Adaptions for the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) for elder and technical unexperienced participants

Christoph Völker[1], Thomas Bisitz[2], Rainer Huber[2] and Stephan M. A. Ernst[1]
[1] *CvO Universität Oldenburg, Cluster of Excellence 'Hearing4all', Email: christoph.voelker@uni-oldenburg.de*
[2] *Kompetenzzentrum HörTech, Cluster of Excellence 'Hearing4all'*

Introduction

The development of new audio codecs or any kind of signal processing schemes, e.g. algorithms for hearing aids, always includes a final stage of evaluation. At this stage the question is answered whether the algorithms work correctly and will bring the anticipated benefit.

The evaluation can be split into an instrumental and a perceptual part. In the instrumental part the algorithms will be evaluated by computer programs, i.e. models mimicking the human perception. In the perceptual part the algorithms are evaluated by real humans rating e.g. the quality of an algorithm. Ideally the models would predict the human ratings with a minimum error, so that the entire evaluation could be performed by the computer. As this ideal has generally not been reached so far, models of human perception have to be trained on real human data to become more accurate. To gather the necessary human data several methods were discussed in the literature [1].

A standardized method for the subjective assessment of intermediate quality level of audio systems is the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) [2, 3]. Among other audio systems the MUSHRA test is used in hearing aid technology (e.g. see [4]). However, the measurement practice shows some difficulties especially with elder and technically unexperienced participants, as commonly used in hearing aid research. Such test subjects are often challenged by the complexity of the method.

In the following the original MUSHRA test is called *MUSHRA classic* (figure 1).

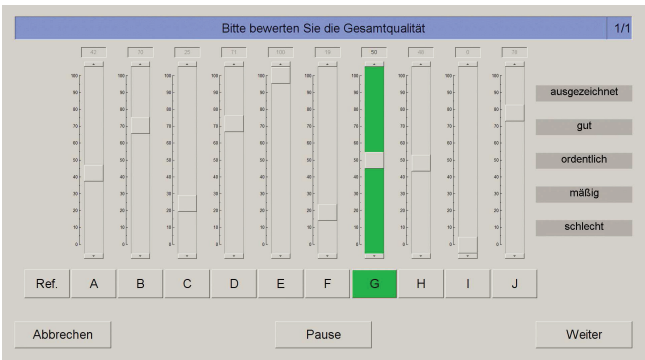

Figure 1: Screenshot of MUSHRA classic

The aims of the presented study are the introduction and evaluation of two adaptions for the original MUSHRA method to maximize its accessibility and consequently the applicability of this method. This will give us the opportunity to extend the user group of this standardized method from experienced and trained subjects for which it was initially developed [3, p. 4] to the target group of elder, technically unexperienced and hearing impaired people in order to improve the development of new assistive hearing systems.

MUSHRA adaptions

The two adaptions for the original MUSHRA method are called *MUSHRA simple* (figure 2) and *MUSHRA drag&drop* (figure 3).

The main differences of *MUSHRA simple* in comparison to *MUSHRA classic* are: *MUSHRA simple* uses eleven discrete buttons for the rating instead of sliders. The total number of stimuli per test-screen is restricted to six: one trial is split into two screens. In the second screen two algorithms from the first screen are presented again with a fixed rating that was given to them in the first screen. These two stimuli are the hidden reference and the hidden low-pass anchor. These two changes are intended to increase the clarity of the user interface and therefore to reduce the complexity of the MUSHRA method.

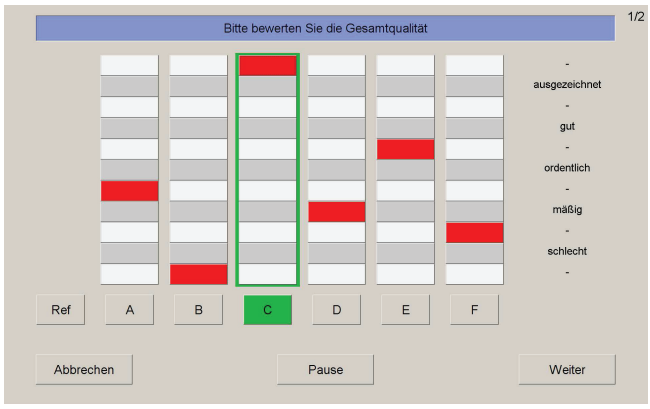

Figure 2: Screenshot of MUSHRA simple

In comparison to the other two versions *MUSHRA drag&drop* uses a drag & drop interface and lets the user sort the stimuli from left to right. This offers the possibility of an instantaneous visualization of the ranking, helping the assessor to easily check their rating. This adaption optimizes the intuitivity of the method.

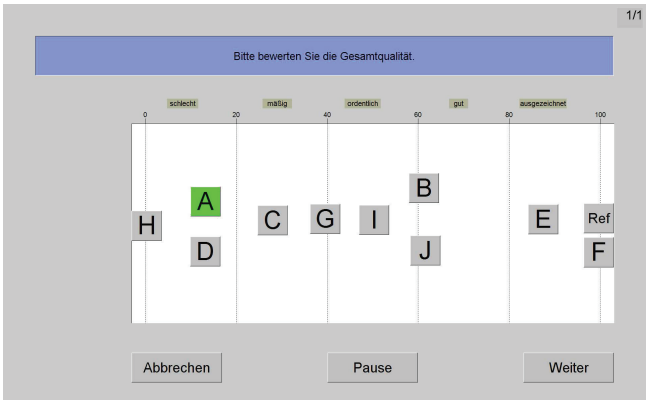

Figure 3: Screenshot of MUSHRA drag&drop

Test subjects and measurement protocol

Testing the hypothesis that our adaptions are especially suitable for elder and technically unexperienced participants the following subject factors were investigated in this study:

- age (young, old),
- hearing ability (normal, impaired),
- technical experience (experienced, unexperienced).

Five subject groups were built out of these factors:

G1: young, normal hearing, technically experienced ('control group'),

G2: old, normal hearing, technically experienced,

G3: old, normal hearing, technically **un**experienced,

G4: old, impaired hearing, technically experienced,

G5: old, impaired hearing, technically **un**experienced.

The categorization for the subjects' technical experience is based on answers to a questionnaire concerning their willingness to adopt new technologies [5]. In each group ten subjects were tested.

By using the original MUSHRA method (*MUSHRA classic*) and our two adaptions all subjects rated the overall quality of seven different noise reduction algorithms designed for the application in hearing aids together with a 'no processing scheme'. Additionally the subjects rated the quality of the hidden reference – supposed to have the best quality – and the hidden low-pass anchor – supposed to show the worst quality.

The subjects rated the noise reduction algorithms in three different realistic noise scenarios. The scenarios consist of scene-specific dialogues in a kitchen, a supermarket and a cafeteria. All noise scenarios were generated using the TASCARpro [6] software having a signal duration of 20 seconds and a signal-to-noise ratio of 2 dB. In the test procedure the signals are looped, and the user can switch between the stimuli instantaneously.

Each subject performed these tests twice on two different days, i.e. every participant gave 36 ratings per algorithm in total (12 ratings using each of the three methods).

Evaluation of adaptions

In this study we present the comparison of assessor performance for the three MUSHRA versions. The assessor performance is calculated based on the *eGauge* (expertise gauge) method [7, 8], that describes a way for the obligatory post-screening and a criterion for the rejection of a test subject from the final data analysis. All our assessors used all three methods in a randomized order, which allowed us to evaluate the test method adaptions by directly comparing the assessor performance.

With eGauge the performance of each assessor is calculated in terms of reliability and discrimination. By using a non-parametric permutation test [9] the 95% significance levels for both metrics are calculated. Assessors with values below these levels, i.e. having problems to distinguish between the test items or to reproduce an earlier rating have to be regarded separately and are therefore excluded from calculations concerning estimates of central tendency for the algorithms.

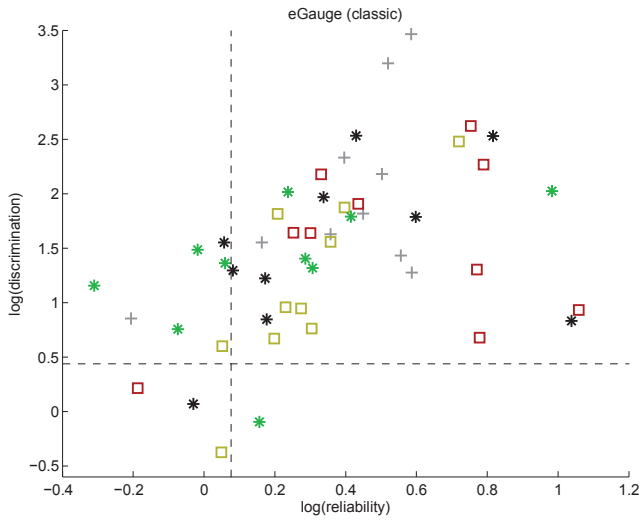

Figure 4: Scatterplot of reliability vs. discrimination for the assessors using MUSHRA classic. *Plus sign*: Control group G1, *stars*: technically experienced subjects (G2&G4), *squares*: technically unexperienced subjects (G3&G5). Also displayed are the 95% levels of significance for both metrics: 11/50 assessors have to be excluded from a mean value analysis.

Results

The performance based on the *eGauge* analysis of the assessors using all three MUSHRA version is displayed in figures 4, 5 and 6.

Figure 4 shows a scatterplot of the metrics reliability and discrimination calculated by the eGauge model for the assessors using MUSHRA classic. Also displayed are the 95% levels of significance for both metrics. Assessors with values below these levels have to be excluded from the final mean value analysis of the ratings. In this sense eleven assessors showed problems using the

MUSHRA classic user interface: one assessor from the control group G1, seven technically experienced assessors (groups G2 and G4) and unexpectedly only three technically **un**experienced subjects.

In figure 5 the scatterplot for MUSHRA simple is displayed. Here 19 assessors showed problems using the MUSHRA simple user interface: two assessors from the control group G1, eight technically experienced assessors (groups G2 and G4) and nine technically **un**experienced subjects.

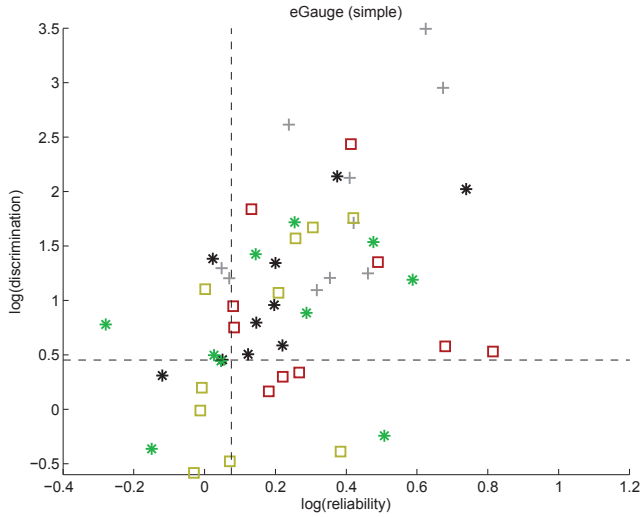

Figure 5: Scatterplot of reliability vs. discrimination for the assessors using MUSHRA simple. *plus sign*: Control group G1, *stars*: technically experienced subjects (G2&G4), *squares*: technically unexperienced subjects (G3&G5). Also displayed are the 95% levels of significance for both metrics: 19/50 assessors have to be excluded from a mean value analysis.

Figure 6 shows the assessor results for MUSHRA drag&drop. Using the drag & drop user interface nine assessors showed problems which would result in an exclusion from the final data analysis: two technically experienced assessors (groups G2 and G4) and seven technically **un**experienced subjects.

The rejection rates in percent for each subject group using each method are summarized in table 1. Additionally the rejection rates for the technically experienced and technically unexperienced assessors are calculated and displayed. The last row shows the rejection rates when considering all groups and all 50 assessors together.

Discussion & Conclusion

In this study we evaluated three possible user interfaces for the MUSHRA method. For this purpose we concentrated on the intuitivity and the accessibility of the user interface for different target groups, including elder and technically **un**experienced people with hearing loss, which is a typical category for test subjects in hearing aid research.

Interpreting the rejection rates for data sets given by the eGauge analysis (table 1), it can be concluded that the

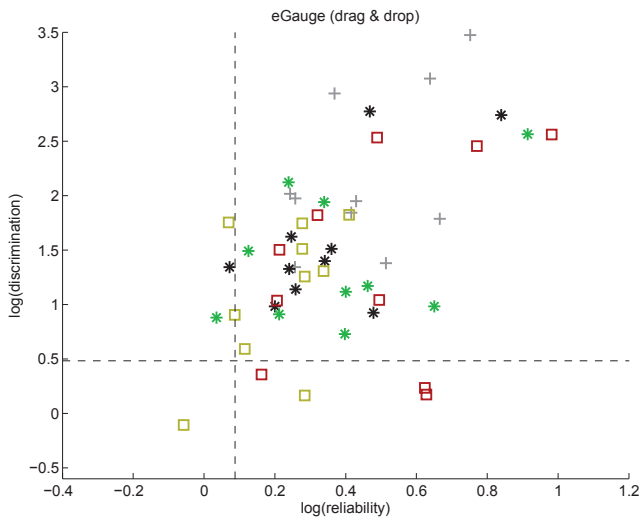

Figure 6: Scatterplot of reliability vs. discrimination for the assessors using MUSHRA drag&drop. *plus sign*: Control group G1, *stars*: technically experienced subjects (G2&G4), *squares*: technically unexperienced subjects (G3&G5). Also displayed are the 95% levels of significance for both metrics: 9/50 assessors have to be excluded from a mean value analysis.

Table 1: Assessor rejection rates

group	method		
	classic	simple	drag&drop
G1	10%	20%	0%
G2	20%	30%	10%
G3	20%	60%	40%
G4	50%	50%	10%
G5	10%	30%	30%
technically experienced	35%	40%	10%
technically un-experienced	15%	45%	35%
all groups	22%	38%	18%

new drag & drop interface shows the highest accessibility resulting in the lowest rejection rate averaged over all subjects. In particular this is true for the group of technically experienced subjects, where only 10% of our subjects showed problems to generate a reliable discrimination of the test items using the interface. However, we could not confirm our initial hypothesis that particularly elder and technically **un**experienced assessors would benefit from the new user interfaces. In this target group the original MUSHRA classic interface performed very well. Thus a differentiated choice of suitable user interface for the targeted assessor group is necessary when performing MUSHRA measurements. This would result in a specific efficient **combination** of test subject and test method.

In future work we want to develop further variations of established evaluation tools. We are interested in approaches that only differ in their user interfaces but not in the underlying test method theory. This would offer different ways of accessibility, while leading to compa-

rable results. In this scenario the optimal test method could be chosen for individual test subjects.

Acknowledgments

We would like to thank the DFG Exzellenzcluster Hearing4all for supporting this work. This work was funded by the German Ministry of Education and Research (BMBF), project „Model-based Hearing Aids“ (13EZ1127D) and the DFG research group (FOR1732). We thank Matthias Vormann and Miriam Kropp from Hörzentrum Oldenburg for performing the measurements.

References

[1] S. Bech, and N. Zacharov (2006): Perceptual Audio Evaluation — Theory, Method and Application. Wiley & Sons, Chichester.

[2] ITU-R BS.1534-1 (2001-2003): Method for the subjective assessment of intermediate quality level of coding systems.

[3] ITU-R BS.1534-2 (06/2014): Method for the subjective assessment of intermediate quality level of audio systems.

[4] C.S. Simonsen and S.V. Legarth (2010): A procedure for Sound Quality Evaluation of Hearing Aids. Hearing Review, 17(13),32–37

[5] F. Neyer, J. Felber and C. Gebhardt (2012): Entwicklung und Validierung einer Kurzskala zur Erfassung von Technikbereitschaft, Diagnostica, 58(2), 87–99

[6] G. Grimm and V. Hohmann (2014): Dynamic spatial acoustic scenarios in multichannel loudspeaker systems for hearing aid evaluations, 17. Jahrestagung der Deutschen Gesellschaft für Audiologie.

[7] ITU-R BS.2300-0 (04/2014): Methods for Assessor Screening.

[8] G. Lorho, G. Le Ray and N. Zacharov (2010): eGauge — A Measure of Assessor Expertise in Audio Quality Evaluations. Proceeding of the Audio Engineering Society. 38th International Conference on Sound Quality Evaluation, Piteå, Sweden, 13-15 June 2010.

[9] G.B. Dijksterhuis and W.J. Heiser (1995): The role of permutation tests in exploratory multivariate data analysis. Food quality and preference 6, 263–270.