Audio quality evaluation of soundbars using the multiple stimulus ideal profile method

Theresa Liebl[1], Sebastian Wakan[2], Oliver Curdt[2]

[1] *Institut für Rundfunktechnik, 80939 München, E-Mail: liebl@irt.de*

[2] *Hochschule der Medien Stuttgart, 70569 Stuttgart, E-Mail: curdt@hdm-stuttgart.de*

Introduction

In 2016, a new method to assess advanced sound systems, the multiple stimulus ideal profile method (MS-IPM), was introduced [1]. Being involved in the standardization process of such methodologies, IRT decided to investigate the characteristics of the method. The subject of the test was an audio quality evaluation of soundbars.

With decreasing sound quality of modern TV sets, soundbars have become a more and more popular alternative for audio playback of TV content at home. The IRT conducted a series of tests to compare the audio quality of soundbars against established playback devices.

This paper presents the method as well as the results of the test. A more detailed description of the soundbar evaluation and the test results can be found in [2].

Multiple stimulus ideal profile method

The MS-IPM is designed to evaluate various systems without an explicit reference. It provides measures of overall subjective quality, as well as characterizing the nature of the systems by using attributes.

The MS-IPM uses the multiple stimulus presentation approach to compare the sound systems under test similar to the MUSHRA [3] approach. The assessors are asked to provide their overall impression of the systems on a 100-point basic audio quality scale. A multiple stimulus comparison is also used for the rating of the attributes. Relevant attributes to describe the differences between the systems are selected by experts prior the test from establishes lexica. Additionally, the method seeks to establish how well the sound systems under evaluation compare to an envisaged ideal. For this purpose, the assessors are asked to rate the ideal level of each attribute, a hypothetical ideal system based on their wishes and experience. Depending on the nature of the systems under test and the attribute ratings, the ideal point may vary from the ratings of the systems. It should not be assumed to yield the same results as the preferred system.

The combination of the basic audio quality and attribute rating allows an in-depth analysis and interpretation of the quality of the systems under test.

Experimental setup

In order to gain experience of the MS-IPM an experiment was performed to evaluate the audio quality of soundbars compared to an ordinary TV setup and a 5.1 speaker system. The aim of this experiment was to study both the test protocol and the audio quality of the systems.

The purpose of the evaluation was to find out whether soundbars may be used as a good alternative for a 5.1 speaker system in a living room environment and whether they can improve the audio reproduction quality compared to a common TV set significantly. The differences of the playback devices were studied in two separated tests for stereo and 5.1 content. Original TV content from different genres (sport, documentation, TV-show, movie and music) was used to test the systems.

Eight soundbars with different audio reproduction technologies and from a wide price range were selected for the test using online rankings. To evaluate the audio quality of these soundbars they were compared to a common TV set with integrated speakers and a high quality 5.1 speaker system.

MS-IPM promised to be the right method for this evaluation because no real reference was given in the experiment. It would have been pure assumption to define the 5.1 speaker system as reference for the test as the performance of the soundbars could turn out to be better in the given circumstances. A standard paired comparison as described in ITU-R BS.1284-1 [4] could also have been a potential method to compare the systems but the required time effort would have been much higher than with MS-IPM. Furthermore, the rating of the basic audio quality and different attributes promised to provide a much more detailed description of the systems and their differences.

Test preparation

Program material

Typical German TV programs from different genres in stereo and 5.1 were selected for the test. TV content was used because the evaluation should show the differences between the systems for a typical use at home in a broadcast context. One short clip in both stereo and 5.1 was selected for each genre sport, documentation, TV-show, movie and music. The latter was represented as classic and pop music. Therefore, twelve samples with a duration from 10 to 21 seconds were selected in total.

Listening room

The listening room (Figure 1) where the tests have been performed was optimized with absorbers to resemble typical living room acoustics.

The 5.1 speaker system was positioned according to the requirements in ITU-R BS.775-3 [5]. The TV was set on top the shelf which was especially built for the evaluation. The shelf was designed to hold four soundbars at once in different positions around the center speaker of the 5.1 system. The position of the soundbars varied during the test for each assessor. Therefore, every soundbar was at every position at least once which was meant to minimize a possible position effect over all assessors. With the positioning of the soundbars and the absorbers on the walls the individual recommendations for all the systems under test were taken into account. During the test the shelf was concealed with acoustically transparent fabric, to avoid optical influences.

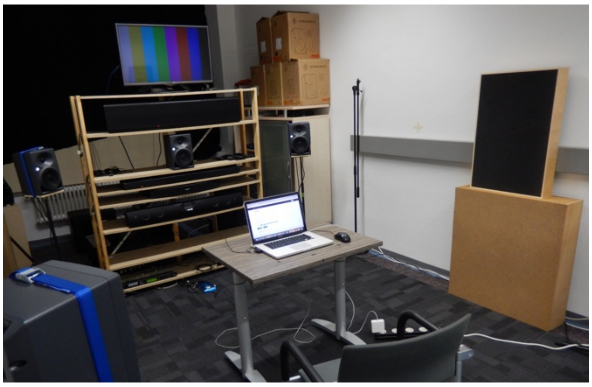


Figure 1: Listening room without concealing fabric

Test software

The test methodology was implemented in IRT’s own browser-based evaluation platform. The software allowed the assessors to switch in real-time between the systems under test, loop the test samples and set a range for the loop. The graphical user interface is shown in Figure 2 and 3.

Attribute selection

An approach described in [1| was used to select relevant attributes for the test. Four expert assessors familiarized themselves with the sound systems and the test samples in the listening room. They reviewed available attributes from a lexicon [6] and discussed the selection. They agreed on five attributes (see Table 1) which were considered to best describe the differences and characteristics of the systems under test. The attributes were translated in German and the descriptions were integrated in the test software.

Table 1: Attribute selection

Attribute	Description	Scale
Envelopment (for 5.1)	Are you surrounded by the reproduced sound and does it give a sense of space around you?	Not enveloping – Completely enveloping
Width (for stereo)	The width of the sound image (expressed as the perceived angle). - The width of the sound sources positions (soundscape width). The width of any reverberation should not be included in the assessment.	Narrow - Wide
Canny	The music sounds like it is being played in a can or tube. The sound is characterized by prominent and narrowband resonances in the midrange.	Not canny - Canny
Natural	Sounds reproduced with high fidelity. Acoustic instruments, voices and sounds, sounds like in reality. The sound is similar to the listener's expectation to the original sound without any timbral or spatial coloration or distortion, "Nothing added - nothing missing." The soundstage is clear in space and brings you close to the perceived original sound experience.	Unnatural - Natural
Detailed	A well-resolved sound rich in detail. Instruments, voices etc. can easily be separated. The music has many details, details that cannot be measured, details that give the music "soul". It may be small audible nuances: Breathing from a singer, fingers wandering across the guitar strings, the flaps from the clarinet, embouchure sound of the saxophone, the impact from the piano's hammers when they hit the strings.	Not detailed - Detailed
Bass strength	The relative level of bass, i.e. the low frequencies, for example male voices, bass guitar, bass drum, timpani and tuba. Should not be confused with bass depth that indicates the low frequency bass extension.	Soft - Loud

Experimental Procedure

The output level during the test was adjusted at 64 dB(A) for all assessors. The test was performed in German by 24 assessors with experience in listening tests. It was conducted in two sessions á 45 minutes. The assessors could stop at any point during the test and continue later.

The test was conducted in the following manner:

- Assessor instruction
- Basic audio quality familiarization
- Basic audio quality rating
- Ideal point and attribute familiarization
- Ideal point and attribute rating

For the first step, the assessors were provided with written and verbal instructions about the test in general and a detailed description of the task. In the second step, the assessors had time to listen to the test samples and familiarize themselves with the systems and the software for the basic audio quality (BAQ) rating (Figure 2).

The rating of the BAQ in step three was conducted for all systems under test in a multiple stimulus comparison. Each trial comprised one test sample. The order of the samples and the systems was randomized for each assessor.

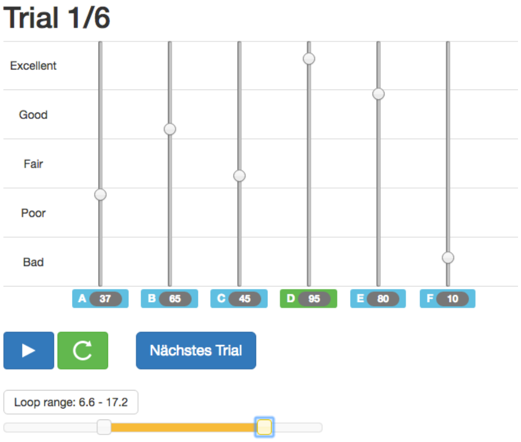
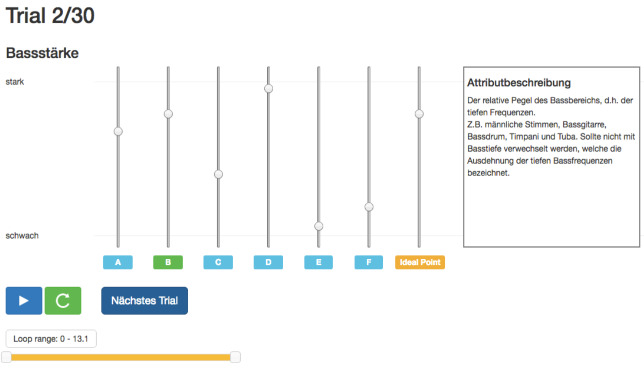


Figure 2: Basic audio quality test graphical user interface

After the BAQ rating the assessors had time to familiarize with the attributes and the ideal point rating (Figure 3). The slider for the ideal point was highlighted in yellow and a detailed description of the attribute under test was included in the software. For each attribute the order of the samples and the systems was randomized. The slider for each system and the ideal point had to be moved at least once to continue to the next trial. The assessors rated all six samples for one attribute before continuing with the next. In total 30 trials had to be completed for the attribute rating.



Trial 2/30

Bassstärke

Attributbeschreibung

Der relative Pegel des Bassbereichs, d.h. der tiefen Frequenzen.
Z.B. männliche Stimmen, Bassgitarre, Bassdrum, Timpani und Tuba. Sollte nicht mit Basstiefe verwechselt werden, welche die Ausdehnung der tiefen Bassfrequenzen bezeichnet.

Loop range: 0 - 13.1

Figure 3: Attribute test graphical user interface

Test results

A number of analyses were performed on the collected data. The Shapiro-Wilk test showed that the data was normally distributed. The applied ANOVA showed significant influence of the systems and no significant influence of the assessors. Moreover, the position of the soundbars within the shelf, which was changed for each assessor, had no influence on the ratings.

Basic audio quality

Figure 4 and 5 illustrate the average BAQ scores for 5.1 and stereo content averaged over all six samples and 24 assessors.

It can be noted that for both, 5.1 and stereo content, the TV set was rated significantly the lowest. The 5.1 speaker system achieved the best ratings but not with a significant difference to one of the soundbars (SB 2), which got the best scores of the soundbars, especially with 5.1 content. The ratings for the other seven soundbars are mainly located in the middle of the scale.

Over all the soundbars were rated slightly better for stereo content than for 5.1.

Attribute rating

In order to obtain a more detailed view on the data, the attribute and ideal point data was studied. The ideal point ratings for each attribute were averaged over all systems and assessors. This creates an ideal profile which illustrates an envisaged ideal system provided by the assessors.

The ratings for 5.1 content for all attributes and each system averaged over the 24 assessors and six samples are presented in combined spider plots in Figure 6.

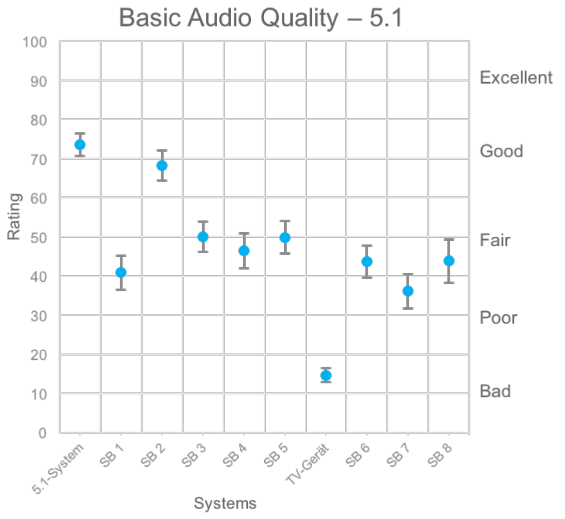


Figure 4: Basic audio quality ratings with 5.1 content, average over all 24 assessors and all samples

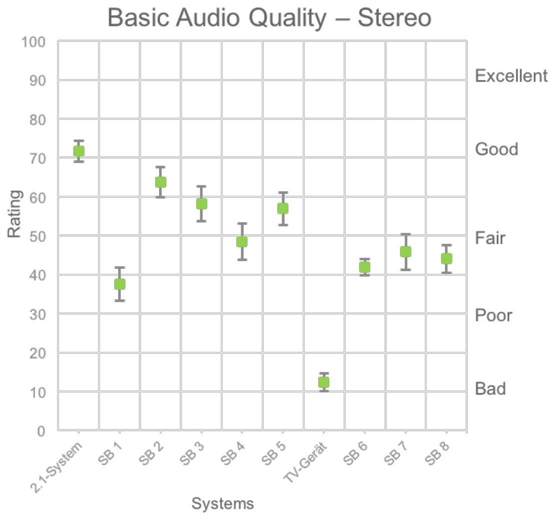


Figure 5: Basic audio quality ratings with stereo content, average over all 24 assessors and all samples

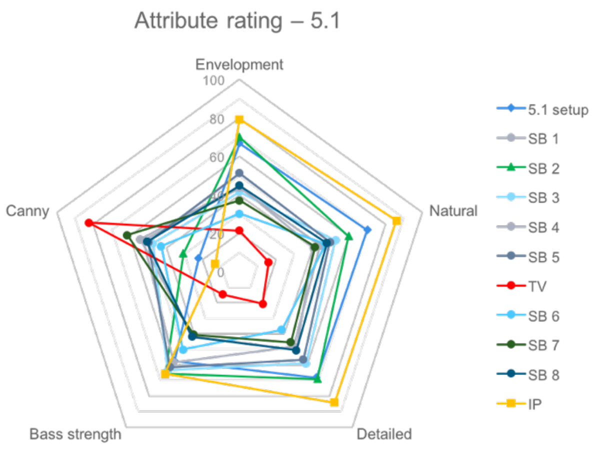


Figure 6: Combined spider plots of the attribute rating per system with 5.1 content, averaged over all 24 assessors and all samples

This data collection explains the performance of the systems better and in more detail. For example, the TV set is found to lack not only transparency characteristics, but there is also a lack of envelopment. Moreover, the system appears to be very canny with nearly no bass strength. The 5.1 system and SB 2 come closest to the ideal profile. This separates them from the rest of the soundbars for most of the attributes, whilst only for bass strength more of the sounbars seem to reach the ideal point.

Figure 7 and 8 show detailed results of the attribute ratings for envelopment and bass strength.

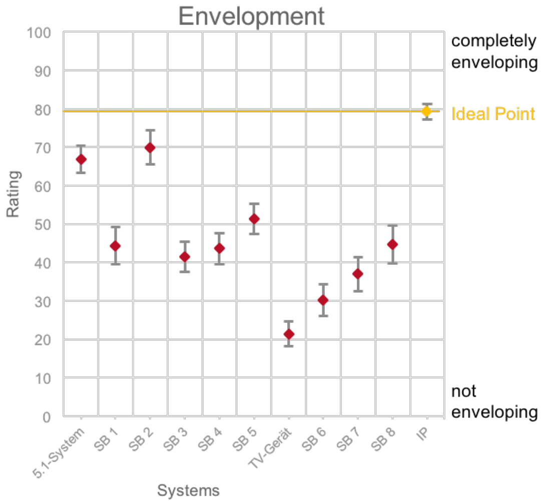


Figure 7: Attribute and Ideal Point ratings for envelopment with 5.1 content, average over all assessors and samples

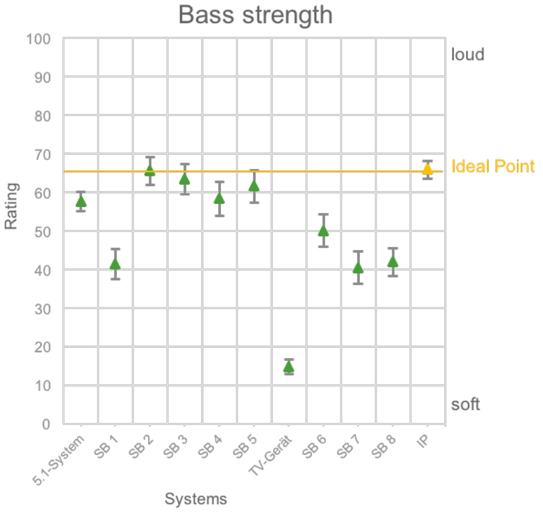


Figure 8: Attribute and Ideal Point ratings for bass strength with 5.1 content, average over all assessors and samples

Figure 7 illustrates the attribute and ideal point ratings for envelopment. Clearly none of the systems reaches the ideal point, highlighted with the yellow line at 79 points. The 5.1 system and SB 2 come close, but all the other systems are far from the ideal provided by the assessors. There is even one soundbar which didn’t perform significantly better than the TV system, which has been rated with the lowest average.

Figure 8 shows the attribute and ideal point ratings for bass strength. The average ideal point for this attribute is lower than for envelopment at 66 points on the rating scale. This shows that for the assessors an ideal system in context of this test has high envelopment but only medium bass strength. Four of the eight soundbars reach the ideal point for this attribute. An extraordinary result of this evaluation was the rating of the 5.1 system which did not reach the ideal point for the attribute bass strength.

A detailed analysis of all attributes as well as an additional principle components analysis (PCA) and preference maps can be found in [2].

Conclusion

The test results show that soundbars can improve the audio quality of an ordinary TV set in a living room environment significantly for typical TV content. For both, 5.1 and stereo content, the basic audio quality as well as the attribute ratings show significant better results for most of the soundbars. Some of the soundbars could even be an alternative for a high level 5.1 speaker system in this environment. Over all, the soundbars showed slightly better ratings for stereo content compared to 5.1 samples.

The MS-IPM proved to be the right choice for the given experiment. The attribute and ideal point ratings provided a better understanding of the quality of the systems under test and the assessor’s expectations in the context of the test. The absence of a reference was a challenge for the assessors but gave better insight in the relations of the systems among themselves and to the assessor’s expectations.

The selection of relevant attributes is a key step to getting meaningful results. A lot of consideration should be invested in this step. Furthermore, the familiarization of the assessors is very important especially if they are not yet familiar with this kind of test.

Literatur

[1] Zacharov, N., Pike, C., Melchior, F., and Worch, T., “Next Generation Audio System Assessment using the Multiple Stimulus Ideal Profile Method”. Proceedings of QoMEX 2016, Lisbon, Portugal, 2016

[2] Wakan, S., “Untersuchungen zur Wiedergabequalität von Soundbars mithilfe der Evaluationsmethodik MS-IPM“. Bachelor Thesis, Hochschule der Medien Stuttgart, 2017

[3] Recommendation ITU-R BS.1534-3, “Method for the subjective assessment of intermediate quality levels of coding systems”. ITU, Geneva, Switzerland, 2015

[4] Recommendation ITU-R BS.1284-1, “General methods for the subjective assessment of sound quality”. ITU, Geneva, Switzerland, 2003

[5] Recommendation ITU-R BS.775-3, “Multichannel stereophonic sound system with and without accompanying picture”. ITU, Geneva, Switzerland, 2012

[6] Pedersen, T. H., and Zacharov, N., “The Development of a Sound Wheel for Reproduced Sound.” In 138th Convention of the Audio Engineering Society, 2015