Comparison of Auditory Testing Environments for Car Audio Systems

Jan Reimes[1], Thomas Deutsch[2], André Fiebig[1], Michael Oehler[3]
[1] *HEAD acoustics GmbH, 52134 Herzogenrath, Germany*
[2] *Institute of Sound and Vibration Engineering (ISAVE), University of Applied Sciences, 40476 Düsseldorf, Germany*
[3] *Osnabrück University, 49074 Osnabrück, Germany*

Introduction

For most car drivers and passengers today, the enjoyment of music is an inherent part of their driving experience. In consequence, car audio systems are getting more important from a commercial point of view. Quality investigations of such systems are commonly based on single component or loudspeaker tests. However, the final judgment should always be the sound quality of the complete audio system as perceived by the user.

This contribution presents a new framework for auditory tests of car audio systems, which is based on binaurally recorded music tracks. Several cars of different manufacturers including a wide range of audio systems were involved in this evaluation. Since the audience should not be influenced by the interior of a certain car cabin, equalized headphone playback in a defined environment is used.

Listening tests can be conducted in a realistic environment, i.e., a driving simulator. While this is recommended from a psychological point of view, the alternative use of a listening studio makes the test process faster and less cumbersome as multiple test subjects can participate simultaneously. To investigate the impact of the testing environment, the test corpus is evaluated in a car cabin as well as in a listening studio.

Auditory Evaluation

A new study of auditory evaluation of car audio systems is introduced. The concept is based on real music playback and diffuse-field-equalized (DF) binaural recordings of head and torso simulator (HATS) according to [1] and/or [2]. Figure 1 exemplarily illustrates the setup inside a real car cabin. Measurements were conducted using nine common cars and their original audio system as a device-under-test (DUT) in silent condition (engine turned off). If applicable, all devices were set to default equalizer settings, possibly available advanced signal processing was turned off.

Real and professionally mixed music tracks are used as test stimuli. In a preliminary study, the songs were selected according to current and popular broadcast playlists, containing styles like rock, classic, jazz, r&b, classic and pop music. Playback inside the car was realized via CD player, which was a common playback method for all audio systems. Overall, seven tracks per DUT were recorded. For the auditory presentation of the stimuli, representative excerpts of 8.0 s - 12.0 s are selected from the source material.

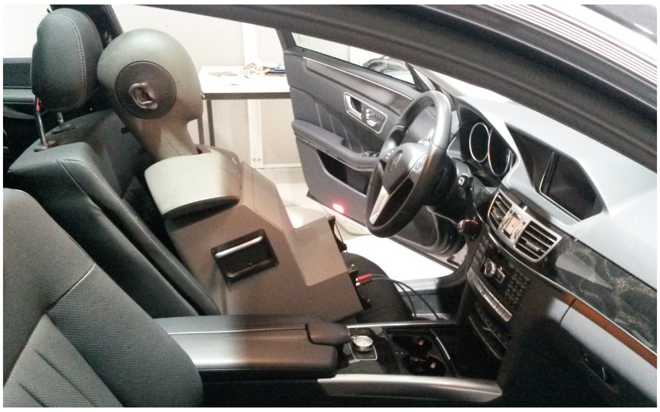


Figure 1: Example of mounting HATS inside car cabin

Two volume settings were evaluated for each DUT. To obtain comparable levels for each audio system, pink noise was played back via CD player and volume control was tuned until a level of 60 dB(A) was approximately achieved. This volume setting was assumed as a standard listening level. Additionally, the same procedure was conducted for 70 dB(A). This corresponds to a much louder playback and could possibly cause more non-linear distortions of the loudspeakers.

Several auditory test methods and scales of different standards are available for testing audio quality in general, e.g. [3], [4] or [5]. Due to the focus of this evaluation on a large amount of test conditions, a 5-point absolute category rating (ACR) test design according to ITU-T P.800 [4] was chosen. Here the stimulus is absolutely judged according to a given attribute list and not e.g., against an explicit reference. Each stimulus is evaluated for the perceived *sound quality*, on a scale from "bad" to "excellent".

Within the pre-processing of the stimuli, all recorded signals are normalized in loudness to 23.0 sone (N5 percentile acc. to ISO 532-1 [6]). With this level alignment, the two volume settings 60 dB(A) and 70 dB(A) seem to be redundant on the first sight. But the intention of this calibration is that only audible differences resulting from e.g. loudspeaker distortions should be included, but without the impact of the increased level itself.

Even though ACR-based tests do not provide an explicit reference signal for comparison, several best possible stimuli should always be included, i.e., where perceived quality is judged with maximum score. For the assessment of such a high quality recording, a binaural measurement with a stereo-triangle and high quality playback equipment in an anechoic chamber was arranged. All songs were played back via this setup, which is illustrated by figure 2.

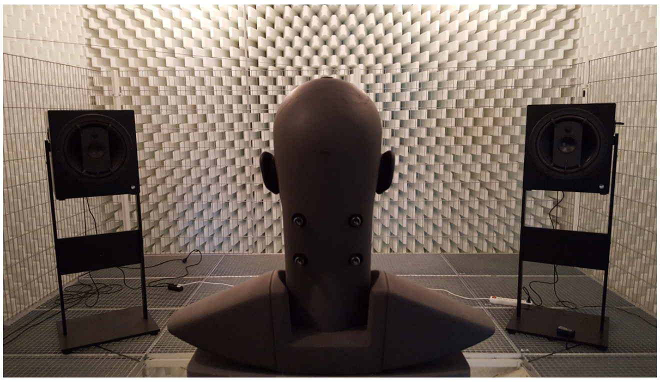


Figure 2: High quality recording in anechoic room

On the other hand, also anchors to the lower end of the scale are recommended. A procedure is, e.g., proposed in [3] which uses two band-limited versions (7.0 kHz and 3.5 kHz) of a best-case reference signal to anchor mid and low quality. These additional two anchors were created based on the high quality reference signals for all songs.

Including all anchor, test and other signals, the test corpus consists of 161 samples in total. The presentation of all stimuli in the auditory evaluation is conducted by hearing-adequate playback of the binaural DF-equalized recordings.

Even though the considered setup may not be 100% realistic due to special systems (like e.g., subwoofer mounted below the seat), this drawback is regarded as an acceptable trade-off. For a more realistic approach, auditory testing of audio systems could also be conducted in the original cars. However, beside the enormous additional effort, test subjects could unnecessarily be influenced by effects like e.g., the interior, the brand or in general any subjective preferences. Thus a common, neutral environment for all analyzed DUTs is strongly desired.

Listening Test Environments

So far, the test design and procedure follows "traditional" guidelines. Usually the auditory evaluation itself is conducted in listening laboratories. However, several psychological aspects are not considered yet. For example, it remains unclear, if test subjects may behave differently when listening to real car audio systems in a real cabin. Furthermore, even the quality perception may not be the same here as in a listening laboratory. Possible differences between such listening test environments are regarded in the following sections.

Laboratory

Figure 3 exemplarily illustrates the impression of the room and the test procedure in a listening laboratory. Such a room usually provides good acoustic parameters (idle noise lower than 20 dB(A)). Depending on the listening test software, also multiple terminals can be used simultaneously. Several ambience parameters like air ventilation, lighting, seating can be controlled by the investigator and the room can be seen as a very "clean" or even environment.

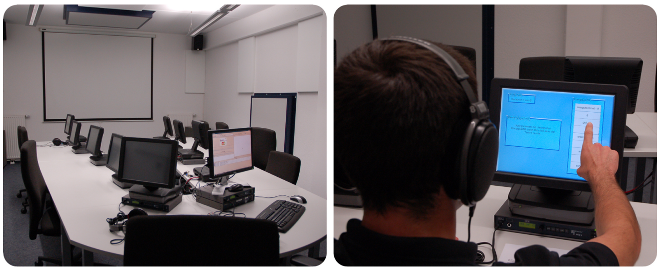


Figure 3: Example of listening laboratory

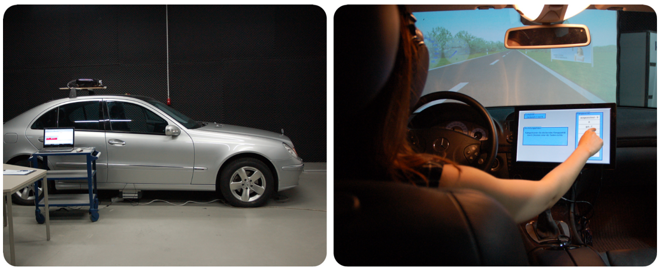


Figure 4: Example of driving simulator

However, it is debatable if such a room is too unrealistic or too artificial for the evaluation of car audio systems. In case of a hearing-adequate playback via headphone, quality perception may already differ here.

Driving simulator

In contrast to the previously described listening laboratory, a more realistic environment can be evaluated. Figure 4 shows an example of a driving simulator which is equipped for auditory testing similar to a listening laboratory. Obviously, this setup represents a much more "lively" ambience. Additionally, a street is projected on a screen in front of the car. Even interactive elements or simulation of a drive could be introduced. However, in the current work, only the static projection without any further interaction was chosen.

In this environment, good acoustic parameters are provided (idle noise lower than 20 dB(A)) as well. On the one hand, this setup introduced several drawbacks compared to the listening laboratory. Only one test person can be evaluated at once and a driving simulator may not always be accessible for auditory testing. On the other hand, this environment obviously seems much more realistic for evaluation of car audio systems.

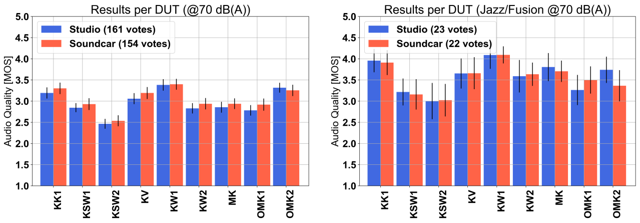
Results of auditory evaluation

In order to investigate the impact of the previously presented listening test environments, the introduced test corpus is evaluated once in the laboratory and once in the cabin of the driving simulator. Altogether, 45 different test subjects consisting of experts, experienced and naive listeners participated in both tests (22 in driving simulator, 23 in listening laboratory). Each participant listened to all 161 stimuli of the corpus, including a longer pause after 80 samples.

General observations

Both conducted tests provide a lot of results. Since this work focuses on the difference between the test environments, only a subset of the large amount of results is presented here. The most obvious analysis is the evaluation per DUT. Figure 5a provides the auditory results for this purpose, averaged over all included songs.

A first observation can be derived from this data, i.e. showing a quite low difference between minimum and maximum obtained results is quite low (about 1.0 MOS between KW1 and KSW2). When expanding the results to the per-song ratings, the song *Jazz/Fusion* shows higher scores for most DUTs. Figure 5b provides the results for this stimuli for all devices. It can be seen that all scores increase compared to the averaged results, but the rank order of the DUTs is approximately maintained.



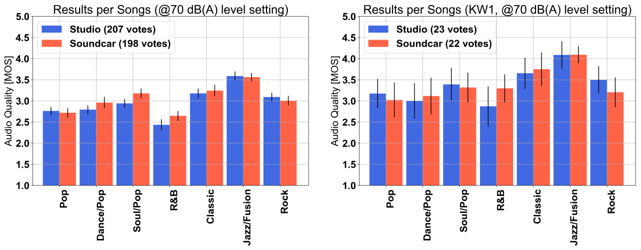
(a) per DUT **(b)** only song *Jazz*

Figure 5: Per DUT results at volume setting 70 dB(A)

To investigate the suitability of certain songs over all DUTs, the averaged per-song results are provided in 6a for the volume setting of 70 dB(A). If all songs would be judged around the same average value, all songs would provide a good representation of the individual DUT. Obviously this is not the case, the difference between worst- and best-rated song is again about 1.0 MOS. The song of type *R&B* with very low frequency content causes stronger audible distortions in some DUTs/loudspeakers, thus the lower rating is not surprising. On the other hand, the source of *Jazz/Fusion* is based on a high-quality music production and includes a wide stereo image.

To investigate if this possibly is resulting from any arbitrary per-song judges, the results are then separated for the best-rated DUT KW1, as shown in figure 6b. In general the scores of all songs improve compared to the evaluation over all DUTs (see figure 5a), again *Jazz/Fusion* obtains the best result. It seems like that this song seems more appropriate for the auditory evaluation than others - possibly due to the aforementioned spatial aspects. Also here, the magnitude between best and worst song amounts only to about 1.0 MOS.

Usually anchor signals in ACR-based auditory evaluations are used to provide several known quality degrees, including the best-case reference. As described in the previous section, a best-case reference signal was recorded with high-quality equipment in order to provide a signal with "ideal / perfect" quality. It is expected that most participants would judge these special anchor signals with best category ("excellent", MOS = 5.0)



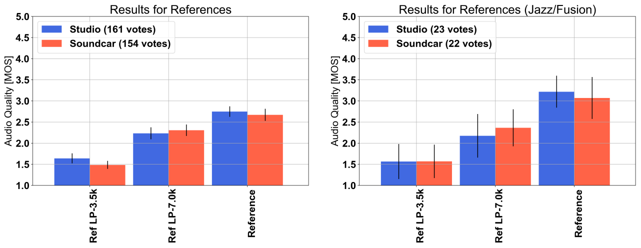
(a) per Song **(b)** only DUT KW1

Figure 6: Per song results at volume setting 70 dB(A)

Figure 7a depicts the averaged results for the three anchor conditions. While the lower (3.5 kHz) and medium (7.0 kHz) anchor conditions obtain values in the expected range, the unfiltered anchor signals are not even close to provide best-possible quality scores. Only MOS ≈ 2.7 is reached for these conditions.

The source of this odd behavior is still under investigation. Even though the high quality recordings inside the anechoic room were carefully conducted, expert listeners confirmed at least some of the lower scores. On the other hand, the different acoustic situation could be another reason for the mismatch, because it does not match the listener's expectation of a car-cabin.

Again, to investigate the impact of the different songs, the best-rated song for these conditions is analyzed in figure 7b. Similar as in the previous results, the track *Jazz/Fusion* performs best here, but still only a maximum value of MOS ≈ 3.2 is obtained.



(a) All songs **(b)** only song *Jazz*

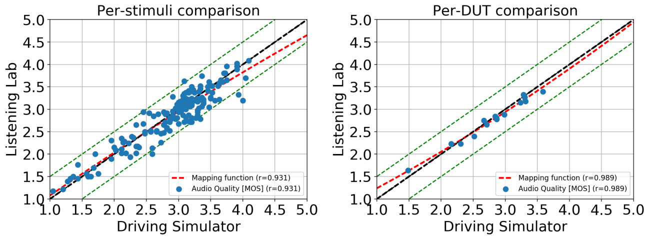
Figure 7: Anchor results

Comparison of environments

So far, several averaged results per DUT or song were introduced. In all bar plots, the difference between both test environments is shown. But this does not provide a clear impression of significant different judgments. Thus, an overall evaluation and comparison is provided in figure 8.

Figure 8a illustrates the comparison by a scatter plot, comparing the averaged 161 per-stimuli results. A third-order mapping function is also provided to indicate that there is no systematic shift between both environments, the observable differences seem to be random with respect to the number of votes per stimulus (22/23).

To base the comparison on a larger group, in figure 8b the results of DUTs and anchor signals are aggregated over all songs to a per-condition score. Obviously, the differences get even more negligible. Table 1 also provides root-mean-square errors for both scatter plots, which also emphasize the high correlation.



(a) Per stimuli

(b) Per DUT

Figure 8: Comparison of listening test environment

To check the statistical significance of these observed differences, a two-sample t-test is conducted between the results of the listening laboratory and the driving simulator. The p-values for different kinds of comparisons are summarized in table 1, which do not indicate any significant differences for the evaluation per DUT or per stimuli. Only when looking into all $161 \cdot 23 = 3703$ (listening laboratory) respectively $161 \cdot 22 = 3542$ (driving simulator) single votes of each auditory test as whole, slight significant differences can be observed ($p = 4.8\% <$ 5% level of significance).

Aggregat	p-value	RMSE	$\lvert d \rvert$
DUT	0.84	0.103	–
Stimuli	0.48	0.236	–
Votes	0.046*	–	0.06

Table 1: Error metrics and significance of differences

In order to analyze the effect size of the rather weak (but existing) significant difference for the votes, the parameter $|d|$ according to [7] and [8] is provided by equations 1 to 3. The indices LL refer to the votes collected in the listening laboratory, whereas DS denotes the driving simulator judgments. $n_{LL/DS}$ is used for the number of votes for the corresponding environment. Similarly, $\overline{\mathrm{MOS}}_{LL/DS}$ is the average of all votes in each test.

$$|d| = \frac{\overline{\mathrm{MOS}}_{LL} - \overline{\mathrm{MOS}}_{DS}}{\sigma} \qquad (1)$$

$$\sigma = \sqrt{\frac{(n_{LL}-1)s_{LL}^2 + (n_{DS}-1)s_{DS}^2}{n_{LL}+n_{DS}-2}} \qquad (2)$$

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\mathrm{MOS}_{j,i} - \overline{\mathrm{MOS}}_i)^2. \qquad (3)$$

The significant differences found by the t-test conducted with all votes are put into perspective by the very low effect size of $|d| = 0.06$ (0.20 would indicate at least a small relevance according to [7]).

Conclusions

A setup of auditory testing for car audio systems based on binaural recordings and hearing-adequate playback was introduced. Nine commercially available cars and their audio playback systems were evaluated as DUTs. A large auditory evaluation was conducted with 161 stimuli in two test environments with 45 participants.

As a first noticeable result, it was observed that the complete quality range was not fully covered. Still there is ongoing work on the reasons for these characteristics. The chosen ACR-test procedure or the too low best-case anchor signals could be discussed as reasons. Another explanation could be that a comparison against explicit reference may be more suitable for this application. However, future work should also consider the approach of a ”optimum quality reference”, which is necessary e.g., in comparison rating tests.

Additionally, the auditory evaluation of a single overall score may not be sufficient to address the problem in its entirety. Audio quality impression is based on multiple dimensions which should be assessed on distinct scales. Attributes like, e.g., reproduction of spatial properties or linear/non-linear distortions could be considered as possible candidates for such a multidimensional approach.

Even though the assessed auditory data did not distribute equally over the quality range, results between listening environments provide a high correlation. At least in this study, no systematic differences can be observed which was proved by several significance metrics. In consequence and for practical reasons, further auditory tests targeting at a comparable setup can be conducted in a listening lab without distorting the results.

References

[1] *Artificial ears*, ITU-T Recommendation P.57, Dec. 2011.

[2] *Use of head and torso simulator for hands-free and handset terminal testing*, ITU-T Recommendation P.581, Feb. 2014.

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

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

[5] *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*, ITU-T Recommendation P.913, Mar. 2016.

[6] ISO/FDIS 532-1, *Acoustics – Methods for calculating loudness – Part 1: Zwicker method*, International Organization for Standardization, 2016.

[7] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, 1988.

[8] J. Hartung, G. Knapp, and B. Sinha, *Statistical Meta-Analysis with Applications*, ser. Wiley Series in Probability and Statistics. Wiley, 2008.