, 1997). However, the impaired auditory system differs markedly from the normal hearing system and shows large, individual and highly unpredictable variation. Therefore, a listening panel cannot be used and targeted task training is relevant.

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.

The need for targeted task training was underlined by the results of a pilot study we conducted. During four visits, six hearing-impaired subjects were asked to discriminate between processed and unprocessed sounds. This was done by an ABX-paradigm, which allows double blind comparison testing (Clark, 1982). In such a test, the subject is presented two known samples A and B (one being the reference, the other processed, randomly assigned) together with a third (X) that is the same as either A or B. The subject needs to identify which sample (A or B) corresponds to X (see Fig. 1 for the Graphical User Interface). Subsequent statistical analysis is performed to show whether subjects are able to discriminate between two sounds above chance level.

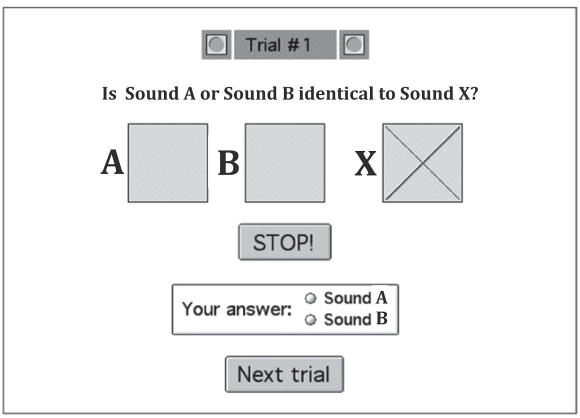

Fig. 1: Graphical User Interface of the ABX-paradigm.

The results of our pilot study, summarized in Fig. 2, seemed to indicate that the discrimination between processed and unprocessed stimuli was mostly not significantly different from chance performance ($p<0.05$, binomial test) and, unexpectedly, even the largest artifacts did not seem to be audible, whereas moderate artifacts showed best performance. We questioned the validity of these results as they seemed to be contradictory. Our concerns regarded possible insufficient training: the training consisted of 24 trials of stimuli with severe artifacts. Total training time was approximately one hour. No feedback was given to whether responses were correct or incorrect. As the results seemed contradictory and the test subjects reported difficulties executing the task, we speculated on two matters:

Had the subjects understood the ABX-task correctly?

Was the outcome due to lack of familiarisation to the processed stimuli?

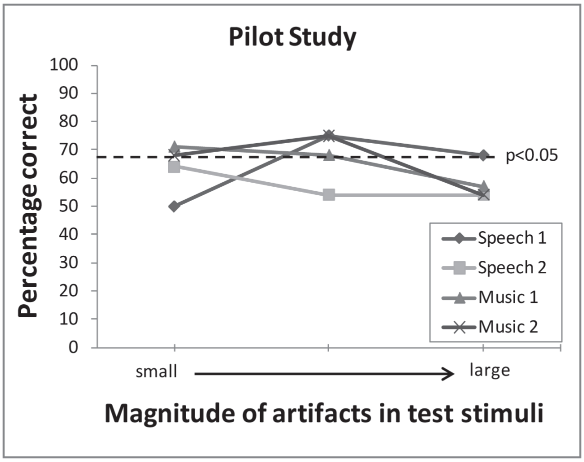

Fig. 2: Before guided stepwise training procedure: Results of the pilot study with six hearing-impaired subjects. Performance is mostly not significantly different from chance performance. Performance for stimuli with the largest artifacts is as poor as for stimuli with the smallest artifacts.

GUIDED STEPWISE TRAINING PROCEDURE

We developed a guided stepwise training program to address the abovementioned two speculations on the ABX-task and the familiarization. We had a number of considerations in these two areas. An overview is given in Fig. 3.

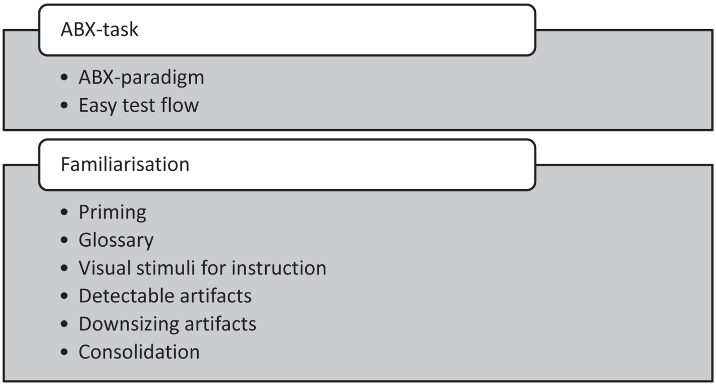

Fig. 3: Overview of considerations on perceptual audio evaluation by hearing-impaired listeners using an ABX-paradigm.

There is a high demand of working memory (Baddeley, 1986) in the ABX-paradigm, as the subject needs to memorize one sound while listening to and comparing to another sound. Meanwhile, the given task of identifying the stimulus that is the *same* as the X-reference needs to be kept in mind. Involuntarily, subjects at times perform the opposite task, i.e. they identify the stimulus that *differs* from the other two stimuli. The fact that the task is presented on a computer screen complicates matters for those subjects unaccustomed to computers. Some of these problems can be solved by using an easy test flow that facilitates learning. An example of such an approach is errorless learning, whereby the task is manipulated to eliminate or reduce errors (Fillingham *et al*., 2003). Baddeley and Wilson (1994) showed in an experiment with anamnestic subjects, young controls and elderly controls, that errorless learning was superior to errorful learning by trial and error for all groups with regard to learning time; subjects needed less trials for maximum performance. With regard to the task we presented to our subjects, an errorless learning approach would make the training phase more effective. An errorless learning procedure can be implemented in the ABX-task by presenting the task in a familiar way, such as pictures laid out on a table and by adding an initial AB-phase, where the task is to describe the difference between two stimuli.

The second speculation concerned familiarization to the artifacts. Several issues play a role: priming, compilation of a glossary, the experimenters role, the magnitude of the artifacts and consolidation. Priming is a key issue for familiarization; it is an experimental technique by which a stimulus is used to sensitize the subject to a later presentation of the same or similar stimulus. A powerful tool in priming is the compilation of a glossary, reflecting the perceived auditory difference between a processed and an unprocessed sound. As mentioned above, this cannot be done in a listening panel when hearing-impaired subjects are to be used; the glossary needs to build on the individual perception and should not be imprinted with perceptions from another auditory system, such as the experimenter's. This can be avoided by using visual stimuli in the instruction of techniques for building up and expanding a glossary. When the task is understood, the link to sound can be made and sound stimuli with severe artifacts can be presented. The subject can then start with the compilation of a glossary, preferably both orally and in writing to enhance memorization. The experimenter's role in this phase is especially important, since only the subject's auditory perception should be allowed. This can be achieved by using techniques such as repetition, rephrasing, paraphrasing or positive affirmative comments, known from other fields (Miller and Rollnick, 2002). The determination of the magnitude and steps of the artifacts to be included in the training phase might need some pilot studying if errorless learning is desired; artifacts should be largest and easy detectable at the onset of the training phase and be reduced subsequently. This downsizing can be challenging, since it should only affect the magnitude but not the perceptual character of the artifacts. Finally, effectiveness of the training can be enhanced by including a consolidation phase, where subjects can test their newly gained skills; an ABX-paradigm with response feedback (right/wrong) is suitable for this purpose.

EXPERIMENT

Test procedure

With the abovementioned considerations in mind, a new procedure was developed (see Table 1), and the same six subjects were tested again. The procedure was divided into four phases: a priming phase, where the subjects were made familiar with the type of artifacts in focus using an AB-paradigm. The second phase downsized the artifacts and introduced the ABX-paradigm. Feedback was given on whether the response was correct or not. In the third phase artifacts were downsized even more and the task situation closely resembled the test phase, as feedback was no longer given. In both the second and the third phase it was required that performance was above chance level before going on to the next phase. In the final phase the actual test stimuli of interest were used, and response feedback was no longer given. Total training time, i.e. time used for phase one to three, was approximately one hour.

Training	**Phase 1. Priming**	
	Task for test subject	Build up a glossary with regard to sound quality both orally and in writing.
	Experimenter	Instruction with pairs of pictures (one distorted, the other not) and link to sound is made. Experimenter is active sparring partner, sitting next to subject.
	Paradigm	AB paradigm, i.e. sounds presented in pairs (processed / unprocessed).
	Stimuli	Profound artifacts in different degrees. Presented with increasing difficulty.
	Phase 2. Consolidation	
	Task for test subject	Expand glossary with regard to sound quality both orally and in writing, as well as identify the sound that corresponds to X.
	Experimenter	The experimenter is only sparring if asked for, placed about one meter behind the subject.
	Paradigm	ABX-paradigm with response feedback (right / wrong).
	Stimuli	Profound and severe artifacts. Presented with increasing difficulty.
	Criterion	High correct score is criterion for passing on to pre-test phase.
	Phase 3. Pre-test	
	Task for test subject	Identify the sound that corresponds to X, use glossary to reach decision.
	Experimenter	The experimenter is outside the test booth.
	Paradigm	ABX-paradigm, no feedback.
	Stimuli	Moderate artifacts in different degrees. Presented in random order.
	Criterion	High correct score is criterion for passing on to test phase.
Test	**Phase 4. Test**	
	Task for test subject	Identify the sound that corresponds to X, use glossary to reach decision.
	Experimenter	The experimenter is outside the test booth.
	Paradigm	ABX-paradigm, no feedback.
	Stimuli	Test-stimuli with mild artifacts. Presented in random order.

Table 1: Outline of procedure used for perceptual audio evaluation by hearing-impaired listeners.

Test subjects

Participants were six hearing-aid users (three men and three women) with symmetrical sensorineural hearing loss, aged between 50 and 76 years (mean=62.0; SD=11.8), see Fig. 4.

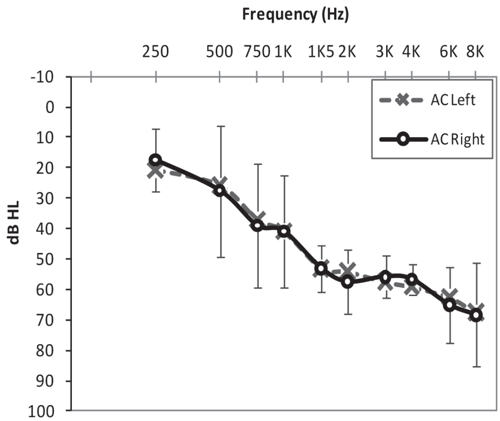

Fig. 4: Mean air conduction thresholds for the six test subjects. Vertical bars indicate ± one standard deviation of the left ear.

Material

Speech samples ("Speech 1") consisted of sentences from the Dantale II-corpus (Wagener *et al*., 2003). Music samples ("Music 3" and "Music 4") included two types of music chosen for rendering maximum artifacts and having a wide frequency spectrum. As our motivation was to study the impact of the developed guided stepwise training procedure, the material for the experiment was chosen to be similar – but not identical – to the pilot study; speech samples were from the same corpus, whereas new music samples were selected based on the same criteria. Half of the stimuli were processed to contain artifacts similar to one type of MPEG4 coding.

During the four phases of the procedure, artifacts were presented with ever decreasing magnitude. Great care was taken to generate artifacts with a similar perceptual character, only changing the magnitude of the artifacts. In the final two phases, i.e. pre-test and test, the stimuli were presented in random order within each phase (see Table 1).

Set-up

The test was conducted in a soundproof room. Stimuli were presented via a single loudspeaker in front of the subject. The subjects were fitted bilaterally with digital hearing aids (Oticon Syncro ITE) having eight-channel slow-acting compression and 0.8 mm comfort vents. Automatic features such as noise reduction and directional microphones were turned off. Audibility was verified by measuring free-field aided thresholds binaurally in an audiometric booth. The test signal used was one-third octave noise with a centre frequency of 0.25, 0.50, 1, 2, 4 or 8 kHz. For each of these frequencies, a target threshold was defined based on the one-third octave speech levels corresponding to an overall level of 65 dB SPL (specified in ANSI S3.5, 1997) to ensure that high-frequency speech sounds would generally be audible. Stimuli were presented at 65 dB SPL. Each condition was tested four times for each of the six subjects, resulting in 24 data points per condition.

RESULTS

For all tested conditions, scores were significantly above chance ($p<0.05$, binomial test), indicating that the subjects could now reliably discriminate between unprocessed and processed sounds for both speech and music samples (see Fig. 5). Even stimuli with the smallest degree of artifacts could easily be identified, as performance was between 80% and 100%.

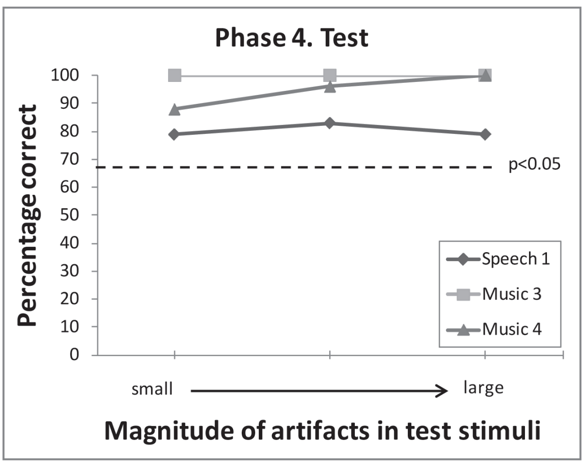

Fig. 5 After guided stepwise training procedure: Results of six hearing-impaired subjects. Subjects can now reliably discriminate between processed and unprocessed sounds for all conditions, including the speech samples they had heard before but could not discriminate in a pilot study (Fig. 2).

DISCUSSION AND CONCLUSION

The results of the presented studies seem to indicate that subjects with mild to moderate hearing loss are able to perform perceptual audio evaluation on stimuli containing artifacts, but only when certain conditions are met. After a guided stepwise training program, performance of the hearing-impaired subjects was significantly different from chance performance, but without this training program the subjects failed on almost all tested conditions. A possible explanation for the better performance in the experiment could be additional training time, as the same six subjects participated in both the pilot study as in the experiment itself. However, it is noteworthy that correct scores were highest on the music samples, which the subjects had not been exposed to before. It would be expected from the pilot study to see poor performance on these stimuli, if training time was the primary influencing factor. This seems to suggest that the guided stepwise training program had a significant impact on the test results. We conclude therefore, that carefully designed task training, especially with regard to priming to the nature of the test stimuli, is essential when hearing-impaired subjects are to be used for perceptual audio evaluation.

The implications for hearing research are rather far-reaching: based on the results from the pilot study we might falsely have accepted the hypothesis, that subjects with mild to moderate hearing loss are unable to perceive certain artifacts, while in fact we measured lack of training rather than limitations of the impaired auditory system. Targeted task training, for example by a guided stepwise training program, can be an important tool to explore these limitations, as results are less likely to be dictated by insufficient training. A question that remains though is how much training is required with regard to magnitude and steps of artifacts, number of trials and number of sessions.

ACKNOWLEDGEMENTS

We wish to thank Søren Bech, Bang and Olufsen A/S, for a useful discussion which inspired this work. Furthermore, we thank Tobias Neher, Research Centre Eriksholm, for providing the user interface for the ABX-paradigm.

REFERENCES

ANSI S3.5 (**1997**). "Methods for calculation of the speech intelligibility index," American National Standard.

Baddeley, A. D. (**1986**). *Working memory* (Oxford University Press, Oxford).

Baddeley, A. D., and Wilson, B. A. (**1994**). "When implicit learning fails: Amnesia and the problem of error elimination," Neuropsychologia **32**, 53-68.

Bech, S., and Zacharov, N. (**2006**). *Perceptual Audio Evaluation - Theory, Method and Application* (Wiley, West Sussex).

Clark, D. (**1982**). "High-Resolution Subjective Testing Using a Double-Blind Comparator," J. Audio Eng. Soc. **30**, 330-338.

Fillingham, J. K., Hodgson, C., Sage, K., and Lambon-Ralph, M. A. (**2003**). "The application of errorless learning to aphasic disorders: A review of theory and practice," Neuropsychological Rehabilitation **13**, 337-363.

Hansen, V. (**1987**). "Establishing a panel of listeners at Bang and Olufsen: a report", In *Perception of reproduced sound*, edited by S. Bech and O. J. Pedersen, (Stougaard Jensen, Copenhagen).

ITU-R (**1990**). "Recommendation BS.1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunications Union Radiocommunication Assembly.

Miller, W. R., and Rollnick, S. (**2002**). *Motivational interviewing; preparing people for change* (The Guilford Press, New York).
Wagener, K., Josvassen J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise", Int. J. Audiol. **42**, 10-17.