Subjective Testing of Car Audio Systems With and Without Parallel Task

Magnus Schäfer[1], Jan Holub[2], Jan Reimes[1], and Tomáš Drábek[2]

[1]HEAD acoustics GmbH, 52134 Herzogenrath, Germany, E-Mail: telecom@head-acoustics.de
[2]FEE CTU Prague, 166 27 Prague 6, Czech Republic, E-Mail: holubjan@fel.cvut.cz

Abstract

This contribution presents an auditory evaluation of car audio systems, providing a comparison between listening tests conducted in different simulated listening situations: listening only in a listening laboratory or a static car and listening while driving a (simulated) car. In all tests, identical signals and identical playback configurations were used leaving the test environment as the only variable in the test. The comparative evaluation of the listening tests shows good agreement between the listening situations for many stimuli. There is one clear trend that can be observed, though: The lower end of the quality scale is not used as frequently when performing a parallel task. This could indicate that a car driver is a less critical listener than a passenger.

Introduction

The sound quality of a car audio system is an important aspect for many consumers when making their buying decisions. Consequently, it is also of significant interest for car manufacturers. There are some studies addressing the testing paradigms that can be utilized for an efficient and realistic evaluation of the perceived sound quality. However, there are no investigations that focus on the differences between only listening and listening while performing a realistic parallel task, i.e. in this case, driving a car. The auditory data that can be acquired by a realistic listening test can then, e.g., be utilized for developing instrumental quality assessment approaches [1].

A previous contribution [2] already investigated the difference between conducting the listening test in a listening lab and in a stationary car (labeled sound car in the following). The main outcome of that investigation was that there is no systematic shift in the results of the test and that, furthermore, there is not even any significant difference between the results – not just when looking at large averages but already for the individual votes.

In both previously investigated test environments, the test subjects could fully focus on listening to the audio signals. The actual use case of car audio systems usually does not allow for this behaviour as driving the car is more important than listening to music.

Sound quality testing

There is no shortage when it comes to specifications for conducting listening tests. Several ITU recommendations (e.g., [3, 4, 5]) describe test methodologies for different testing needs. All of these tests require relaxed and focused naïve test subjects. As mentioned before, this can usually be achieved for the listening test – but the realistic usage scenario for the tested systems or devices can be very different.

A current ETSI work item (Procedures for Multimedia Transmission Quality Testing with Parallel Task Including Subjective Testing) addresses this gap in the range of available listening tests. Within this work item, a methodological description of how (and for which cases) to conduct listening tests with the parallel task approach is being developed. The investigation described in this contribution directly integrates into the work item, as it allows to cleanly identify the impact of an appropriate parallel task on the outcome of an auditory evaluation.

Subjective testing under parallel task conditions

One common feature of the aforementioned, currently standardized quality assessment methods is the requirement to perform all experiments in defined and precisely maintained laboratory environment. Technical parameters as the acoustic background noise level, room reverberation time, headphone type and calibration etc. are reported and compared with standardized required limits. The test subjects are comfortably seated and are required to fully attend to the quality test procedure.

Such a rigid laboratory environment is far from a realistic situation where the devices under test (e.g., telecommunication systems or car audio systems) are typically used. Thus, the question of the practical value of regular quality measurement procedures arises. A parallel task to distract test subjects from full concentration to the quality listening test testing is therefore introduced to bring the test procedures (and results) closer to real situations.

There are three fundamental classes of parallel tasks that can be applied depending on the situation:

- **Physical**
  Purely physical activities requiring little to no mental effort, e.g., running or cycling.
- **Mental**
  Purely mental activities requiring little to no physical effort, e.g, arithmetic operations or memorizing digits and words.

- **Hybrid**
  A mix of both areas, activities that require both mental and physical effort, e.g., driving a car.

For the assessment of car audio systems, it is fairly straightforward to choose driving a (simulated) car as the parallel task. However, it is worth noting that it is not necessary to have the parallel task mimic the actual usage scenario of the device under test. The parallel task should be chosen from a class that fits the real-world usage of the device: E.g, when testing small, portable audio players, the parallel task should be mostly physical. When testing stationary hands-free telephones, a mental parallel task is more fitting.

Test design

Three auditory tests have been designed and conducted using an identical set of stimuli. The listening database contains 161 samples of excerpts (length approximately 8 s) from binaural audio recordings of seven different music signals. The devices under test are the audio systems of different cars. A description of the measurement setup and the signals can be found in [2].

The first test run comprised the tests in the listening laboratory and the sound car, it followed the standardized P.800 [3] methodology and was presented and analyzed in [2]. The main focus of this contribution is on the second test run which utilized a parallel task – simulated car driving. All tests were performed in critical listening environment conforming to ITU-T P.800 [3]. Open headphones (Sennheiser HD650) with diffuse field equalization and professional distribution amplifier were used in all environments for playing the stimuli. A professional voting software was used for collecting the subjective votes. In all tests, a rating scale with opinion scores ranging from 1 (Quality is very bad) to 5 (Quality is excellent) with nine possible ratings in steps of 0.5 on the opinion score scale was used.

The first experiments were conducted with 45 listeners split into two groups of 23 test subjects for the listening lab and 22 test subjects for the sound car.

The group for the parallel task experiment consisted of 32 listeners (17 female and 15 male test subjects, average age 28.7 years, minimum 18, maximum 65). Additional information regarding the listening experience and the preferences of the test subjects have been acquired by questionnaires filled by them after the test was finished. The parallel task used a PC-based car driving simulator.

Test Difficulty

As one aspect of the aforementioned questionnaires, the test subjects were asked to assess the rating difficulty on a scale of 1 (not difficult at all) to 5 (very difficult). The answers to this question are depicted in Figure 1 and given as percentages of the entire group of test subjects in each test environment.

It can be observed that the listening lab seems to be the easiest environment while the driving simulator is most

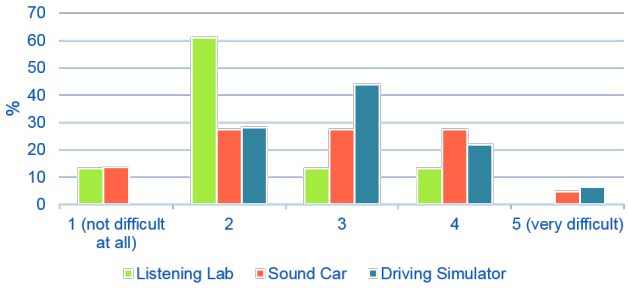

Figure 1: Perceived difficulty of the listening test in different environments

difficult. Since listening lab and sound car are very similar in their demands on the test subjects, it was expected that the answers to this question should be fairly similar as well. Surprisingly though, the answers for the sound car are much closer to the answers from the driving simulator. The reason for this is not entirely clear. The ergonomics in the sound car were slightly worse than in the listening lab as the touch screen for rating the stimuli was not directly in front of the test subjects due to the position of the steering wheel. As this arrangement was far from uncomfortable, though, it is most probably not the only reason for this increase in perceived difficulty.

Test Results

To get a first overview of the results of the perceptual evaluation, the average mean opinion score (MOS) values for the three test environments are given in Figure 2. The black vertical lines on each bar represent the corresponding confidence interval.

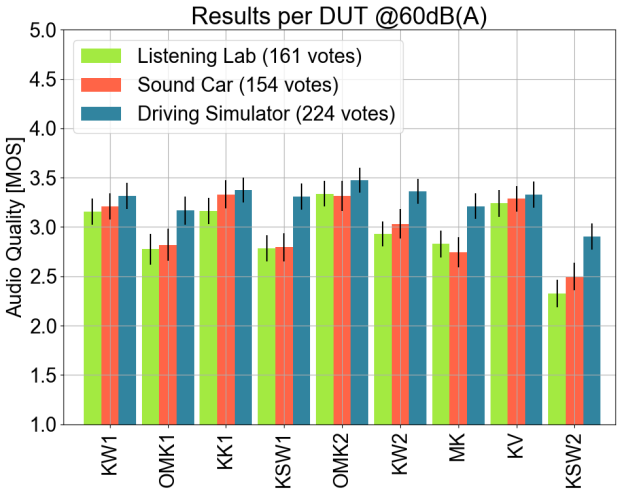

Figure 2: Average MOS results of the listening test in different environments

In general, the averaged values for the driving simulator are higher than for the listening lab and the sound car. Two groups can be distinguished in the results: For the better systems (KW1, KK1, OMK2 and KV), the results are fairly similar for all three test environments. For the other systems (OMK1, KSW1, KW2, MK and KSW2),

though, the values in the driving simulator are clearly higher. The detailed reasons for this are not visible in this representation due to the strong averaging.

A closer look at the values for the individual stimuli is given in Figure 3. The scatter plot compares the results in the sound car with the results in the driving simulator. The results for the listening lab are omitted here as they are very similar to the results for the sound car (cf. [2]).

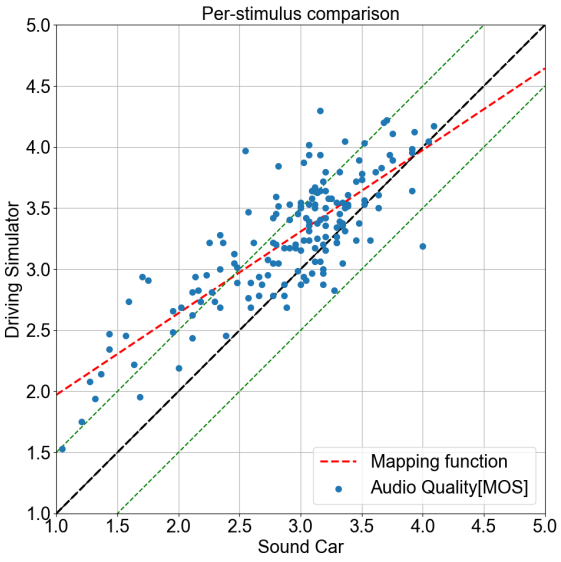

Figure 3: Scatter plot of averaged results per stimulus for sound car and driving simulator

The mean relation between the results in the sound car and the driving simulator can be expressed by a linear mapping.

$$\mathrm{MOS}_{\mathrm{DrivingSimulator}} = 0.67 \cdot \mathrm{MOS}_{\mathrm{SoundCar}} + 1.3 \qquad (1)$$

This mapping function is depicted by the red dashed line in Figure 3. Applying this mapping function to the MOS values from the sound car leads to the relation that is shown in Figure 4.

The results from both environments are very similar after mapping and there are only very few outliers where the results differ by more than 0.5 MOS points. Since there are no stimuli in the area above 4.3 MOS (where the mapping curve is below the main diagonal), the mapping mostly addresses the voting behaviour in the lower quality regions. The test subjects were fairly reluctant to use the lower end of the rating scale in the driving simulator – on average, a stimulus that was rated as "bad" in the sound car was only rated as "poor" in the driving simulator.

The same observation is possible when analyzing per test condition, this includes some reference and anchor conditions along with the previously shown devices under test. In analogy to Figure 3, the results for each test condition are shown in Figure 5 along with the corresponding mapping curve.

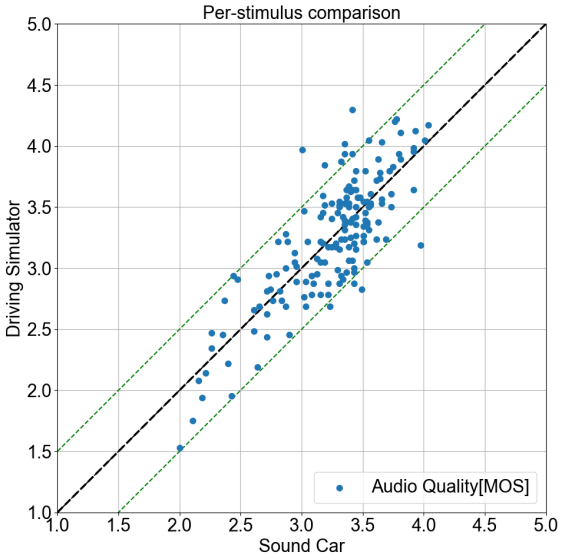

Figure 4: Scatter plot of averaged results per stimulus for sound car and driving simulator after mapping

Due to the averaging, the optimum mapping curve for these values is similar to but slightly less steep than the curve in Figure 3. The results after applying the mapping are shown in Figure 6.

The results per test condition after mapping are very similar for the two environments. This seems to indicate that the difference between the testing environments can be completely explained by the linear mapping which approximately compresses the results for the sound car from the original rating scale between one and five to the range between two and five as only this area was used in the driving simulator.

Discussion

Similar studies on the effect of a parallel task in different domains are available in [6, 7, 8, 9]. In [9], an interesting concept is proposed, providing a possible explanation for differences in the results of an assessment task. A human mind concept of two characters is introduced, called System 1 and System 2. System 1 includes involuntary and effortless mental operations (e.g., counting or simple arithmetic operations) while System 2 is used for reasoning and logical thinking, not requiring quick (intuitive) solutions. In some cases, these two systems provide different results, contradicting each other.

Laboratory testing without parallel is a typical System 2 situation, where subjects are fully focused on the quality assessment task, analyzing the stimulus as thoroughly as possible. The parallel task arrangement may force the test subject to rely on System 1 for quality tests as their System 2 capacity is occupied by the parallel task activity. This would explain different results of certain test conditions in a parallel task situation.

In this investigation, the results with and without the parallel task do not differ dramatically. There is clear

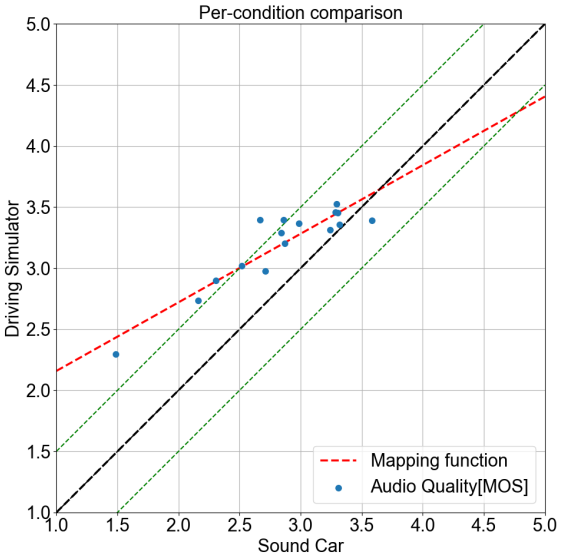

Figure 5: Scatter plot of averaged results per test condition for sound car and driving simulator

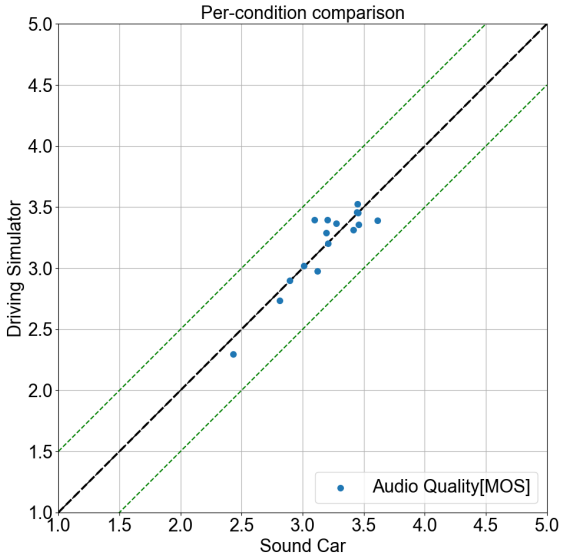

Figure 6: Scatter plot of averaged results per test condition for sound car and driving simulator after mapping

reluctance of the test subjects to use the lower end of the rating scale with the parallel task – however, a simple linear mapping can explain most of the differences between the results.

Conclusions

An identical auditory evaluation was carried out in three different environments: a listening laboratory, a sound car and a driving simulator. While the evaluation in listening laboratory and sound car was conducted without a parallel task, the test subjects were also tasked with car driving while assessing the audio quality in the driving simulator.

While the results are very similar between the listening laboratory and the sound car, a systematic shift can be observed in the results for the driving simulator. The test subjects were not as critical towards the lower quality devices in that environment. A linear mapping could be derived for this situation that compensates for the differences between the results – in particular for the comparison of devices.

The outcome of this investigation is two-fold: The actual auditory results can be utilized for modeling human quality perception of audio systems. The main point in this contribution is on the methodological side, though. It was shown that the auditory evaluation of car audio systems with parallel task gives comparable results to an auditory evaluation without parallel task. A possible interpretation of the systematic shift in the results is that the evaluation without parallel task allows for a more precise analysis, while the results with parallel task are closer to the quality perception in the realistic usage scenario of the audio system.

References

[1] Magnus Schäfer. An approach for instrumental quality evaluation of car audio systems. In *Fortschritte der Akustik - DAGA 2017*. DEGA e.V., Berlin, 2017.

[2] Jan Reimes, André Fiebig, Thomas Deutsch, and Michael Oehler. Comparison of Auditory Testing Environments for Car Audio Systems. In *Fortschritte der Akustik - DAGA 2017*. DEGA e.V., Berlin, 2017.

[3] ITU-T Recommendation P.800. *Methods for subjective determination of transmission quality*, Aug. 1996.

[4] ITU-T Recommendation P.835. *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, Nov. 2003.

[5] ITU-R Recommendation BS.1534. *Method for the subjective assessment of intermediate quality levels of coding systems*, October 2015.

[6] David Navon and Daniel Gopher. On the economy of the human-processing system. *Psychological review*, 86(3):214, 1979.

[7] Christopher D Wickens. Processing resources and attention. *Multiple-task performance*, 1991:3–34, 1991.

[8] Sian L Beilock, Thomas H Carr, Clare MacMahon, and Janet L Starkes. When paying attention becomes counterproductive: impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology: Applied*, 8(1):6, 2002.

[9] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.