Distributed Pair Evaluation in Context of MUSHRA Listening Tests

Maximilian Neumayer, Michael Schoeffler and Jürgen Herre
International Audio Laboratories Erlangen, A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Am Wolfsmantel 33, 91058, Erlangen, Germany

Abstract. MUSHRA (Multi-Stimulus Test with Hidden Reference and Anchor) is an established listening test method that is widely used for evaluating basic audio quality (BAQ) of audio codecs or audio systems in general. In many MUSHRA listening tests, the differences in BAQ between the audio codecs under test are hard to judge. Therefore, so-called expert listeners are needed to evaluate the audio codecs.

In software development exists a technique, called (distributed) pair programming, in which two programmers work as a pair together on the same code. Applying this technique is known to improve code quality. In this work, a distributed pair programming approach is integrated into the MUSHRA listening test method. In order to investigate the new approach in more detail, a preliminary listening test was conducted in which participants evaluated the BAQ of audio codecs. The participants were assigned into pairs and collaboratively evaluated the BAQ of the audio codecs. When participants worked in pairs, they were spatially separated from each other, but able to communicate by video-, voice-, and text-chat.

1 Introduction and Related Work

Pair programming is a software development technique in which two programmers collaboratively work on the same code on the same computer. Many studies have been carried out to investigate the benefits of pair programming and revealed that this technique can significantly increase code quality [1]. Distributed pair programming extends the methodology of pair programming by spatially separating two programmers who still collaboratively work on the same code. Stotts et al. describe distributed pair programming as feasible, effective, and pleasant for the participants [2].

The main contribution of this paper is to bring the idea of distributed pair programming into the context of evaluating audio codecs. Related to the development of new audio codecs, subjective evaluation methods (e. g., listening tests) are indispensable, however, they are very time- and cost-consuming. A major reason for this is that in many listening tests so-called expert listeners are needed, since only they have sufficient experience to detect very small audible differences between different audio codecs. Unfortunately, expert listeners are sometimes not available in a sufficient number in order to get statistically significant results. As an alternative, naïve listeners could be used for an evaluation, however, their ratings are known to be less statistically significant than those ratings given by expert listeners [3].

In this work, a framework is presented that enables to collaboratively evaluate audio codecs as a pair of two assessors. The validity of the presented framework is evaluated by a listening test in which participants rated the basic audio quality (BAQ) of short audio excerpts that were encoded in MP3 at different bit rates.

The conducted listening test is based on the MUSHRA (Multi-Stimulus Test with Hidden Reference and Anchor) methodology [4]. In a MUSHRA listening test, a reference stimulus is presented to the participants. The participants are asked to compare the reference to various conditions: the audio codecs under test, the hidden reference, and two anchors. The anchors are low-passed versions of the reference with cut-off frequencies at 3.5 and 7 kHz. Furthermore, the conditions are presented in random order without any information that would identify the condition being an audio codec under test, the hidden reference, or an anchor. Participants can instantaneously switch between the reference and the conditions when listening. For each condition, participants rate its BAQ. BAQ is defined as single, global attribute that is used to judge any and all detected differences between the reference and the condition. The BAQ is rated on a scale ranging from 0 to 100. A condition that sounds as the reference is expected to be rated with a score of 100. The conducted listening test consisted of two sessions:

Single session

A single participant is evaluating the stimuli according to the MUSHRA methodology.

Paired session

Two spatially separated participants are evaluating the stimuli according to the MUSHRA methodology. In addition to the single session, they can communicate via text-, video-, and voice-chat.

2 Method

2.1 Participants

Eight participants (three females and five males) volunteered to participate in the listening test. The participants had an average age of 29.1 years ($SD = 4.1$) and six identified themselves as audio professionals[*]. One participant indicated being an expert listener in timbre, another participant indicated being an expert listener in spatial audio, and two participants indicated being both. In the paired session, having a pair of an experienced listener and an inexperienced listener was expected to give ratings that are mainly influenced by the experienced listener. Therefore, pairs were arranged such that each pair consists of participants having similar levels of experience in evaluating audio codecs.

2.2 Stimuli

Four music excerpts (items) were selected to generate the stimuli : the song “Tom’s Dinner” performed by Suzanne Vega, the song “I feel fine” performed by The Beatles, a recording of a pitch pipe playing three different tones, and a recording of a stadium announcer with background noise from the audience. Each excerpt had a duration of about ten seconds. All excerpts were unencoded WAV files with a sampling frequency of 48 kHz and bit depth of 16. The unencoded excerpts served as references in the listening test.

For each reference, seven conditions were processed. Each reference was encoded into MP3 files having a constant bit rate (CBR mode) of 64, 80, 96, 112, and 128 kbps. Any of these bit rates is known to be distinguishable from a reference signal at least by expert listeners [5]. Higher bit rates, especially starting at 192 kbps, are known to be not distinguishable anymore [5]. Furthermore, two low-passed versions (3.5 and 7 kHz cut-off frequency) of the reference were processed which served as anchors. An EBU-R128 loudness normalization (see [6]) was applied to all stimuli to have the same perceived loudness during the listening test.

2.3 Materials and Apparatus

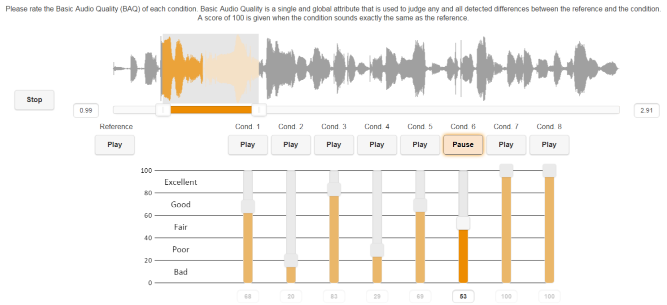
The used MUSHRA listening test software was based on webMUSHRA, an open-source software that can be customized to fulfill the needs of unique listening test designs [7]. In the single and paired session, a listening room was equipped with two 24” widescreen monitors, a Logitech HD C310 webcam, keyboard, mouse, and Beyerdymamic DT 770 250 Ω Headphones connected to a RME Babyface sound interface. The webcam and the second widescreen monitor were disabled in the single sessions. In the paired session, two desks having the same equipment were prepared in two different rooms. Furthermore, a modified version of webMUSHRA was used as listening test software. The modified version extended webMUSHRA by a text-chat window as well as a video- and a voice-chat. The video- and voice-chat was shown on the second widescreen monitor. In addition, the graphical user interfaces of the participants were synchronized. The synchronization included the navigation through the listening test and the ratings of the conditions. Moreover, participants could send each other a selected excerpt of the stimuli they were currently listening to. For example, such an excerpt could contain an audible artefact of a condition.

2.4 Procedure

First, participants were assigned to pairs. One participant of a pair had to take part in the single session before the paired session. The other participant took the paired session first and then the single session.

At the beginning of each session, instructions were shown to the participants. The instructions contained some information about the MUSHRA procedure, especially about the definition of BAQ. In the paired session, also some information about the communication features were shown to the participants. After reading the instructions, the participants could adjust the volume individually. Next, the participants rated the four items where each item consisted

[*] SD = standard deviation.



Please rate the Basic Audio Quality (BAQ) of each condition. Basic Audio Quality is a single and global attribute that is used to judge any and all detected differences between the reference and the condition. A score of 100 is given when the condition sounds exactly the same as the reference.

Figure 1 A screenshot of the graphical user interface used in the single and paired session.

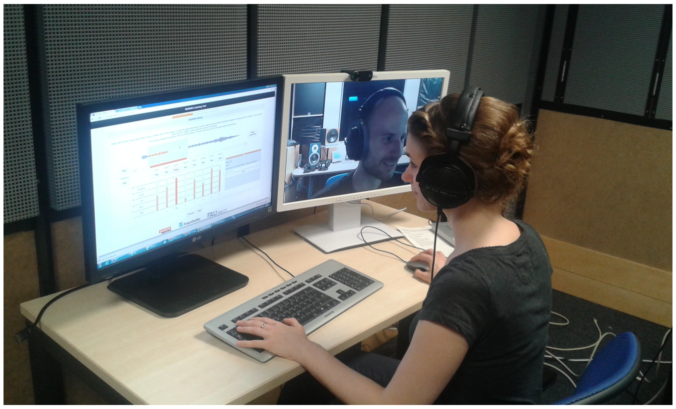


Figure 2 A picture of the paired session.

of eight conditions (hidden reference, five mp3 encoded files, and two anchors). A screenshot of the graphical user interface is shown in Figure 1. A picture of the paired session showing a participant rating the conditions and communicating with her partner is shown in Figure 2. Finally, the participants were asked about their age, gender, whether they are an expert listener in judging timbre, in spatial audio, in how many listening tests they have volunteered before, and whether they have a background in professional audio.

3 Results and Discussion

The average time needed to complete the single session was 15.5 min (*SD* = 6.1). In the paired session, participants needed 34.1 min (*SD* = 25.1) to complete the listening test.

In Figure 3, a boxplot is depicted that shows the ratings of the participants grouped by the two different sessions. As one can see, most of the conditions could be easily distinguished from the reference by the participants. According to a pairwise t-test, the ratings of the anchors and the bit rates up to 96 kbps were significantly[†] different to the ratings of the reference in the single session ($p \leq .001$) as well as the paired session ($p \leq .023$). No significant differences between the reference and the 112 kbps condition were found in the single ($p = .478$) and paired session ($p = .128$). The same applies to the 128 kbps condition (single: $p = .758$; paired: $p = .887$). The p-values are mainly influenced by the effect size of a condition and the number of ratings. Since the listening test was designed with the purpose to initially investigate the idea of distributed pair evaluation, only a low number of participants was recruited which resulted in rather high p-values. As a consequence, no reliable statement can be made whether distributed pair evaluation has a significant benefit compared to traditional evaluations. However, an interesting fact is that the pair of participants with the lowest experience (both indicated to have taken part in only two MUSHRA tests before) could not distinguish between a reference and a condition encoded at the highest bit rate, 128 kbps, a single time in the single session. Whereas in the paired session, they were able to correctly distinguish between a reference and a condition encoded at the highest bit rate at least once. The other (more experienced) participants were able to distinguish between a reference and a condition encoded at 128 kbps at least once in the single session. Therefore, the question arises whether distributed pair evaluation brings a benefit when the listeners are rather inexperienced. Answering this question, as well as identifying statistically significant benefits of distributed pair evaluation, requires to carry out more listening test with a higher number of participants. Nevertheless, our results show that ratings obtained from a "normal" MUSHRA procedure are at least in accordance to ratings obtained from a MUSHRA procedure in which distributed pair evaluation has been used.

[†] In this paper, the significance level is set to $\alpha = 0.05$

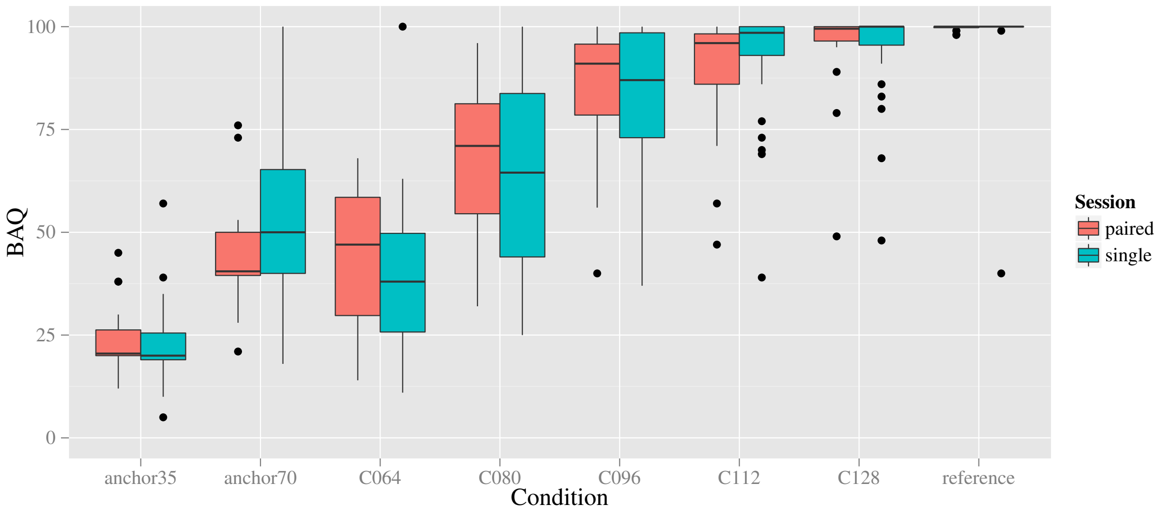


Figure 3 The BAQ ratings of the different conditions depicted as a boxplot.

4 Conclusion

In this work, distributed pair evaluation has been presented as a new approach for evaluating audio systems. A preliminary listening test showed that distributed pair evaluation leads to similar results as a traditional approach. However, in order to test whether distributed pair evaluation has a significant benefit, more listening tests have to be carried out.

References

1 A. Cockburn and L. Williams. Extreme Programming Examined. In G. Succi and M. Marchesi, editors, *The Costs and Benefits of Pair Programming*, pages 223–243. Addison-Wesley Longman Publishing Co., Inc., Boston, United States, 2001.

2 D. Stotts, L. Williams, N. Nagappan, P. Baheti, D. Jen, and A. Jackson. Virtual Teaming: Experiments and Experiences with Distributed Pair Programming. In F. Maurer and D. Wells, editors, *Extreme Programming and Agile Methods - XP/Agile Universe 2003*, volume 2753 of *Lecture Notes in Computer Science*, pages 129–141. Springer Berlin Heidelberg, 2003.

3 N. Schinkel-Bielefeld, N. Lotze, and F. Nagel. Audio Quality Evaluation by Experienced and Inexperienced Listeners. *Proceedings of Meetings on Acoustics*, 19(1), 2013.

4 International Telecommunication Union. ITU-R BS.1534 (Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems), 2014.

5 A. Pras, R. Zimmerman, D. Levitin, and C. Guastavino. Subjective Evaluation of MP3 Compression for Different Musical Genres. In *Audio Engineering Society Convention 127*, New York, United States, 2009.

6 European Broadcasting Union. Practical Guidelines for Production and Implementation in Accordance with EBU R 128 (Version 2.0), 2011.

7 M. Schoeffler, F. Stöter, B. Edler, and J. Herre. Towards the Next Generation of Web-based Experiments: A Case Study Assessing Basic Audio Quality Following the ITU-R Recommendation BS.1534 (MUSHRA). In *1st Web Audio Conference*, Paris, France, 2015.