Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Automotive Audio
2022 June 8–10, Dearborn, MI, USA

Automotive Audio System Evaluation over Headphones Based on the BVIRs of Different Listening Positions: A Case Study of a Specific Audio System

Yukun Pei[1], You Li[1 and] Pablo Ripollés[2,3,4],

[1] *New York University Steinhardt, 35 West 4th Street, New York, NY 10012*

[2] *Music and Audio Research Laboratory (MARL), New York University, 370 Jay Street, New York, NY, 11201*

[3] *Center for Language, Music and Emotion, New York University, 6 Washington Place, New York, NY, 10003*

[4] *Department of Psychology, New York University, 6 Washington Place, New York, NY, 10003*

Correspondence should be addressed to Yukun Pei (yp635@nyu.edu)

ABSTRACT

In recent years, car manufacturers have consistently upgraded the audio systems of their vehicles, with audio aficionados adding further modifications to them. The acoustics of the vehicle cabin and the sound effects of the audio systems have become one of the most important topics of the Research and Development Departments of manufacturers. For example, the selected vehicle in this experiment has implemented different spatial audio algorithms in its models to achieve a better listening experience. By capitalizing on impulse responses, binaural audio technology provides the opportunity and the flexibility to virtually generate the sound effects of a particular space without the requirement of physically being in that space. We used binaural technology and the vehicle to develop a standardized procedure for the evaluation of car audio systems. A perceptual listening test was integrated into this study to verify the procedure and to further evaluate this specific audio system.

1 Introduction and Background

People have enjoyed pre-recorded music through reproduction systems ever since the phonograph was invented. In recent years, with the development of recording technology, the mono playback system has been developed further into stereo and multichannel systems to suit different needs. In addition, digital processing technology has improved the sound quality and the listening experience. While the industry has been focusing on developing products like personal headphones and home audio systems for consumers for decades, the investment for a better audio system on automotive vehicles has mostly been limited to high-end vehicle models and audiophiles. This creates an opportunity for automotive manufacturers to stand out among competitors by improving their audio systems. In addition to digital signal processing algorithms that read and reproduce different audio formats, the acoustics of the vehicle cabin need to also be considered, as for example, listeners in different seating positions can have different sound perceptions.

Binaural audio technology is widely used to examine spatial environments and has the advantage of virtually generating sound effects in a specific space through the impulse response of headphones *without the need to be in that particular space*. Previous

research has used binaural technology and binaural vehicle impulse responses (BVIRs) to simulate the acoustics of an automotive audio reproduction system in the laboratory, including testing BVIRs at different head angles and under different specific noises (e.g., engine, road, wind, etc.) [1,2,3]. In this work, we capitalize on BVIRs to reproduce the interior sound environment of a specific audio automotive system under different *seating positions and different digital processing algorithms*.

2 Acquiring Impulse Responses

The audio system of the selected vehicle allows for different surround audio effects: Off, Standard, and High. This is to create a surround effect even when the audio is not in surround format [4]. The tested vehicle was not modified, and the BVIRs recording was conducted with the vehicle parked in a garage with all the electronics off, the windows and doors closed, and with no other item nor person in the cabin.

We chose a logarithmically time-varying sinusoid sweep as the stimuli for acquiring the impulse responses, as it is more robust to artifacts when conditions are not ideal [5]. Five-second-long unity-gained sine sweeps were generated for the left channel only stereo, right channel only stereo, and in mono, and exported as Wave files in 44.1k Hz sample rate and 16-bit depth (this is the audio spec that most free streaming services use).

A Neumann KU 100 Dummy Head Binaural Microphone was placed at the height of average sitting adults to record the sine sweep at the investigated seating position (see Fig. 1). The position of the Dummy Head was fixed facing straight forward, and no head rotation was taken into consideration. Since the vehicle can only play external audio via Bluetooth from a cell phone or a USB drive that is plugged into the vehicle, the three generated sine sweeps were imported into a USB drive to be played sequentially and automatically by the vehicle audio system. The three sine sweeps were played and recorded individually under the three different spatial audio algorithms (Off, Standard and High). The recording session was on a setup with a 44.1k Hz sampling rate and 16-bit depth, the same as the sine sweeps. Since a delay before the audio was played was required to allow enough time to get out of the vehicle, five seconds of silent audio were inserted before each sine sweep. However, since the input and the output of the audio system do not share the same interface, we manually added a "pop" sound after the five-second silence. This sound was played just before the sine sweep, so that the beginnings of the sine sweeps could be later aligned. Furthermore, the end of the sine sweep recording was located 6 seconds after the start of the sine sweep, to make sure it captured all the reverb tails. We manually marked the -60dB position as the end of the recordings. We then cropped the recording according to the start and end points. Finally, we performed a fast deconvolution on the cropped recordings to obtain the impulse responses (see Fig. 1).

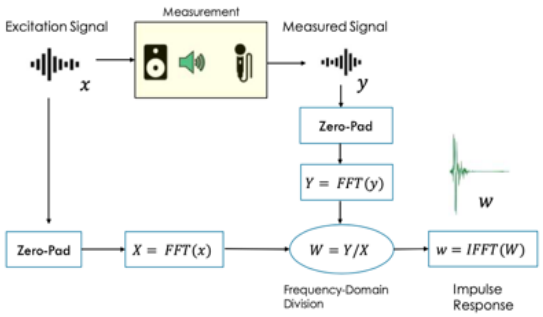

Figure 1. Fast Deconvolution Flow Diagram.

The impulse responses were further processed to ensure data quality and to tackle two main issues: a low amplitude response and a delay in the acquired impulse responses.

The dynamic range of 16-bit audio is roughly 96dB. When the peak of the impulse response is lower than -36dB, we cannot use the RT60 to judge the end of the reverberation or directly use the lower bound -96dB as the noise floor. Then, to tackle this, we globally normalized the resulting impulse response, setting the highest peak among all tracks to 0dB, and normalizing the other tracks proportionally. With this normalization, we can have a better view of the noise floor (which is amplified as well) while not changing the reproduction results (convolution is linear-shift invariant) and preserving the amplitude ratio within all tracks. After that, we calculated the noise floor of

each channel by taking the mean of the RMS value of a silent period at the end of the track.

However, the shift caused by fast deconvolution can lead to uncertain starting points of the acquired impulse responses. Therefore, it is crucial to find not only when the reverb ends, but also when the reverb starts. The approach we took started with the calculation of the RMS of the entire track for each channel. From the first window to the last, when the RMS of a certain window was greater than the noise floor +6dB, we considered the attack kick-in, and we took the starting point of the previous window as the starting point of the impulse response. Similarly, when the RMS of a window was less than the noise floor +6dB (indicating that the end of the reverberation time was approaching), we took the end point of the next window as the end point of the impulse response.

Speaker position\Seats Measurement	Distance (inches)	Distance (cm)	Time (ms at 9°C)
Front Seat Position 3 (same)	20	50.8	1.509
Front Seat Position 3 (opposite)	27	68.58	2.037
Front Seat Position 9 (same)	20	50.8	1.509
Front Seat Position 9 (opposite)	23	58.42	1.735
Back Seat Position 8 (same)	27	68.58	2.037
Back Seat Position 8 (opposite)	27	68.58	2.037
Back Seat Position 9 (same)	35	88.9	2.64
Back Seat Position 9 (opposite)	36	91.44	2.715
Mid Seat Position 8 (same)	28	71.12	2.112
Mid Seat Position 8 (opposite)	32	81.28	2.414
Mid Seat Position 9 (same)	41	104.14	3.093
Mid Seat Position 9 (opposite)	43.5	110.49	3.281

Table 1. Speakers Placement Measurement

Finally, to ensure that the binaural playback system can reflect the time difference of different channels as much as possible (as shown in Table 1) and reduce the impact on the Interaural Time Difference (ITD), we decided to choose a global starting point for the impulse response. We chose the earliest one among all channels and ensured that all channels also used this starting point. Furthermore, to prevent the "pop" sound or other artifacts from appearing, we introduced a half-window-length linear fade-in and fade-out at the beginning and end of the impulse responses.

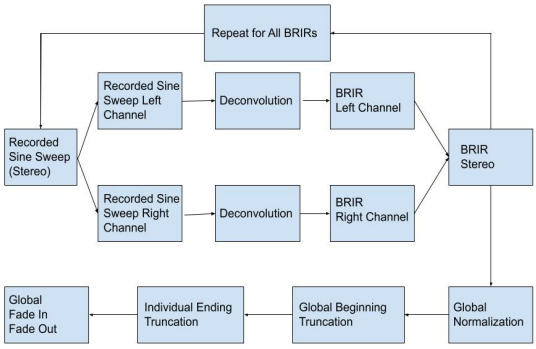

Figure 2. BVIR Signal Processing Flow Diagram.

After all the processing (see Fig. 2), all BVIRs were ready for convolution, with each audio file beginning at when the earliest impulse response started with a fade-in to the on-set that was 6dB louder than the noise floor and ending at its own noise floor fade-out.

3 Test Stimuli

The song clips to be convolved with the BVIRs were chosen from different five genres of music (Ambience, Classical, Jazz, Pop, and Solo) so that the effects of the algorithms to different genres could be investigated. The duration of the clips was based on the tempo of the selected songs and was chosen with two 8-bar measures so that the listeners had enough time to distinguish the differences between the tracks without introducing listening fatigue. Each of the clips was between 20 to 30 seconds long. The audio spec was the same as the previous audio files (44.1k Hz sampling rate and 16-bit depth).

To simulate the sound effects of the audio system and the vehicle cabin at each seating position, the obtained BVIRs were convolved with the selected tracks, using a self-developed Python script (see Fig. 3). Specifically, the left and the right channels of the stereo audio files were read into individual files (left

channel only stereo, right channel only stereo). Then, the convolution was done by individually convolving the left and right channels of the songs with the BVIRs obtained from the sine sweep recordings at each seating position and with each different algorithm. Then, the audio convolved with the left channel of the BVIRs was summed together, and the same was done for the right channel, finally obtaining the left channel and the right channel of the binaural audio. Lastly, all the convolved tracks were globally normalized to a standard dBFS level to eliminate the possibility of misleading the listeners with unbalanced levels and audio quality among tracks.

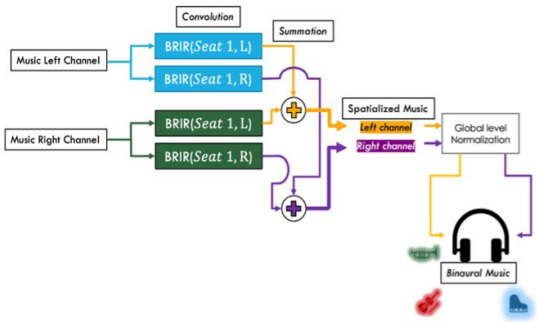

Figure 3. Binaural Audio Generation Flow Diagram.

4 Perceptual Listening Test

Three songs were selected for each genre. Each of the songs represented different artists either from different time periods or different production styles. Combining the different surround algorithms (Off, Standard and High) with the 5 different seating positions, 15 different versions of the same song were generated, for a total of 45 different tracks per genre. One perceptual online listening test per genre was developed using oTree, a Python-based framework for the development of controlled online behavioural experiments [6]. The online perceptual tests were presented to participants as an HTML webpage running in Google Chrome. Participants were first presented with a few lines summarizing the task and then with an informed consent page. Upon acceptance, detailed instructions for the test were presented. Participants were instructed to complete the listening test using their own reproduction systems, headphones (closed-back and high-resolution headphones were preferred) and in a quiet and controlled environment.

To ensure compliance with these requirements, participants first completed a headphone check test [7]. Then, the 45 different tracks of a particular genre were presented in randomized order. Participants were not given any information regarding the algorithm used or the specific seating position. After listening to each track, participants were asked to provide behavioural ratings using a 1-10 scale for several measures, including: pleasure, level of noise, loudness, clarity (muffled/piercing), immersiveness, tone (dark/bright), distance to the sound source, and "outside feeling" (how much did you feel the music coming from "outside" of your head/body?). Familiarity ratings were also collected as a control. These questions were selected to assess the subjective timbral attributes and spatial impression of the tested tracks. Common terminologies were used due to the diverse background of the participants. The experiment was coded so that participants could not advance to the next page unless the song had been played in full (participants had no control over the music player) or all answers had been provided.

Participants were recruited from United States using Amazon Mechanical Turk (AMT), a platform for the acquisition of large online behavioural datasets (participants were required to have 99% of previous submissions approved on AMT to ensure data quality). Participants provided demographic information (gender, age, education), and completed tests assessing their musical training (Gold-MSI; important to control for musical abilities) [8], and their sensitivity to musical reward (the Barcelona Music Reward Questionnaire, BMRQ; important to screen out participants who cannot experience emotion from music, that is, participants with specific musical anhedonia) [9]. The two questionnaires included an attentional checks (e.g., *Please, select the option "Agree"*). Participants were paid $15.

5 Statistical Analyses

Analyses were performed by implementing linear mixed modelling (LMM) in R (version 4.0.2) using

the lmer4 package [10]. In each analysis we first generated an empty model, which contained only a random intercept for participants. Next, we generated 12 different models (see Fig. 4) by adding the main conditions (musical genre, seating position in the car, surround algorithm) and their interactions, and the demographic variables (age, gender, education, musical training with the Gold-MSI and sensitivity to musical reward with the BMRQ).

1) empty model
2) model with only Genre variable and demographic data
3) only Position and demographic data
4) only Algorithm and demographic data
5) Genre, Position, and demographic data
6) Genre, Algorithm, and demographic data
7) Algorithm, Position, and demographic data
8) Genre, Algorithm, Position
9) Genre * Position and demographic data
10) Genre * Algorithm and demographic data
11) Algorithm * Position and demographic data
12) Genre * Algorithm * Position and demographic data

Figure 4. LMMs. * Reflects an interaction.

For each behavioural rating, we selected the best model explaining the variance in the data using the Akaike information criterion (AIC). We considered a model different from another if the difference in AIC was greater than 2, to balance complexity and goodness of fit. If models were separated by less than two AIC, we selected the model with fewer factors, as this explains the same amount of variance in the data using fewer variables. The effects of the different predictors were assessed using Type III Wald Chi-Square tests with post-hoc contrasts being calculated using the emmeans package [11] with Tukey correction for multiple comparisons.

6 Results

Data from a total of 90 participants (22 per genre) was collected. Twenty-three participants were excluded for: 1) scoring less than 50% in the headphone check (33% chance level); 2) failing the attention checks of the questionnaires; 3) answering that the model of headphones used was “loudspeakers” (note that the huge variability in the type of headphones precludes a more in-depth analysis that fully explores the effect of this variable); and 4) if they had musical anhedonia (scoring less than 63 in the BMRQ) [9]. The final sample was composed of 67 participants (22 women, age=38.6 ±10.2 years).

Regarding pleasure scores, the model with the best fit was the one containing algorithm and the demographic factors. There was a trend (p=0.06) for the choice of algorithm to affect the pleasure reported by the participants, regardless of the position of the car or the genre. The comparison between the different surround algorithms shows that the pleasure reported for music treated with the Standard algorithm is marginally higher than when processed with the High one (p=0.055; see Fig. 5).

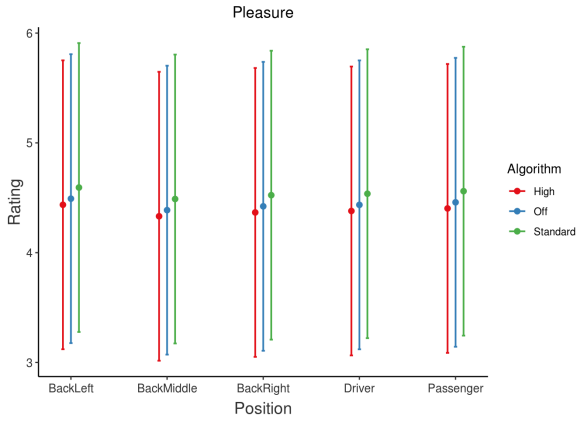

Figure 5. Behavioral results for pleasure with 95% confidence intervals. The difference between Standard and High algorithms trends towards significance, regardless of genre or seat (p=0.055).

For immersiveness, the best model contained genre, seating position, and the demographic factors (that is, the choice of surround algorithm did not have a particular effect). Genre significantly predicted immersiveness (p=0.043), while the seating position trended towards significance (p=0.055). Comparing the different levels within these two factors showed that Jazz was the genre inducing more immersiveness (marginally more immersive than the Solo genre, p=0.066) and that the Back Middle seat induced the most immersive experience (significantly more immersive than the Driver seat, p=0.025; see Fig. 6). For the seating position, it is interesting to note that the Back Middle seat is the one having the most open space within the car, with this possibly affecting the

immersive experience. Lastly, the lack of an effect of surround algorithm in the reported immersiveness is surprising and might indicate that either our processing pipeline did not successfully reconstruct the effects of each algorithm, or that the different algorithms do not significantly increase the subjective level of immersiveness perceived by the listeners.

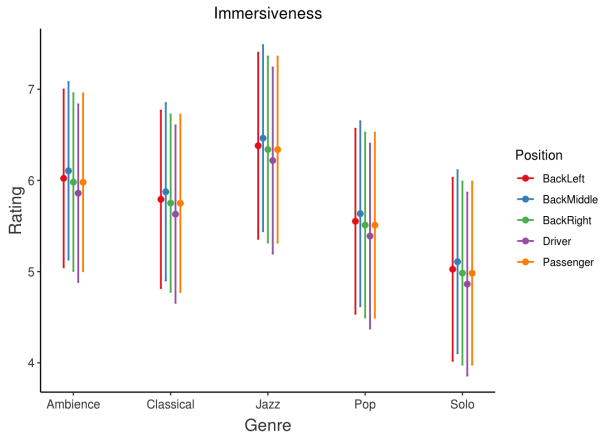

Figure 6. Behavioral results for immersiveness with 95% confidence intervals. Jazz was marginally more immersive than Solo (p=0.066). The Back Middle seat was significantly more immersive than the Driver seat (p=0.025) regardless of algorithm or genre.

Regarding “outside feeling” (i.e., How much did you feel the music coming from "outside" of your head/body?), the model with the best fit was the one that included genre as a main factor and the demographic variables (that is, neither algorithm not seating position had an effect). Genre was significantly (p=0.007) predictive of the reported “outside feeling”, with Jazz receiving the highest scores, which were significantly different than those reported for Classical music (p=0.011; see Fig. 7).

The model with the best fit to the loudness scores was the one containing seating position as a main factor and the demographic parameters, regardless of the choice of algorithm or the genre. As with immersiveness, the seating position was predictive of the loudness reported (p=0.002; see Fig. 8), with the Back Middle seat having the highest rating, significantly higher than the Driver (p<0.001) and the Back Left seats (p=0.043). The pattern is similar as that reported for immersiveness, which suggests that immersiveness and loudness might be related.

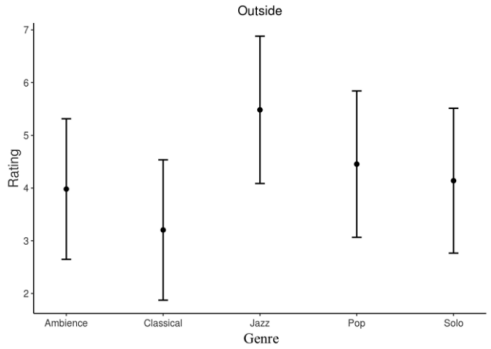

Figure 7. Behavioral results for “outside feeling” with 95% confidence intervals. The difference between Jazz and Classical is significant (p=0.011), regardless of algorithm or seating position.

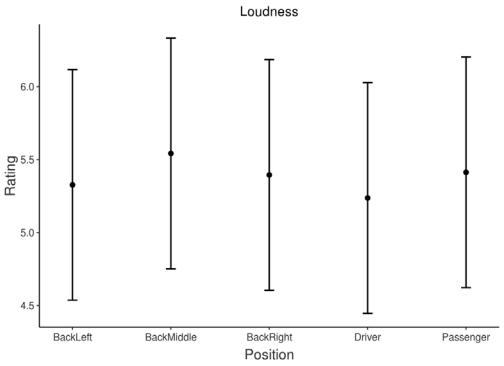

Figure 8. Behavioral results for loudness with 95% confidence intervals. The difference between the Back Middle and the Driver (p<0.001) and the Back Left seats is significant (p=0.043), regardless of genre and algorithm.

The results for clarity also mirror those reported for loudness and immersiveness, with the best fit being the model containing seating position and the demographic factors. There was a significant effect of seating position (p=0.02; see Fig. 9), with the Back Middle seat having the most piercing sound, significantly more piercing than the Driver seat (p=0.043). These results show that the Back-Middle seat induced the highest ratings of immersiveness, loudness, and clarity, further suggesting that that the more symmetrically opened space around the middle

seat allows the listener to perceive the sound wave more evenly and effectively.

For tone, the model with the best fit was the one containing genre and the demographic factors (that is, there were no effects of the surround algorithm or the seating position). Indeed, genre significantly predicted tone (p=0.0015; see Fig. 10), with Jazz having the brightest tone, significantly brighter than Classical (p=0.015) and Pop music (p=0.036). This result suggests that the only factor that affected the tone of the stimuli was the tone of the original songs themselves, regardless of the surround algorithm used or the seating position.

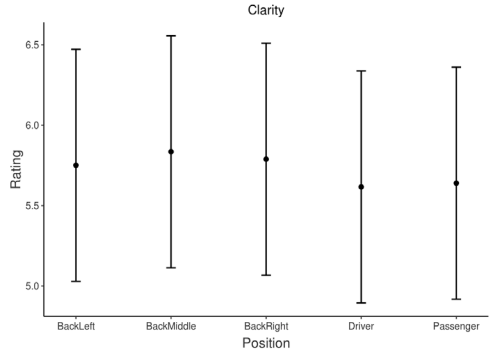

Figure 9. Behavioral results for clarity with 95% confidence intervals. The difference between the Back Middle seat and the Driver seat is significant (p=0.043), regardless of algorithm and seat.

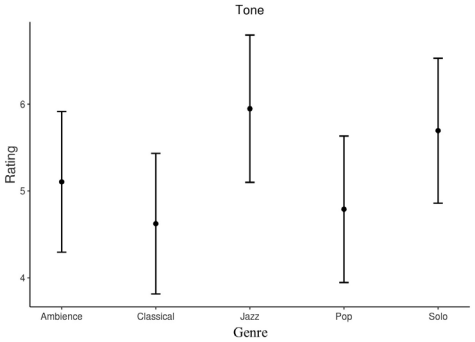

Figure 10. Behavioral results for tone with 95% confidence intervals. The difference between Jazz and Classical (p=0.015) and Pop music is significant (p=0.036), regardless of seat and algorithm.

Interestingly, none of the main factors (seating position, genre, and algorithm) significantly predicted the reported distance to the source, the perceived noise, or the familiarity with the songs. Regarding noise, since the sine sweeps were recorded under the same environment with minimum fluctuation, the noise should be relatively constant, and no factors should affect it. This result further suggests that our proposed pipeline successfully recreated the environment of where the sine sweeps were recorded, and that it did not add any artificial noise effects. The fact that there was no familiarity effect suggests that participants' previous knowledge of the music stimuli did not affect the results.

Finally, we built a LMM to predict pleasure using all other behavioural scores provided by the participants, regardless of the main factors (seating position, genre, and surround algorithm). The model included the demographic variables and subject as a random intercept. While participants liked a song more if it was more familiar, less noisy, less loud, and clearer (all $p<0.001$), the best predictor of pleasure was the reported immersiveness ($p<0.001$; immersiveness had the largest estimate: 0.41). This result highlights the crucial role of immersiveness in enhancing enjoyment for music listening inside of a car.

7 Conclusions

The purpose of this study was to capitalize on BVIRs to reproduce the interior sound environment of a specific automotive system and to assess its acoustics under different seating positions and digital processing algorithms. The spatial algorithms (Off, Standard, High) in this vehicle offer the passengers the ability to choose what kind of immersive effect they want to experience. However, the results from the perceptual listening test did not show any significant effect of algorithm in the immersiveness reported by the participants. The *pleasure* reported by the participants was slightly affected by the surround algorithm, although, surprisingly, the High surround algorithm was the one that elicited the lowest pleasure scores. This might have been caused by: 1) our signal processing procedure; 2) sound coloration from the recording equipment; 3) the collected data from the

listening test; 4) the reproduction systems themselves; 5) factors of noise.

Regarding the seat location, different seat positions had significant effects on the reported immersiveness, loudness, and clarity, with the Back Middle seat providing the best acoustic experience. Second, musical genre had a significant effect on immersiveness, “outside feeling” and tone, with Jazz being reported as the top genre for these scores. Third, distance to the sound sources, familiarity and perceived noise were not affected by any predicting factors. Lastly, participants reported more pleasure for tracks which sounded more immersive. The developed procedure and results support the use of binaural processing technology for the assessment of spatial audio, especially when the objective is to recreate the acoustics of real-life environments so that an end user does not have to physically be in a particular space to experience the sound effects of that space.

Our study suffers from several caveats. First, the reported results could be further confirmed by performing a listening test inside of the car (instead than in front of a computer while using headphones) that includes head-tracking. Second, while the postprocessing of the recordings minimizes the risk for inconsistencies in the synchrony of the recordings, our capture method might introduce small inaccuracies by not accounting for the clock drift between reproduction and capture systems and for system non-linearities. Third, there might be a mismatch between the objective of the digital algorithms provided by this vehicle audio system and the questions of the listening test (e.g., "outside feeling" might not necessarily be one of the design targets of the surround algorithms).

Despite these caveats, we suggest that the reported processing pipeline, along with the proposed perceptual listening test, can assist researchers to further exam any vehicle audio system. With more customizations and improvements, the proposed procedure can provide endless possibilities for the audio and automotive industry.

References

[1] F. Christensen, M. Lydolf, G. Martin, P. Minnaar, B. Pedersen, & W.K. Song, “A listening test system for automotive audio-Part 1: System description,” *Audio Engineering Society Convention 118*, (2005).

[2] P. Hegarty, S. Choisel, and S. Bech, "A Listening Test System for Automotive Audio – Part 3: Comparison of Attribute Ratings Made in a Vehicle with Those Made Using an Auralization System," *Audio Engineering Society Convention*, (2007).

[3] S. Choisel, P. Hegarty, F. Christensen, B. Pedersen, W. Ellermeier, J. Ghani, and W. Song, "A Listening Test System for Automotive Audio - Part 4: Comparison of Attribute Ratings Made by Expert and Non-Expert Listeners," *Audio Engineering Society Convention*, (2007).

[4] M. R. Bai & C. C. Lee, “Comparative study of design and implementation strategies of automotive virtual surround audio systems,” *Journal of the Audio Engineering Society* vol. 58, no. 3, pp. 141—159 (2010).

[5] A. Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” *Audio engineering society convention 108*, (2000).

[6] D. L. Chen, M. Schonger, C. Wickens, “An open-source platform for laboratory, online, and field experiments,” *Journal of Behavioral and Experimental Finance* vol. 9, pp. 88—97 (2016).

[7] K.J.P. Woods, M.H. Siegel, J. Traer & J.H. McDermott. Headphone screening to facilitate web-based auditory experiments. *Atten Percept Psychophys*. Oct;79(7):2064-2072 (2017).

[8] D. Müllensiefen, B. Gingras, J. Musil, & L. Stewart, “Measuring the facets of musicality: The Goldsmiths Musical Sophistication Index (Gold-MSI),” *Personality and Individual Differences* vol 60, no. S35 (2014).

[9] E. Mas-Herrero, J. Marco-Pallares, U. Lorenzo-Seva, R. J. Zatorre, & A. Rodriguez-Fornells “Barcelona Music Reward Questionnaire,” *Music Perception* (2013).

[10] D. Bates, M. Mächler, B. Bolker, S. Walker “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, 67(1), 1–48 (2015).

[11] R. Lenth “emmeans: Estimated Marginal Means, aka Least-Squares Means”. *R package version 1.7.3,* <https://CRAN.R-project.org/package=emmeans>