Performance Review of an Expert Listening Panel

Matthias Frank, Alois Sontacchi
Institut für Elektronische Musik und Akustik , Email: frank@iem.at
Universität für Musik und darstellende Kunst Graz, Austria

Introduction

Within a research project in the last 3 years, an Expert Listening Panel (ELP) of 41 subjects was recruited [1], trained [2] and employed for several experiments [3, 4, 5]. These experiments used different evaluation methods and studied quality of codecs and earphones, speech intelligibility and spatial attributes of sound fields.
As it is desirable for an ELP to yield only small deviations in the results, both consistent and agreed answers are necessary. Thus, the product of interrater agreement and intrarater reliability is used as performance measure in this review (like in the previous recruitment). For each experiment and ELP member, intrarater reliability and interrater agreement are derived and compared. Furthermore, the review investigates other relations, like the effect of general and specific training (preliminary tests) and the individual duration of the experiments on the performance, as well as the comparison of the performance in the experiments and in the recruitment.

Performance Measures

The interrater agreement is derived from the correlation between the individual ratings and the mean ratings. As the experiments used different evaluation methods and paradigms, the intrarater reliability is mostly derived from different parameters. In order to normalize the performance measure for each experiment, it is transformed into a scale between 0 and 1: The subject with the best performance gets a value of 1, the subject with the worst performance gets 0. In the case that a subject did not participate in an experiment, the mean value of all other subjects is used as the performance value. The following paragraphs provide a short description of each experiment and the computation of its measure for the intrarater reliability.

Experiment 1: Codec Evaluation 1 [3]

The first experiment evaluated the sound quality of a low-latency codec for wireless transmission on stage. A training phase served as preliminary test by investigating the audibility thresholds of codec artifacts for different sounds. For the detection of the thresholds, a 3-Alternative-Forced-Choice triangle-test [6] was used to compare encoded sounds to the original sounds. The training level referred to the resolution of the codec.
In the experiment, a double-blind triple-stimulus with hidden reference method [7] was used with ratings on a 5-point continuous scale. As measure for the intrarater reliability, the relative amount the subject judged the reference with the best possible quality is used. Intrarater reliability and the duration of the experiment correlate with 20%, the intrarater reliability and the interrater agreement correlate with 43%.

Experiment 2: Codec Evaluation 2 [3]

This experiment continued the first experiment by a four-dimensional comparison. The direct comparison was based on the MUSHRA standard [8]. In this experiment, the same measure as in the first experiment is used for the intrarater reliability. There was no distinct training for the second experiment. Intrarater reliability and the duration of the experiment correlate with 21%, the intrarater reliability and the interrater agreement correlate with 20%.

Experiment 3: Earphone Evaluation

The third experiment evaluated the influence of different processing algorithms on earphones by four attributes. The ratio between the standard deviation of all conditions and the standard deviation of the repetition of the same condition is used as measure for the intrarater reliability. The subjects were prepared for the experiment by a training phase with the same attributes as in the experiment. The correlation between intrarater reliability and interrater agreement in the experiment is 22%.

Experiment 4: Speech Intelligibility with Active Noise Cancellation (ANC) Headphones

This experiment investigated the speech intelligibility test using ANC headphones. The performance measure for this experiment is the relative amount of correct answers. The comparison to the number of semantic pairs in the verbal fluency test of the recruitment shows a correlation of 17%. As some of the subjects of this experiment were not part of the ELP, a comparison to non-expert listeners can be drawn: On average, they show similar results.

Experiment 5: Spatial Audio 1 [4]

Experiment 5 consisted of 2 parts that were conducted as pair-wise comparisons about source width. The measure for intrarater reliability is based on the range of the Thurstone scales [9] that result from the pair-wise comparison matrices. Comparing N stimuli using a value of -1 for A>B and 1 for A<B, would result in a scale that ranges from -(N-1) to (N-1). By dividing the scale values by 2(N-1) results in a range of the scale between 0 and 1. These values are only possible, if one stimulus is always preferred over all other stimuli and if one stimulus is never preferred over all other stimuli and no cyclic triads occur. Thus, the difference between the smallest and the greatest value of a scale is a measure for the relia-

bility of the subject. Intrarater reliability and interrater agreement correlate with 66% in the first part and 48% in the second.

Experiment 6: Spatial Audio 2 [5]

The first part of experiment 6 was similar to experiment 5. In the second part, the pairs were compared on continuous scales for five spatial attributes. Therefore, an adaptation of the reliability measure is necessary that incorporates not only the range of the scale but also for the distribution of the answers on the scale. Thus, the measure for the intrarater reliability in the second part is chosen to be the product of the difference between the smallest and the greatest value of each individual scale and the standard deviation of this scale. The correlation between intrarater reliability and interrater agreement is 38% in the first and 58% in the second part.

Training

The training of the ELP was done with a level-based training software [2] at the ELP members' homes. The training time and the maximum achieved training level correlate with 91%. There was general training and specific training that served also as preliminary tests for experiments 1 and 3. The overall training time correlates with 54% and 82% to the time for the two specific trainings and with 83% to the general training.

Correlations and Discussion

The performance in all experiments is correlated, except for the speech intelligibility using ANC headphones. This means that some subjects always performed better than others. This ranking is also correlated with the performance in the recruiting, i.e. a careful design of the recruitment process is very important for the performance of an ELP. The exceptional case with the ANC experiment can be explained by the fact that everybody is well trained in speech intelligibility by daily verbal communication. This is underlined by the comparison to the performance of non-expert listeners in this experiment. Subjects with a high intrarater reliability show also a high interrater agreement. Subjects who took their time in the experiments tended to perform better.

The training time for all general and specific training is correlated. Thus, it was always the same group of subjects which trained more or less than the average subject. Training time and achieved training level yield a strong correlation, i.e. the subjects who trained more, achieved higher training levels. This correlation suggests statements about the attitude of the subjects: Most subjects succeeded in their training and did not need many repetitions of the exercises to achieve the next level. On the other hand, if a subject did not achieve the next level, he or she mostly did not try again. For further prove of this statement, the training software would have to be modified in a way that it saves how often and how long each subject trained on each level.

Experiments with specific training benefited from this training phase. The general training shows weaker influence on the performance in the experiments. This can be explained by the fact, that all subjects are professional musicians and audio engineers or students in these fields. Thus, they train their ears and listening skills every day anyway. Nevertheless, the general coloration training had a positive influence on experiments about coloration.

Conclusion

A proper recruitment process is important for the performance of an ELP. Subjects who take more time in the experiments tend to perform better. Interrater agreement correlates with intrarater reliability. In experiments with specific training, the better trained subjects perform better. On the other hand, general training shows not much effect. This is because ear training is a part of each ELP member's daily life, anyway. Further updates of the training should care for these findings by increasing the amount of specific training and decreasing the amount of general training.

Acknowledgments

This study was part of the project AAP, which is funded by Austrian ministries BMVIT, BMWFJ, the Styrian Business Promotion Agency (SFG), and the departments 3 and 14 of the Styrian Government. The Austrian Research Promotion Agency (FFG) conducts the funding under the Competence Centers for Excellent Technologies (COMET, K-Project), a programme of the abovementioned institutions.

References

[1] A. Sontacchi, H. Pomberger, and R. Höldrich: Recruiting and Evaluation Process of an Expert Listening Panel, NAG/DAGA Rotterdam, 2009

[2] M. Frank, A. Sontacchi, and R. Höldrich: Training and Guidance Tool for Listening Panels, DAGA Berlin, 2010

[3] M. Frank, A. Sontacchi, T. Lindenbauer, and M. Opitz: Subjective Sound Quality Evaluation of a Codec for Digital Wireless Transmission, AES 132nd Convention Budapest, 2012

[4] M. Frank, G. Marentakis, and A. Sontacchi: A simple technical measure for the perceived source width, DAGA Düsseldorf, 2011

[5] F. Zotter, M. Frank, G. Marentakis, and A. Sontacchi: Phantom Source Widening With Deterministic Frequency Dependent Time Delays, DAFx-11 Paris, 2011

[6] ISO: Sensory analysis - Methodology - Triangle test, ISO 4120:2004, 2004

[7] ITU, ITU-R BS.1116-1 Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, 1994–1997

[8] ITU, ITU-R BS.1534-1 Method for the subjective assessment of intermediate quality level of coding systems, 2001–2003

[9] L. L. Thurstone: A law of comparative judgment, Psychological Review, Vol. 101, No. 2, 1994