Recruiting and Evaluation Process of an Expert Listening Panel

A. Sontacchi, H. Pomberger and R. Höldrich
Institute of Electronic Music and Acoustics, Email: sontacchi@iem.at
University of Music and Performing Arts Graz, Austria

Introduction

During the last few years an increased desire to assess and quantify acoustical properties of technical products can be recognized. Even the objective is obvious several aspects have to be considered. In [1] several related perspectives concerning the term sound quality are discussed. If the requested definition is given the related properties have to be determined and evaluated. In common the assessment can be obtained via objective or subjective evaluation among defined framework conditions. Objective evaluation methods are based on modeling the human sound perception. They can deliver qualitative and quantitative acoustical descriptions of the investigated products. However, only a limited number of application areas can be treated reliably [2]. The predominant crucial aspect due to the complex interaction of acoustical stimulus and auditory perception is neither linear and additive nor time-invariant. Overlapping and competitive acoustical stimuli can be instructively assessed by a group of so called expert listeners, cf. [3]. However, specific demands on the listening panel are claimed to extrapolate the results of the panel to the general population involved with the evaluated product. Moreover, a well trained and proper selected panel will show consistent vocabulary usage and reliable rating behavior. Reduced variability within evaluation rating will shrink the confidence interval. Reliable and satisfying statistical conditions are obtained even with a reduced number of subjects and this will save money and time, cf. [4]. Within this article we discuss the selection process of designated listeners with special abilities to assess specific acoustical properties and to quantify overall affective properties. The addressed listeners are musicians and audio engineers. Based on reported selection procedures in [5, 6, 7] an adapted version of the selection process has been developed and accomplished with a group of 62 persons.

Panel Selection Strategies

Selecting subjects for an expert listening panel is a quite time consuming task. Depending on the purpose of the panel different aspects have to be considered, such as hearing ability and listening skills. The so-called *generalized listener selection* (GLS), cf. [5, 6], outlines a general framework for conducting a selection process. In the same way as other reported selection processes, cf. [8, 7], it applies a three-stage procedure, consisting of a questionnaire followed by audiometry and one or more test experiments. In the following, the particular accomplished stages are reported in detail.

Questionnaire

The purpose of the questionnaire is twofold. On the one hand it covers the necessary collection of the candidates contact dates, age and gender as well as further information, such as profession, experience in listening to or performing music, etc., which is eventually necessary for later analyses. On the other hand it provides a fast and efficient way to preselect the candidates. If a large number of volunteers enrolls during recruitment thus only a limited number of the most suitable candidates may be chosen for the subsequent administered test sessions. Otherwise, at least subjects who are definitely unsuitable, e.g. due to their temporal availability, can be excluded from the further selection stages, cf. [8, 5, 7]. Our call was addressed to a very specific group of listeners. Only musicians and audio engineers (students as well as teachers) at the University of Music and Performing Arts Graz were invited to apply for the selection process. Thus none of them was excluded due to the questionnaire. This predefined focus will ensure educated critical listeners but apart from that it will represent a biased sampling of the general population.

Audiometry

Individual hearing threshold levels of subjects can be determined by several psychophysical measurements, cf. [2]. For practical reasons we used a method of adaptive stimuli realized by the 3-down-1-up procedure [9]. Thereby, a sequence of two sound sources A and B is presented to the subject. During each trial only one of them is active and the subject's task is to indicate the active one. After 3 consecutive correct responses the stimuli level is decreased or increased after an incorrect one. The step-size starts with a defined value and is halved after the first turnaround. This adaptive procedure is stopped, if the variance of the last 3 turnarounds is below a defined value, averaging these last 3 turnarounds yields the hearing threshold level. The audiometry was conducted solely via headphones within an acoustically adapted room fulfilling the requested specifications in [10]. Results of the audiometry are summarized in Figure 1.

Listening skills

Panel selection processes usually apply a battery of test experiments to evaluate the listening skills of the candidates.

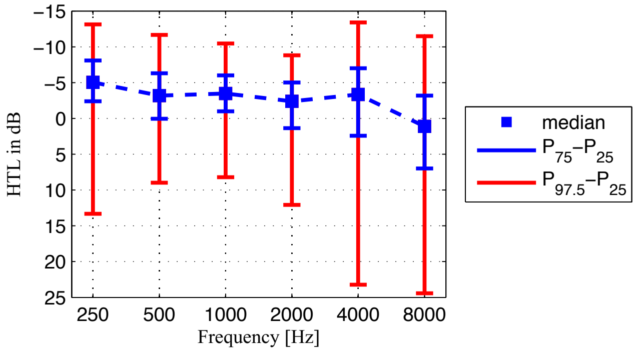


Figure 1: Summarized results of the audiometry, median and inter-percentile-ranges of the measured hearing threshold loss (HTL). Positive values indicate a reduced hearing ability.

Test paradigm

In the test experiments a paired comparison procedure with four levels of stimuli was employed. Whereby the stimuli include an upper and a lower anchor and two closely spaced stimuli in between. This yields a set of 6 sample pairs from which the pair containing the two anchors is clearly discriminable, and the one formed by the two middle stimuli challenges the subjects. Each sample pair was judged repeatedly by each subject to evaluate the subjects (intrarater) reliability. Moreover the collected data allows for studying the (interrater) agreement between the subjects. The test paradigm was adopted form the GLS, but instead of repeating each pair of sample equally often, as in the original paper [5], the number of repetitions has been chosen based on the assumed discrimination difficulty. Repeating the easy pairs less often than the challenging ones is beneficial for two reasons. Firstly, it avoids to fatigue the subjects by judging many repetitions of unambiguous pairs. This also results in a shorter overall test duration. Secondly, reliability errors in less often repeated pairs are implicitly weighted stronger in the subsequent analysis. Both permutations of each sample pair appeared equally often during an experiment to avoid biasing the results by the presentation order. Thus the repetitions are even numbers only. How often the six different sample pairs were repeated is given in Table 1(b).

Stimuli

Altogether, four experiments where conducted concerning an elementary set of auditory attributes (loudness, timbre, audio quality, and stereo width). The stimuli levels used in each experiment are summarized in Table 1(a). The stimuli used in the first experiment (loudness) are similar to the proposed set in [5], but the middle stimuli were placed more off-center to achieve a maximum amount of different intervals between the six sample pairs. For the second experiment (timbre), a likewise set of stimuli was arranged. The formant region of the vowel "a" was raised in a sample of pink noise by filtering. This provides easy-to-describe sound colorations of variable intensity. The third experiment (audio quality) is also similar to one reported in [5], but a more up-to-date set of speech codecs was applied. The last experiment (stereo width) deals with spatial audio. The idea was taken from [7] but instead of an adaptive procedure it was fit into the 4-level paired comparison test paradigm.

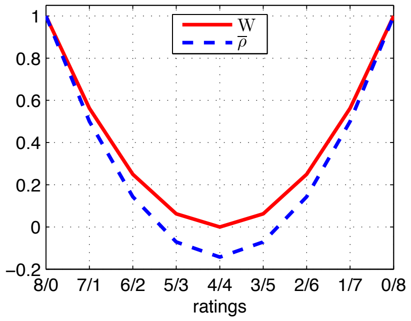


Figure 2: Kendall's coefficient of concordance W and the average Spearman rank correlation coefficient $\overline{\rho}$ for all possible outcomes of a 8 times repeated paired comparison.

Analysis of results

Intrarater reliability In the GLS the suggested measure for the intrarater reliability is the averaged Spearman rank correlation coefficient $\overline{\rho}$. Therefore the rank correlations were computed between the repetitions of six different sample pairs in each experiment for each subject. Since the averaged correlation coefficient $\overline{\rho}$ may also results in negative values, in [6] it is recommended to take its absolute value. A more elegant way to address this problem is to use Kendall's coefficient of concordance W as measure of agreement between the repetitions of a sample pair. The values of W range between 0 and 1, and Rae and Spencer showed in [11] that W is directly related to $\overline{\rho}$ by

$$W = \frac{1}{k} + \overline{\rho}\,\frac{(k-1)}{k} \qquad (1)$$

where k is the number of rankings. Figure 2 shows the difference between the two proposed measurement parameters evaluated for all possible outcomes of a 8 times repeated paired comparison. If the absolute value of the averaged Spearman rank correlation coefficient is used, additional erroneous ambiguity will be introduced. Thus as a more appropriate measure of intrarater agreement W was averaged over the 6 different sample pairs in each experiment for each subject.

Interrater agreement In the GLS the rank correlations between the subjects for all evaluated sample pairs are calculated. To measure the interrater agreement the averaged correlation for each subject in each experiment is used. Probably, this leads to an increased correlation with the intrarater agreement. A more decorrelated measure of interrater agreement is obtained if the paired comparison data is first transformed into a ranking of the stimuli for each subject in each experiment, e.g. by summing the rows of the preference matrix. Subsequently, the rank correlation between a subject's ranking and the ranking of the remaining subjects is averaged. Thereby the maximum achievable interrater

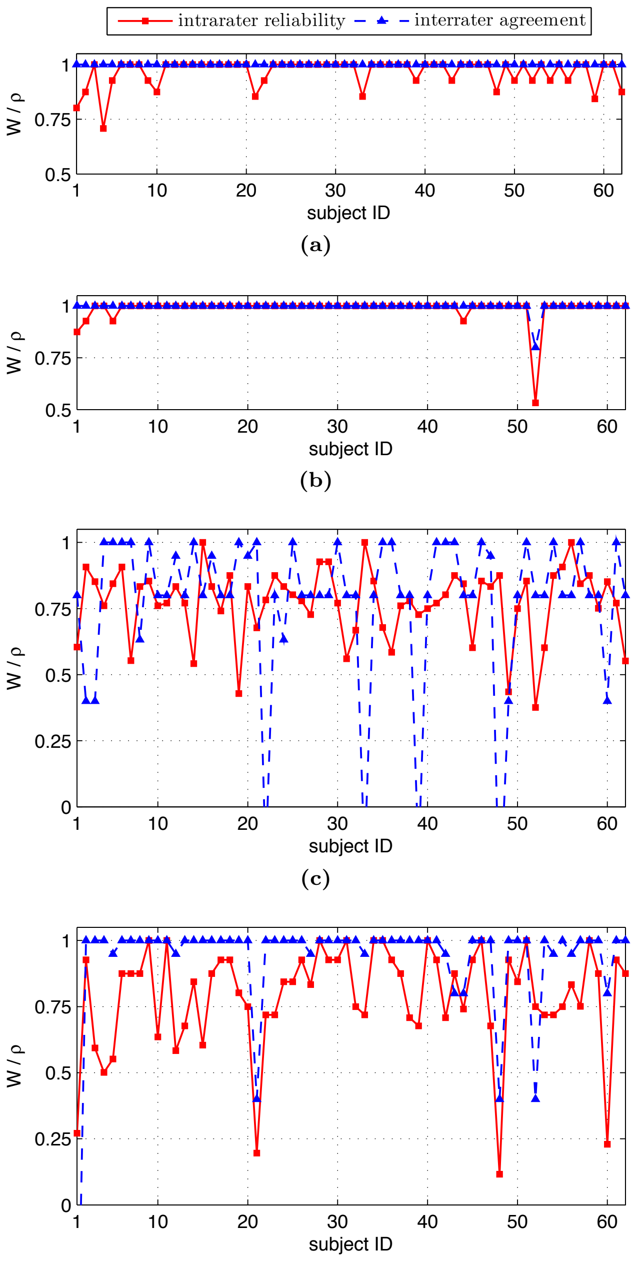
	Loudness	Timbre		Audio quality	Stereo width
Sample	Pink noise	Pink noise		Phonetically balanced speech at telecommunications bandwidth (300-3400Hz)	Classical music
		@1kHz	@1.4kHz		
Level 1	7dB	+8dB	+4dB	PCM	100%
Level 2	5dB	+4dB	+2dB	ILBC (13,33 kbit/s)	60%
Level 3	4dB	+3dB	+1.5dB	AMR (10,2 kbit/s)	50%
Level 4	0dB	+0dB	+0dB	AMR (5,15 kbit/s)	0%

(a)

Pair	(2,3)	(1,2)	(1,3)	(3,4)	(2,4)	(1,4)
Repetitions	8	8	6	4	4	2

(b)

Table 1: Summary of the stimuli levels used in the subjective experiments (a) and how often the sample pairs were repeatedly assessed by the subjects (b) sorted by increasing stimuli difference.



(a)

(b)

(c)

(d)

Figure 3: Intrarater reliability and interrater agreement of the subjects for the four test experiments: (a) loudness, (b) timbre, (c) audio quality, and (d) stereo width.

agreement for a single subject depends on the concordance within the whole group of subjects. To avoid this biasing by the group, instead of averaging we calculated the interrater agreement as rank correlation coefficient between a subject's ranking and a reference ranking. Whereby the reference was defined as the ranking shared by most of the subjects, which in this study is identical to the presumed stimuli ranking in all experiments.

Experiment results Figure 3 displays intrarater reliability and interrater agreement of all subjects for the four test experiments. The proposed loudness experiment is much more challenging as in [5], anyhow similar results are obtained. Experiment 2 exhibits that the presented timbre deviations can be easily identified by our listeners. Possible reasons might be that students at our University have to proceed an entrance exam and they are trained through various courses. Both consistency of subjects and their agreement is reduced when assessing the audio quality of speech codecs. The first is caused by small deviations in the used stimuli and possible switches in decision strategy during the experiment due to the multidimensionality of audio quality. The latter can be explained that even subjects give consistent answers their decision can relay upon different preceptive attributes of audio quality. In the last experiment subjects show good agreement. The distribution of reliability measures is due to small but perceivable differences. Audio engineers are used to the term *stereo width* and showed significantly better reliability results than musicians, thus these skills can be trained.

Probabilistic Choice Models The paired comparison paradigm also allows for analysis of the data by probabilistic choice models. In [12] a Matlab function to estimate choice model parameters is presented. These parameters reflect the distribution of the stimuli along the perceptual continuum. The Bradley-Terry-Luce (BTL) model, cf. [12], fits well (p=0.93) for Exp. 1. In case of Exp. 2 the stimuli can be nearly perfectly discriminated. The stimuli produce disjunct distributions along the perceptual continuum, thus probabilistic choice models have to be rejected. In Exp. 3 the multidimensional stimuli can be explained, using reliable subjects (reliability > 0.5) by an elimination by aspects (EBA) model (p=0.19). In presence of a reduced signal to noise ratio some subjects rated these stimuli pairs obviously different to the general affective measure. This might be avoided through more clearly defined instructions and training preceding the experiment. Regarding subjects that agree on the same ranking, for Exp. 4 a BTL model (p=0.17) can be used.

Verbal Abilities

Members of a listening panel should show good verbal abilities for expressing their sensations. To asses the verbal abilities of the candidates an alternating verbal fluency test (category switching) was conducted, cf. [7]. Subjects were instructed to name as many different terms as possible (in German) within a limited time of 60s, alternately belonging to one of two different semantic categories. To get comparable results, the selected categories were "fruits" and "animals". Before the test a familiarization session with two different categories was absolved. For the fluency score the number of correct terms was counted, whereas repetitions and category preservations were not included. Results show a similar distribution (mean = 17.4, std = 3.8) as reported in [7].

Selection Process

Based on considerations about reduced discrimination abilities due to increased hearing thresholds, cf. [2], subjects with severe deviations form the normal hearing threshold, i.e. HTL values above 20dB, are excluded. To achieve 30 qualified panelists the remaining subjects are further evaluated. The reliability and agreement results as well as the normalized verbal fluency score (vf) are aggregated to an overall score,

$$\text{overall score} = \frac{1}{5}\left(\text{vf} + \sum_{i=1}^{4} \text{intra}_i \cdot |\text{inter}_i|\right). \qquad (2)$$

Based on the overall high-score list 30 candidates were preselected. Furthermore, listeners performing most unsatisfactorily in any experiment are excluded until 30 candidates are left. These two lists of candidates are merged, i.e. listeners fulfilling both criteria are selected. Caused by the fact that the merging process yields only 26 subjects, the remaining 4 were hand-picked.

Conclusion

For the selection of expert listeners we proposed improved measures for intrarater reliability and interrater agreement. Results show that unprepared but experienced listeners can discriminate well small deviations in loudness and colored noise. Moreover, the mean uncertainty concerning loudness differences is far below 1dB and even soft colorations within the speech related area can be detected easily. We mentioned that instructions considerably influence the rating behavior of listeners. The plausibility of stimuli ratings were examined based on probabilistic choice models. We propose a merged selection criteria resting upon a multi-dimensional data set obtained by performing an audiometric test, four basic experiments and an additional verbal fluency test.

Acknowledgements

This project is funded by the Austrian Research Promotion Agency (FFG), the Styrian Government and the Styrian Business Promotion Agency (SFG) under the COMET programme. Thanks to the people of the CNBH at the Univ. of Cambridge, UK, for providing us with audiometry software, to all subjects conducting our experiments, and to all project partners for assistance.

References

[1] U. Jekosch. Sound Quality Assessment in the context of Product Engineering. In *Proc. of the 4th European Conf. on Noise Control*, 2001.

[2] S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application.* Wiley, 2006.

[3] ITU-R-BS.1116-1 Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, 1997.

[4] S.E. Olive. Differences in Performance and Preference of Trained versus Untrained Listeners in Loudspeaker Tests: A Case Study. *J. Audio Eng. Soc*, 51:806–825, 2003.

[5] V.V. Mattila and N. Zacharov. Generalized listener selection (GLS) procedure. In *Proc. of the AES 110th Conv.*, 2001.

[6] D. Isherwood, G. Lorho, V.V. Mattila, and N. Zacharov. Augmentation, application and verification of the generalized listener selection procedure. In *Proc. of the AES 115th Conv.*, 2003.

[7] F. Wickelmaier and S. Choisel. Selecting participants for listening tests of multichannel reproduced sound. In *Proc. of the AES 118th Conv.*, 2005.

[8] S. Bech. Selection and training of subjects for listening tests on sound-reproducing equipment. *J. Aud. Eng. Soc*, 40:590–610, 1992.

[9] H. Levitt. Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, 49:467, 1971.

[10] ISO 8253-1 Pure tone audiometric test methods. Part 1: Basic pure tone air and bone conduction threshold audiometry, 1989.

[11] G. Rae and J.E. Spencer. Average Spearman's Rho, Concordance, and the Matching Problem: Some Relationships. *The American Statistician*, 45(2):161–162, 1991.

[12] F. Wickelmaier and C. Schmid. A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers*, 36(1):29–40, 2004.