D. Koya, R. Mason, M. Dewhirst and S. Bech,
"A Perceptual Model of Spatial Quality for Automotive Audio Systems"
J. Audio Eng. Soc., vol. 71, no. 10, pp. 689–706, (2023 October).
DOI: https://doi.org/10.17743/jaes.2022.0079

A Perceptual Model of Spatial Quality for Automotive Audio Systems

DAISUKE KOYA,[1] **RUSSELL MASON,**[1] *AES Member*, **MARTIN DEWHIRST,**[2] **AND SØREN BECH**[3]
(dkoya@mac.com) (r.mason@surrey.ac.uk) (r.mason@surrey.ac.uk) (sbe@bang-olufsen.dk)

[1] *University of Surrey, Guildford Surrey GU2 7XH, United Kingdom*
[2] *Focusrite Audio Engineering Ltd., Artisan Hillbottom Road, High Wycombe Buckinghamshire HP12 4HJ, United Kingdom*
[3] *Bang & Olufsen, Peter Bangs Vej 15 Struer, 7600, Denmark*

A perceptual model was developed to evaluate the spatial quality of automotive audio systems by adapting the Quality Evaluation of Spatial Transmission and Reproduction by an Artificial Listener (QESTRAL) model of spatial quality developed for domestic audio systems. The QESTRAL model was modified to use a combination of existing and newly created metrics, based on—in order of importance—the interaural cross-correlation, reproduced source angle, scene width, level, entropy, and spectral roll-off. The resulting model predicts the overall spatial quality of two-channel and five-channel automotive audio systems with a cross-validation R^2 of 0.85 and root-mean-square error (RMSE) of 11.03%. The performance of the modified model improved considerably for automotive applications compared with that of the original model, which had a prediction R^2 of 0.72 and RMSE of 29.39%. Modifying the model for automotive audio systems did not invalidate its use for domestic audio systems, which were predicted with an R^2 of 0.77 and RMSE of 11.90%.

0 INTRODUCTION

The acoustic environment of automotive audio systems presents a challenge for the ideal reproduction of spatial audio. This is due to many factors including the small volume of the automobile cabin, the combination of both highly reflective and highly absorptive surfaces, and multiple transducers being located throughout the cabin and auditioned from offset seating positions. This acoustic environment results in perceived spatial degradations such as a lack of spaciousness, widened auditory source widths, and skewed localization [1–3]. These spatial degradations make designing and optimizing automotive audio systems challenging.

Assessing perceived spatial quality involves comparing to a known reference changes in spatial characteristics such as spaciousness, auditory source widths, and localization. Comparing the perceived spatial quality of many automotive audio systems using listening tests takes much time and effort, and statistical analyses which follow the listening tests also consume much time and resources. Effective perceptual models are a beneficial alternative because they can predict the perceived quality of stimuli quickly and reliably within the scope of the target application [4, pp. 11–13], saving time and effort. Perceptual models have shown their utility in predicting loudness [5], perceptual audio codec quality [6], and speech quality [7]. A perceptual model for automotive audio systems could reduce time and effort compared to formal listening tests while retaining similar reliability, either as a replacement for or an aid to listening tests, which would be conducive to the short development cycles demanded by the automotive industry.

A literature review revealed that the overall sound quality of automotive audio systems has been modeled using metrics related to perceived timbral, spatial, distortion, and speech quality [8]. However, a perceptual model that specifically predicts the overall spatial quality of automotive audio systems had not been developed. The Quality Evaluation of Spatial Transmission and Reproduction by an Artificial Listener (QESTRAL) model already existed [9–12], which predicts the overall spatial quality of consumer multichannel audio systems in domestic environments. Hence, this model was taken as a starting point to widen its applicability for the automotive audio environment.

Fig. 1 shows the general procedure of developing a model for predicting perceptual audio quality [9]. SEC. 1 deals with the lower path of Fig. 1, which describes the design of a listening test for collecting overall spatial quality ratings of automotive audio systems, and the results of this. SEC. 2 deals with comparing the listening test results to predictions by the QESTRAL model, to evaluate the effectiveness of the original model. SEC. 3 deals with the final stage of Fig. 1, which presents how listening test results and metrics were combined to calibrate a regression model specifically for

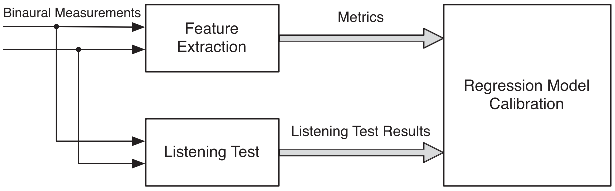


Fig. 1. General procedure of developing a model for predicting perceptual audio quality (adapted from Rumsey et al. [9]).

automotive audio. SEC. 4 deals with assessing the performance of the calibrated regression model, and SEC. 5 deals with assessing whether the calibrated regression model is still valid for domestic audio systems.

1 LISTENING TEST

The calibration of a perceptual model that predicts the overall spatial quality of automotive audio systems requires the ratings of this attribute from a listening test. The design and results of this listening test are described in this section.

1.1 Setup

To reliably compare different automotive audio systems, a headphone-based auralization system was employed. Using an auralization system allows the listening tests to be administered blind (which minimizes biases associated with the appearance of the reproduction system [13]), and allows rapid switching between stimuli (to avoid the problems of auditory memory retention, which decays after around 1 s [14, 15]). Head tracking was incorporated to improve localization by enabling the perceived sound images to remain stationary regardless of head rotation [16]. Previous experiments have shown that headphone auralization with head tracking does not result in substantial differences compared to experiments using in-situ automotive audio environments [17–19].

The auralization system employed binaural room impulse responses (BRIRs) of domestic and automotive audio environments measured with a Brüel & Kjær Head and Torso Simulator 4100 between ±30° in 1° increments. The BRIRs were convolved with program items to synthesize the stimuli. The auralization hardware included Sennheiser HD 650 headphones, an Xsens head tracker, an RME Hammerfall digital signal processing (DSP) sound card, and a Dell Dimension E520 PC with a solid-state drive running Windows XP Professional. A headphone filter—which compensates for the transfer function between the headphone transducer and the blocked ear canal [20]—was applied to the headphones. A MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) interface was created in MATLAB, and the convolution and head tracking were handled by AM3D Convolution Box.

1.2 Stimuli

The test stimuli were composed of BRIRs convolved with program items. A reference system based on a 3/2 stereo system [21] housed in an ITU-R BS.1116 compliant listening room [22] specified a defined level of overall spatial quality to which the stimuli were compared. The "3" in "3/2" refers to the three front channels—left (L), right (R), and center (C)—and the "2" refers to the two surround channels—left surround (LS) and right surround (RS). The reference system was chosen because it was used in the development of the QESTRAL model.

Table 1. Automotive audio systems for the listening test. For the Experimental System, "Tuned" refers to all DSP configurations enabled, whereas "Untuned" refers to all DSP configurations bypassed.

		DSP Configurations							
		Tuned			Intermediate Tune				
		One Seat	Front Seats	Rear Seats	Frequency Equalization	Level Alignment	Time Alignment	Untuned	OEM Systems
Automobiles	Audi A8		X						
				X					
	Experimental System	X							
					Bypassed	X	X		
					X	X	Bypassed		
					X	Bypassed	X		
					X	Bypassed	Bypassed		
								X	
	Audi A6								X
	VW Golf								X

OEM = Original Equipment Manufacturer.

Table 2. SAPs for the listening test and their mean ratings from [23].

SAP	Mean Rating
3/2-Channel to 3/1-Channel Downmix	96
3/2-Channel to 2/0-Channel Downmix	74
3/2-Channel to 1/0-Channel Downmix on All Channels	40
3/2-Channel to 1/0-Channel Downmix	16
3/2-Channel to 1/0-Channel Downmix Combined with 500-Hz High-Pass Filter on All Channels	10

There were three categories of BRIRs: automotive audio systems, spatial audio processes (SAPs), and hidden anchors. Ten two-channel and five-channel automotive audio systems—which were composed of different automobiles or DSP configurations—were employed to investigate their degree of differentiation in the presence of the SAPs and hidden anchors (Table 1).

The automotive audio systems were broadly categorized as either commercially available or experimental, where the former includes an Audi A8, Audi A6, and VW Golf, and where the latter includes a setup which allowed DSP combinations of frequency equalization, level alignment, and time alignment. The Audi A8 featured soundfields that were optimized for either the front or rear seats. The experimental system represents the stages of tuning automotive audio systems in the real world, beginning with an untuned configuration, followed by various configurations of intermediate tune, and ending with a tuned configuration. Although the Audi A8 and Audi A6 were from the same manufacturer, the former featured waveguide-loaded tweeters located on the dashboard near the base of the A-pillars, while the latter featured conventional tweeters located in the upper-door area. The Audi A6 and VW Golf were two-channel systems, and the Audi A8 and experimental system were five-channel systems. The automotive audio system BRIRs were measured for the driver's seat (i.e., front-left).

The SAPs were a subset of the spatial impairments employed to develop the QESTRAL model [23]. These spatial impairments included those commonly encountered in consumer multichannel audio systems, such as downmixing, altered loudspeaker locations, and interchannel level misalignment. Five SAPs with mean ratings that spanned the entire range of the assessment scale as evenly as possible—between 0 and 100—were employed. The SAPs were chosen because they have known scores that the listening test results can be compared to. These known scores are valid only in their original context (i.e., in the configuration of stimuli they were evaluated). Table 2 lists the SAPs and their mean ratings.

Hidden anchors can minimize potential biases in the MUSHRA method [24] by providing perceptual references throughout the assessment scale. High, middle, and low hidden anchors—which were chosen from the hidden anchors and SAPs employed to develop the QESTRAL model [23]—were employed to calibrate the top, middle, and bottom of the scale. The hidden anchors were functionally different from the SAPs; the former calibrated the assessment scale, and the latter were impaired domestic audio systems which were compared with automotive audio systems. Table 3 lists the hidden anchors.

Table 3. Hidden anchors for the listening test.

Hidden Anchor	Description
High Anchor	3/2-Channel Reference System
Middle Anchor	Channel Order Randomized
Low Anchor	1/0-Channel Downmix Reproduced Asymmetrically by the Left Surround Loudspeaker Only

Table 4. Program items for the listening test.

Genre Type	Description
Music (Classical)	Baroque music excerpt from Johann Sebastian Bach, "Concerto No.4 in G-Major." Wide continuous front stage including localizable instrument groups. Ambient surrounds with reverb from front stage.
Music (Pop)	Excerpt from "Faith" by Sheila Nicholls. Wide continuous front stage, including guitars, bass, and drums. Main vocal in center loudspeaker. Harmony vocals, guitars, and drum cymbals in left-surround and right-surround loudspeakers.
TV Sport (Tennis)	Wimbledon tennis match. Commentators and clapping. Commentators panned midway between the left, center, and right loudspeakers. Audience clapping in 360°.

Table 5. Number of subjects who perceived a difference between BRIRs of original and truncated lengths.

BRIR	Truncation (Samples)	Number of Subjects
3/2-Channel Reference System	12,000	2 out of 6
3/2-Channel to 1/0-Channel Downmix	12,000	1 out of 6
Audi A8, Front Seats	8,000	1 out of 6
Audi A6	8,000	0 out of 6

Three five-channel program items used to develop the QESTRAL model [23]—which aimed to span a representative range of ecologically valid material—were used in the listening test. Table 4 lists the program items.

The BRIRs were truncated in length to minimize the demands of real-time convolution, with an A/B-comparison informal listening test undertaken to ensure that artifacts were inaudible [25]. Using the tennis program item, four BRIRs—two domestic audio-based and two automotive audio-based—of original length and those truncated to either 12,000 samples (0.250 s) or 8,000 samples (0.167 s) were compared over one trial. Six assessors could not consistently perceive (Table 5) and describe any differences between the original and truncated BRIRs, which suggested

Table 6. Reverberation times (T20) for BRIRs of original and truncated lengths, averaged over 500 Hz and 1 kHz.

	Reverberation Time (s)	
BRIR	Original	Truncated
3/2-Channel Reference System	0.104	0.104 (12,000 samples)
3/2-Channel to 1/0-Channel Downmix	0.127	0.127 (12,000 samples)
Audi A8, Front Seats	0.050	0.052 (8,000 samples)
Audi A6	0.077	0.077 (8,000 samples)

that the truncation effects were minimal compared with the differences between the BRIRs.

Reverberation times (T20 [26]) of the four BRIRs were calculated based on BRIRs of the left channel and left ear, for a 0° head angle (Table 6). The results averaged over the mid-frequency octave bands of 500 Hz and 1 kHz were identical between the original and truncated BRIRs except for Audi A8, Front Seats, which differed by 0.002 s. The BRIRs were truncated at minimum around twice the longest reverberation time. Although the truncation lengths differed here, in the main listening test (SEC. 1.4), all the BRIRs were truncated to 12,000 samples to ensure the stability of AM3D Convolution Box.

Most comfortable listening levels for headphone reproduction were established for each program item to prevent fatigue over time [25]. Six assessors auditioned each of the three program items with the reference system, which resulted in the classical item being attenuated by 2 dB, the pop item by 0 dB, and the tennis item by 5 dB. Each result was informally auditioned by the experimenter to confirm its acceptability. The levels resulting from the experiment were reproduced during the listening tests by configuring each program item, the RME Hammerfall, and AM3D Convolution Box.

The BRIRs were equalized in loudness using the GENESIS [27] implementation of the loudness model for time-varying sounds by Glasberg and Moore [28]. In the absence of a model that predicts binaural loudness, the left and right ear signals were summed before they were entered into the loudness model, as binaural loudness can be estimated by summing the loudness of each ear [5]. The predicted loudness of the BRIRs were within ±0.5 phons of the reference system.

1.3 Assessors

The listening test employed eleven assessors from the University of Surrey that consisted of three undergraduate Tonmeister students and seven Ph.D. research students from the Institute of Sound Recording, and one Ph.D. research student from the Centre for Vision, Speech, and Signal Processing. The ITU-R BS.1116 and ITU-R BS.1534 standards [22, 24]—which cover the perceptual assessment of small and intermediate auditory degradations, respectively—suggest using expert listeners as assessors. The above assessors can be considered expert listeners from their experience with sound recording, formal listening tests, or both. The assessors were not checked for normal hearing and their ages were not recorded, although their listening test results were screened prior to statistical analysis. They were remunerated and participated in three test sessions.

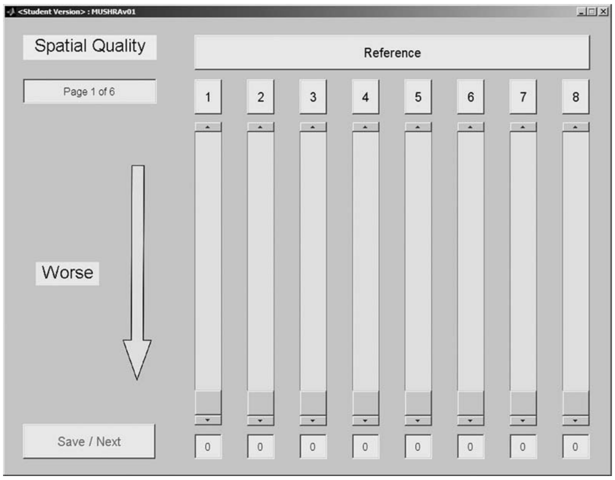


Fig. 2. User interface of the modified MUSHRA method.

1.4 Procedure

The assessors' task was to compare each BRIR to the reference system and then rate the perceived overall spatial quality, which is the perceived magnitude of difference in the spatial domain between a reference and degraded stimuli with a subjective judgement of acceptability [25]. The rating was primarily a fidelity evaluation (i.e., one measuring the degree of similarity to the reference) but also allowed assessors to give an opinion about the extent to which any differences were inappropriate, unpleasant, or annoying. The assessors were provided a list of changes in spatial characteristics they might perceive and incorporate in their evaluation, such as location, width, distance, depth, envelopment, and spaciousness [29]. A modified version of the MUSHRA method [23] was employed, which intended to minimize some potential biases in listening tests [30]. Fig. 2 shows the user interface of the modified MUSHRA method, which employed a label-free assessment scale with a range between 0 and 100.

The test was administered over three sessions, one for each program item, to avoid listener fatigue. A total of 48 BRIRs were rated in each session over six pages of evaluations. Each page contained eight BRIRs, where five were automotive audio systems and SAPs, and three were hidden anchors that appeared on every page. Each BRIR evaluation was repeated. The presentation order was randomized for the program items and BRIRs. Each session took less than 30 min, which was lower than the recommended maximum duration of 40 min [4, pp. 301–303]. The assessors who participated in the listening test performed less than 1 h of listening tests per day, which was lower than the recommended total of 2 h per day [4, pp. 301–303].

Prior to each session, the assessors participated in a familiarization session to reduce errors in the main session results, which could occur if they are unfamiliar with the task, user interface, stimuli, or a combination of these. The results of the familiarization sessions were checked to confirm that the assessors used the entire range of the assessment scale (i.e., between 0 and 100). The familiarization sessions employed the same modified MUSHRA interface as in the main sessions, and the BRIRs were administered randomly. The BRIRs included all three hidden anchors, an SAP, and four automotive audio systems.

1.5 Results

The main reason for the listening test was to establish whether the QESTRAL model in its current form is capable of predicting the overall spatial quality of automotive audio systems. Post-screening of the assessors was performed to assess the reliability of their ratings. The suitability of the data for comparison to predictions by the QESTRAL model was assessed.

1.5.1 Assessor Post-Screening

Post-screening of the assessors was performed using PanelCheck [31] to identify any who provided unreliable overall spatial quality ratings and hence should be removed before performing statistical analysis [25]. The discrimination ability of the assessors was investigated by the Tucker-1 correlation loadings, eggshell, and correlation plots. The consistency of the assessors was investigated by the mean square error (MSE) plot.

The post-screening revealed that one assessor demonstrated low discrimination ability and consistency. The Tucker-1 correlation loadings plot showed that most assessors used the rating scale similarly, apart from the one divergent assessor who used different criteria to rate overall spatial quality. The eggshell plot showed that most assessors ranked the stimuli in a similar order, again apart from the one divergent assessor who ranked the low-quality and middle-quality stimuli differently. The correlation plots showed that each assessor displayed acceptable discrimination between the stimuli, except for the divergent assessor whose ratings spanned a very wide range for the low-quality and middle-quality stimuli. The comments by the divergent assessor revealed that timbral criteria were used to rate overall spatial quality, which was contrary to instructions, and therefore their results were removed from the statistical analysis.

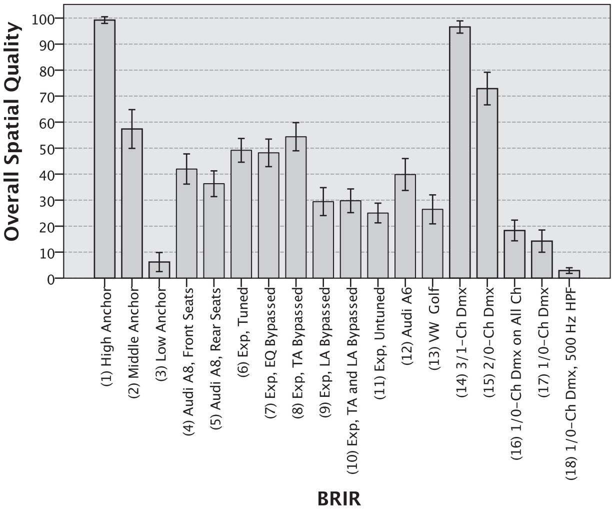


Fig. 3. Means and 95% confidence intervals for BRIR, averaged over program item, assessor, and repetition. Exp = experimental system; TA = time alignment; LA = level alignment; Dmx = downmix; and HPF = high-pass filter.

1.5.2 Statistical Analysis

Statistical analysis was performed to identify the statistically significant main effects of experimental factors and their interactions, and to identify the magnitude of an observed effect (Table 7). A mixed ANOVA model [4, pp. 203–215] was employed, where “BRIR” and “Program Item” were treated as fixed factors and “Repetition” and “Assessor” were treated as random factors. The “Repetition” factor was removed because it and its interactions were not statistically significant or borderline statistically significant at the 0.05 level. The analysis revealed that all the experimental factors and interactions were statistically significant with effect sizes of partial eta squared (η_p^2) spanning from medium (i.e., $\geq$0.06 and <0.14) to large (i.e., $\geq$0.14) [32]. The effect size ranges are guidelines and need to be interpreted within the context of the research field. An R^2 measure based on the likelihood ratio [33] showed that the model was a good fit to the data (i.e., $R_{LR}^2 = 0.913$, where R_{LR}^2 ranges from 0 to 1, and 1 indicates a perfect fit).

Fig. 3 shows the mean overall spatial quality ratings and 95% confidence intervals, which was averaged over program item, assessor, and repetition. The BRIR * Program Item interaction was statistically significant, which means that certain BRIRs were rated differently depending on the program item. The (5) Audi A8, Rear Seats and (12) Audi A6 BRIRs had fairly large interaction effects with program items. The BRIR * Assessor interaction was statistically significant, which means that certain BRIRs were rated differently depending on the assessor. One assessor rated the majority of BRIRs consistently lower than other assessors. These few causes of the statistically significant interactions with large effect sizes were not considered to substantially affect the mean ratings of the listening test results.

The highest-rated automotive audio system (i.e., (8) Experimental System, Time-Alignment Bypassed) was about half the overall spatial quality of the domestic audio-based reference system (i.e., (1) High Anchor). The means of the automotive audio systems spanned between around 25 and 55 points, which appears to be a result of using domestic audio-based BRIRs of much higher and lower relative quality. The mean width of the confidence intervals is 9.09 points, which is comparable to the average listener rating error of around 10 points by Conetta [23], who conducted a similar listening test over loudspeakers.

Table 7. ANOVA table.

	F-ratio	Significance	Effect Size
BRIR	$F(17, 153.000) = 88.98$	$p < 0.001$	$\eta_p^2 = .908$
Program Item	$F(2, 18.000) = 5.38$	$p < 0.05$	$\eta_p^2 = .374$
Assessor	$F(9, 27.584) = 3.28$	$p < 0.01$	$\eta_p^2 = .517$
BRIR * Program Item	$F(34, 306.000) = 5.22$	$p < 0.001$	$\eta_p^2 = .367$
BRIR * Assessor	$F(153, 306.000) = 1.63$	$p < 0.001$	$\eta_p^2 = .449$
Program Item * Assessor	$F(18, 306.000) = 2.21$	$p < 0.01$	$\eta_p^2 = .115$
BRIR * Program Item * Assessor	$F(306, 540.000) = 1.70$	$p < 0.001$	$\eta_p^2 = .491$

1.5.3 Listening Test Result Suitability

When the ratings of the automotive audio systems were combined with those of the hidden anchors and SAPs, the listening test results spanned a wide range of ratings. As mentioned previously, the mean confidence interval width of the listening test results was similar to the average listener rating error by Conetta [23]. The listening test results, which evenly span the entire assessment scale, and their mean confidence interval width, which is similar to that resulting from listening tests to develop the QESTRAL model, suggest that the listening test results can be employed for a more thorough evaluation of model performance and potential modification of the model if necessary.

2 PERFORMANCE OF QESTRAL MODEL FOR AUTOMOTIVE AUDIO SYSTEMS

The listening test results were compared to predictions by the QESTRAL model to evaluate its ability to predict the overall spatial quality of automotive audio systems. Previous versions of the model used BRIRs produced from anechoic simulations. The latest version of the model [34]—which was modified to accept measured BRIRs—was employed to predict the overall spatial quality:

$$\begin{aligned}
&\text{Predicted Overall Spatial Quality}\\
&\quad = -0.66\ \text{iacc\_9band}\\
&\qquad - 0.60\ \text{front\_angle\_diff}\\
&\qquad - 15.88\ \text{mean\_entropy}\\
&\qquad + 0.012\ \text{std\_spectral\_rolloff}\\
&\qquad + 341.66\ \text{max\_rms\_diff}\\
&\qquad + 100.00. \qquad (1)
\end{aligned}$$

Table 8 describes the metrics in the model. The process of deriving these metrics involves binaural measurements of soundfields using a set of probe signals [9], followed by the appropriate calculations to determine a single figure of merit.

Fig. 4(a) shows predicted overall spatial quality scores compared to perceived overall spatial quality scores (i.e., listening test results). The overall trend is correct but the fit to some data points is poor: the correlation (R^2) was 0.72 and the RMSE was 29.39%. The best-fit line has a large Y-intercept and shows a large tilt compared with the ideal-relationship line. On average, the automotive audio systems [Fig. 4(b)] show a larger difference (i.e., 30.79) between predicted and perceived scores compared with the difference (i.e., 19.95) for hidden anchors and SAPs [Fig. 4(c)]. The model was incapable of accurately predicting the overall spatial quality of automotive audio systems because it could not account for the aspects of spatial quality specific to them. In the next section, the model is modified to investigate whether prediction accuracy can be improved.

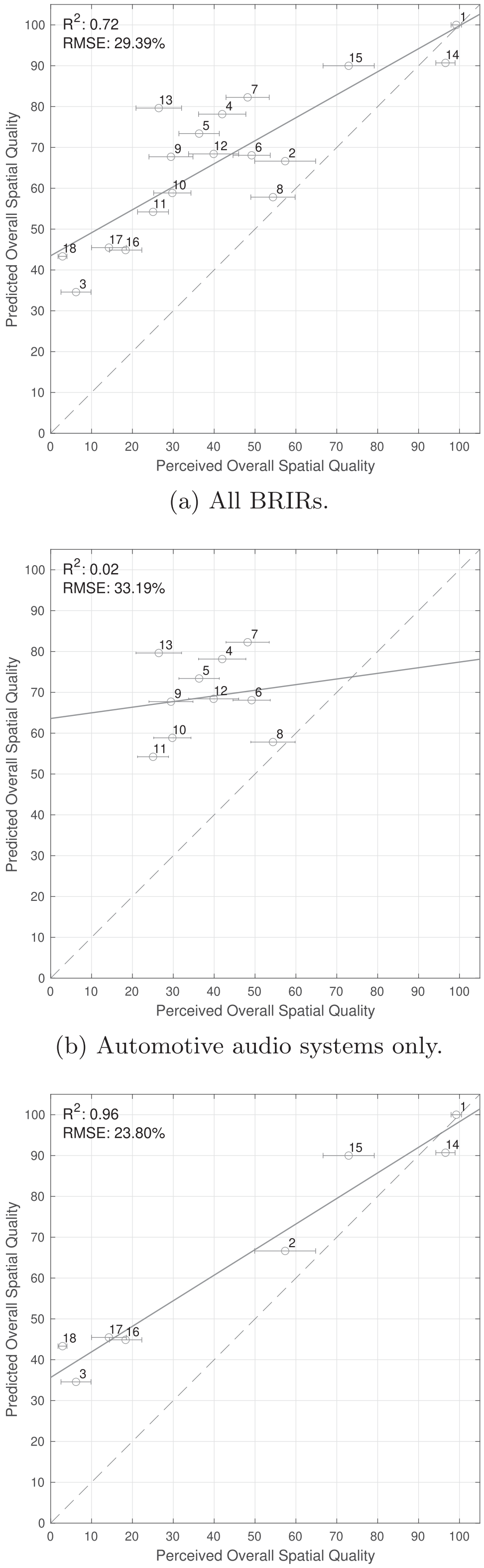
Table 8. List of the metrics in the QESTRAL model.

Metric Name	Description
iacc_9band	The mean value of an array of IACC values for nine critical bands between 570 and 2,160 Hz [35].
front_angle_diff	The mean value of the angle differences between the reference system and a DUT in the localization of seven sound sources in the frontal audio scene (i.e., ±30°) [35].
mean_entropy	The mean value of the left-ear and right-ear signal entropies [35].
std_spectral_rolloff	The standard deviation of the high-frequency spectral roll-off over the total number of time frames in the binaural signals [25].
max_rms_diff	The maximum value of the RMS level differences between the reference system and a DUT for an array of 36 angles spanning ±180° [35].

IACC = interaural cross-correlation; DUT = device under test.

3 MODEL CALIBRATION

This section covers the development of a modified version of the QESTRAL model that accounts for the spatial characteristics of automotive audio systems. Partial least squares (PLS) regression was selected as the model calibration method, which uses a set of orthogonal factors called latent variables (i.e., PLS components) to predict response variables (e.g., perceived overall spatial quality) [36]. The latent variables are decomposed from both predictor variables and response variables, which result in latent variables that best predict the response variables. PLS regression was chosen for modifying the QESTRAL model for automotive audio systems because it can be more accurate than other similar methods such as principal component regression and multiple linear regression, and was successfully employed to develop the QESTRAL model [23]. The nonlinear iterative partial least squares algorithm was cho-



(a) All BRIRs.

(b) Automotive audio systems only.

(c) Hidden anchors and SAPs only.

Fig. 4. Performance of the QESTRAL model. The dashed line shows the ideal relationship, and the solid line shows the best fit. The error bars show 95% confidence intervals. Refer to Fig. 3 to identify BRIRs.

sen to calculate PLS components because it is an accurate and computationally simple method [37]. The listening test results from SEC. 1.5 were employed as a calibration dataset because they were determined to be suitable based on the analysis performed in that section.

Model calibration proceeded over four stages. These stages were: recalibration to the dataset from this experiment using the original QESTRAL metrics (Appendix A.1); replacement of the QESTRAL metrics with more robust equivalents (Appendix A.2); evaluation of existing additional metrics (Appendix A.3); and evaluation of new metrics to reflect automotive-specific degradations (Appendix A.4).

3.1 Modified QESTRAL Model

In the first stage of model calibration, the original QESTRAL model (Eq. (1)) was recalibrated using the listening test results from SEC. 1.5, where the cross-validation results suggested that the model may not generalize to other automotive audio systems. Leave-one-out cross-validation was used to assess the generalizability of the model, where one point is left out from the dataset and predicted by a model created from the remaining data points [38]; this is repeated for each point in the dataset, and the correlation and error are calculated across the validation data points.

In the next stage, a metric in the original model based on maximum values (i.e., max_rms_diff) was replaced with a related metric based on mean values (i.e., mean_rms_diff), which improved both calibration and cross-validation performance compared with the recalibrated model. Then, existing additional metrics were iteratively employed to select an optimal number of metrics and PLS components that resulted in a potentially generalizable model which included six metrics and two PLS components. Finally, to further improve the potential generalizability of the model, five new metrics were created and assessed, where when a new metric that accounted for a wide scene width (i.e., front_hemisphere_scene_width) replaced an existing metric that accounted for a narrower scene width (i.e., front_scene_width), both calibration and cross-validation performance improved.

The newly created front_hemisphere_scene_width metric was incorporated in the six metric/two PLS component model because acceptable performance was achieved in terms of a potentially generalizable model. Eq. (2) shows the regression equation for this result, hereafter referred to as the modified QESTRAL model.

$$\begin{aligned}
&\text{Predicted Overall Spatial Quality}\\
&\quad = -0.68\ \text{iacc\_9band}\\
&\qquad - 0.75\ \text{front\_angle\_diff}\\
&\qquad - 16.19\ \text{mean\_entropy}\\
&\qquad - 0.0043\ \text{mean\_spectral\_rolloff}\\
&\qquad + 2514.65\ \text{mean\_rms\_diff}\\
&\qquad - 0.17\ \text{front\_hemisphere\_scene\_width}\\
&\qquad + 99.45,
\end{aligned} \tag{2}$$

Table 9. Progression of model calibration.

	Prediction		Calibration		Cross-Validation			
Stage	R^2	RMSE (%)	R^2	RMSE (%)	R^2	RMSE (%)	Mean VIF	Max VIF
Original Model (Sec. 2)	0.72	29.39	N/A	N/A	N/A	N/A	N/A	N/A
Recalibrated Model (Appendix A.1)			0.74	13.48	0.51	19.59	N/A	N/A
Existing Metric Replacement (Appendix A.2)			0.80	11.73	0.63	17.11	N/A	N/A
Large Metric Set, 16 Metrics/3 PLS Components (Appendix A.3.1)			0.89	8.93	0.74	14.46	9510.850	115704.307
Large Metric Set, 8 Metrics/2 PLS Components (Appendix A.3.2)			0.88	9.22	0.79	12.91	3.370	7.898
Large Metric Set, 7 Metrics/2 PLS Components (Appendix A.3.3)			0.88	9.06	0.80	12.69	2.740	5.937
Large Metric Set, 6 Metrics/2 PLS Components (Appendix A.3.4)			0.89	8.85	0.81	12.19	1.809	2.557
Modified Model (Sec. 4)			0.91	8.10	0.85	11.03	1.765	2.496

VIF = variance inflation factor.

where iacc_9band, front_angle_diff, mean_entropy, mean_spectral_rolloff, mean_rms_diff, and front_hemisphere_scene_width are the metrics chosen for the model. The iacc_9band, front_angle_diff, and mean_entropy metrics were defined in Sec. 2, while mean_spectral_rolloff is defined in Appendix A.3.1, mean_rms_diff is defined in Appendix A.2, and front_hemisphere_scene_width is defined in Appendix A.4.5.

The modified QESTRAL model contains metrics that are identical to the original QESTRAL model (i.e., iacc_9band, front_angle_diff, and mean_entropy), those that are altered from the metrics in the original model (i.e., mean_spectral_rolloff and mean_rms_diff, which replaced std_spectral_rolloff and max_rms_diff, respectively), and those that are newly added (i.e., front_hemisphere_scene_width). The metrics that were altered or newly added were found to be more relevant for predicting the overall spatial quality of automotive audio systems.

The mean_spectral_rolloff metric replaced std_spectral_rolloff because it was believed to be more perceptually relevant to perceived distance and perceived envelopment in automotive audio systems, had higher Pearson's correlation to perceived overall spatial quality in automotive audio systems, and was found to contribute toward a more potentially generalizable perceptual model by employing fewer PLS components. The mean_rms_diff metric replaced max_rms_diff because it was assessed to be more robust to extreme values, which possibly accounted more accurately for the alteration of sound-source levels caused by the acoustic environment of automotive audio systems. The addition of front_hemisphere_scene_width in the modified model possibly accounted more accurately for the scene-width characteristics of automotive audio systems, which are mainly rendered through a pair of front speakers that are oriented in an angular range wider than that of the domestic audio reference system (i.e., ±30°).

4 PERFORMANCE OF MODIFIED QESTRAL MODEL FOR AUTOMOTIVE AUDIO SYSTEMS

Table 9 summarizes the progression of modifying the QESTRAL model for automotive audio systems and Fig. 5 shows for the modified QESTRAL model, the predicted overall spatial quality scores compared to the perceived overall spatial quality scores. Two PLS components were chosen for the modified model because this resulted in the lowest cross-validation residual Y-variance, which minimizes the chance of an overfitted model. The modified model was considered to be a good fit to the listening test results because the majority of the residuals formed a straight line in a graph of normal probability Y-residuals. The modified model achieved similar cross-validation performance compared to the original QESTRAL model (R^2 = 0.78 and RMSE = 12.00%) [35]. The performance of the modified model improved considerably compared with that of the original model [Fig. 4(a)]: R^2 increased by 0.19 and RMSE decreased by 21.29%.

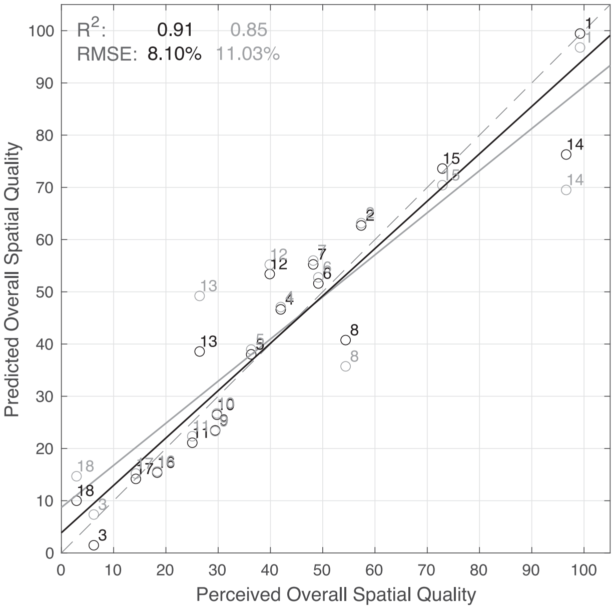


Fig. 5. Calibration and cross-validation performance of the modified QESTRAL model. The dashed line shows the ideal relationship and the solid lines show the best fit. The darker data refer to calibration results, while the lighter data refer to cross-validation results. Refer to Fig. 3 to identify BRIRs.

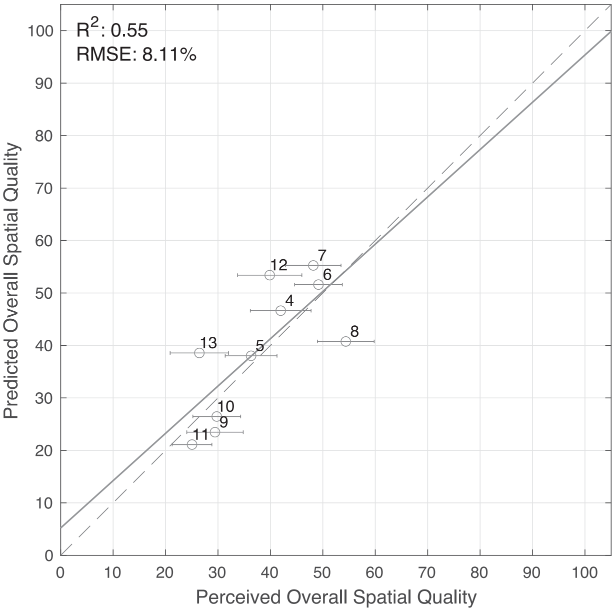


Fig. 6. Performance of the modified QESTRAL model for only the automotive audio systems. The dashed line shows the ideal relationship and the solid line shows the best fit. The error bars show 95% confidence intervals. Refer to Fig. 3 to identify BRIRs.

Fig. 6 shows for only the automotive audio systems, the predicted overall spatial quality scores by the modified QESTRAL model compared to the perceived overall spatial quality scores. The performance of the modified model improved considerably compared with the original model [Fig. 4(b)]: R^2 increased by 0.53 and RMSE decreased by 25.08. The modified model predicts automotive audio systems with an RMSE of 8.11% and the best-fit line follows the ideal-relationship line closely.

Table 10. Standardized coefficients of the modified QESTRAL model for automotive audio systems.

Metric Name	Standardized Coefficient
iacc_9band	−0.4531
front_angle_diff	−0.4480
front_hemisphere_scene_width	−0.3611
mean_rms_diff	0.2725
mean_entropy	−0.1600
mean_spectral_rolloff	−0.1291
constant term	3.6387

Compared with the other automotive audio systems, BRIRs 8 and 13 had worse calibration and cross-validation performance (Fig. 5). Fig. 10(a) shows that these two BRIRs lie on the extremes along the third PLS component, which could be interpreted as reflecting the changes in the high-frequency roll-off point of each BRIR (Appendix A.4.4). Frequency responses above 1 kHz of the two BRIRs show broad peaks and dips. A new metric—such as one that combines the mean_spectral_rolloff and log_rolloff_slope (Appendix A.4.4) metrics—may account more accurately for the high-frequency characteristics of the two BRIRs. However, as new metric development—including front_angle_raw_diff, front_angle_std, and surround_angle_diff—has shown (Appendix A.4), improving the prediction performance of a few outlier BRIRs can worsen the prediction performance of other BRIRs. Hence, a new metric could worsen overall prediction performance. The modifications to the original QESTRAL model showed that collectively, the ten automotive audio systems were predicted with an RMSE of 8.11% (Fig. 6).

Table 10 shows the standardized coefficients of the modified QESTRAL model. The largest standardized metric values were −0.4531 for iacc_9band and −0.4480 for front_angle_diff, which suggest that changes in perceived source width and changes in localization of sound images ±30° in front of the listener, respectively, are about equally the most important in predicting overall spatial quality. The next largest metric value was −0.3611 for front_hemisphere_scene_width, which suggests that changes in perceived scene width ±90° in front of the listener are the next most important in predicting overall spatial quality.

5 VERIFICATION OF MODIFIED QESTRAL MODEL FOR DOMESTIC AUDIO SYSTEMS

Modifications to the QESTRAL model for automotive audio systems could have invalidated its use for domestic audio systems. To evaluate the efficacy of the modified model for domestic audio, as well as to conduct an additional validation, the modified model predictions were compared to listening test results of SAPs from previous research [23] (Fig. 7). For these SAPs, which were based on domestic audio systems, the modified model achieved similar performance ($R^2 = 0.77$ and RMSE = 11.90%) to the original QESTRAL model predicting domestic audio systems ($R^2 = 0.78$ and RMSE = 12.00%) [35].

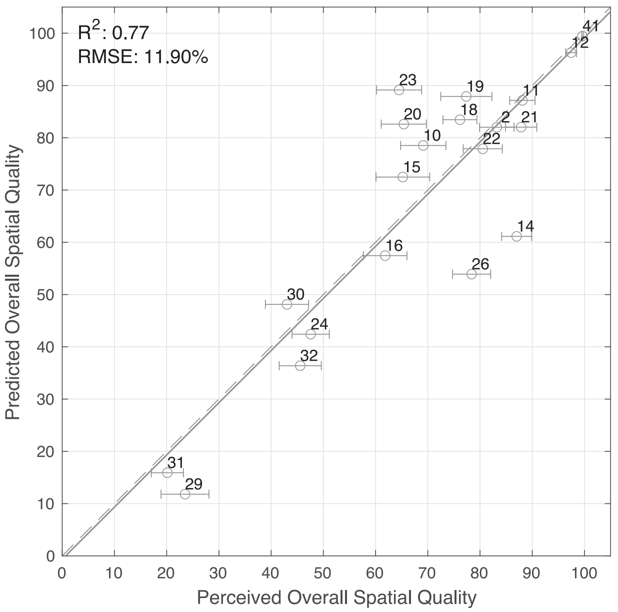


Fig. 7. Performance of the modified QESTRAL model for domestic audio systems. The dashed line shows the ideal relationship and the solid line shows the best fit. The error bars show 95% confidence intervals. Refer to Table G2 in Conetta [23] to identify SAPs.

The reasons why modifications to the QESTRAL model have not invalidated its use for domestic audio systems could be explained by comparing the metrics between the original and modified models [Eqs. (1) and (2), respectively]. First, three metrics were identical between the models (i.e., iacc_9band, front_angle_diff, and mean_entropy).

Second, two metrics in the modified model were altered versions of those in the original model (i.e., mean_spectral_rolloff and mean_rms_diff, which replaced std_spectral_rolloff and max_rms_diff, respectively). The mean_spectral_rolloff metric replaced std_spectral_rolloff because it was believed to be more perceptually relevant to perceived distance and perceived envelopment in automotive audio systems. This perceptual relevance could also apply to the SAPs evaluated in Fig. 7, for example SAP 24, which applied a 3.5 kHz low-pass filter on all five channels (i.e., L, R, C, LS, and RS). The mean_rms_diff metric replaced max_rms_diff because it was assessed to be more robust to extreme values, which possibly accounted more accurately for the sound-source level differences between the reference system and automotive audio systems. The mean_rms_diff metric also possibly accounted accurately for the sound-source level differences between the reference system and the SAPs evaluated in Fig. 7, for example, SAP 16, which rotated the channel order of the reference system one channel counterclockwise.

Third, the new addition of the front_hemisphere_scene_width metric in the modified model possibly did not affect the prediction accuracy of the SAPs evaluated in Fig. 7. This is because the majority of their spatial degradations could be accounted for by changes in scene width between ±90°.

6 CONCLUSION

A perceptual model that predicts the overall spatial quality of automotive audio systems was developed. Such a model is useful for rapid development of automotive audio systems that aim to match the spatial quality of a reference for domestic audio systems.

A listening test was conducted to collect overall spatial quality ratings of automotive audio systems. Statistical analysis of the results supported that they were reliable and hence suitable to be compared to predictions by the QESTRAL model.

To determine whether the QESTRAL model is capable of predicting the overall spatial quality of automotive audio systems, predictions by the model were compared to the listening test results. The model, in its original form, was found to be incapable of achieving this aim because it could not account for the aspects of spatial quality specific to automotive audio systems.

Modifications to the original QESTRAL model were carried out to improve its prediction accuracy for automotive audio systems. Metrics created during the development of the original model and those that were newly created for automotive audio systems—particularly to reflect the scene width across a larger range of angles—were applied in an iterative way to develop a model for automotive audio systems that achieved similar performance to the original model predicting domestic audio systems. The cross-validation performance of the modified model suggested that it can generalize to other automotive audio systems. In addition, modifications to the original model did not invalidate its use in domestic audio systems.

7 FUTURE WORK

Cross-validation is a mathematical estimate of how accurately a perceptual model can generalize outside the calibration context. Formal validation—which employs new listening test results—is more reliable than cross-validation to establish model generalizability. The formal validation could employ different automotive audio systems, program items, expert listeners, or a combination of these.

Other future work could include an extension of the modified QESTRAL model to incorporate immersive audio formats such as Dolby Atmos and MPEG-H Audio in automotive audio systems. The headphone-based auralization system will need to be validated for these formats and new metrics—such as those that incorporate height characteristics—may need to be developed.

8 ACKNOWLEDGMENT

This work was carried out as part of a project funded by Bang & Olufsen. The authors would like to thank the research staff at Bang & Olufsen and the listening test participants at the University of Surrey.

REFERENCES

[1] D. Clark, "Time Delay Imaging for Automotive Sound Systems," in *Proceedings of the SAE International Congress and Exposition*, pp. 79–84 (Detroit, Michigan) (1989 Feb.).

[2] R. Shively and W. House, "Perceived Boundary Effects in an Automotive Vehicle Interior," presented at the *100th Convention of the Audio Engineering Society* (1996 May), paper 4245.

[3] W. House, "Aspects of the Vehicle Listening Environment," presented at the *87th Convention of the Audio Engineering Society* (1989 Oct.), paper 2873.

[4] S. Bech and N. Zacharov, *Perceptual Audio Evaluation – Theory, Method and Application* (Wiley, Chichester, UK, 2006).

[5] B. Moore, B. Glasberg, and T. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240 (1997 Apr.).

[6] T. Thiede, W. Treurniet, R. Bitto, et al., "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29 (2000 Jan./Feb.).

[7] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "PESQ - The New ITU Standard for End-to-End Speech Quality Assessment," presented at the *109th Convention of the Audio Engineering Society* (2000 Sep.), paper 5260.

[8] A. Azzali, A. Farina, G. Rovai, G. Boreanaz, and G. Irato, "Construction of a Car Stereo Audio Quality Index," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), paper 6306.

[9] F. Rumsey, S. Zielinski, P. Jackson, et al., "QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction Using an Artificial Listener," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), paper 7595.

[10] R. Conetta, F. Rumsey, S. Zielinski, et al., "QESTRAL (Part 2): Calibrating the QESTRAL Model Using Listening Test Data," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), paper 7596.

[11] P. Jackson, M. Dewhirst, R. Conetta, et al., "QESTRAL (Part 3): System and Metrics for Spatial Quality Prediction," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), paper 7597.

[12] M. Dewhirst, R. Conetta, F. Rumsey, et al., "QESTRAL (Part 4): Test Signals, Combining Metrics, and the Prediction of Overall Spatial Quality," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), paper 7598.

[13] F. Toole and S. Olive, "Hearing is Believing vs. Believing is Hearing: Blind vs. Sighted Listening Tests, and Other Interesting Things," presented at the *97th Convention of the Audio Engineering Society* (1994 Nov.), paper 3894.

[14] N. Guttman and B. Julesz, "Lower Limits of Auditory Periodicity Analysis," *J. Acoust. Soc. Am.*, vol. 35, no. 4, pp. 610–610 (1963 Apr.). https://doi.org/10.1121/1.1918551.

[15] C. Darwin, M. Turvey, and R. Crowder, "An Auditory Analogue of the Sperling Partial Report Procedure: Evidence for Brief Auditory Storage," *Cogn. Psychol.*, vol. 3, no. 2, pp. 255–267 (1972 Apr.). https://doi.org/10.1016/0010-0285(72)90007-2.

[16] D. Begault, *3-D Sound for Virtual Reality and Multimedia* (Academic Press, Cambridge, MA, 1994).

[17] 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," presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), paper 7224.

[18] F. Postel, P. Hegarty, and S. Bech, "A Listening Test System for Automotive Audio: PART 5 - The Influence of Listening Environment on the Realism of Binaural Reproduction," presented at the *130th Convention of the Audio Engineering Society* (2011 May), paper 8446.

[19] T. Welti and S. Olive, "Validation of the Binaural Room Scanning Method Using Subjective Ratings of Spatial Attributes," in *Proceedings of the AES 48th International Conference: Automotive Audio* (2012 Sep.), paper 1-2.

[20] H. Møller, D. Hammershøi, C. Jensen, and M. Sørensen, "Transfer Characteristics of Headphones Measured on Human Ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217 (1995 Apr.).

[21] ITU, "Multichannel Stereophonic Sound System with and without Accompanying Picture," *ITU-R BS.775-3 Recommendation* (2012 Aug.).

[22] ITU, "Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems," *ITU-R BS.1116-1 Recommendation* (1997 Oct.).

[23] R. Conetta, *Towards the Automatic Assessment of Spatial Quality in the Reproduced Sound Environment*, Ph.D. thesis, University of Surrey, Guildford, UK (2011 Dec.).

[24] ITU, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," *ITU-R BS.1534-1 Recommendation* (2003 Jan.).

[25] D. Koya, *Predicting the Overall Spatial Quality of Automotive Audio Systems*, MPhil thesis, University of Surrey, Guildford, UK (2017 Sep.).

[26] ISO, "Acoustics - Measurement of Room Acoustic Parameters - Part 1: Performance Spaces," *ISO Standard 3382-1:2009* (2009 Jun.).

[27] GENESIS, "Loudness Online," http://genesis-acoustics.com/en/loudness_online-32.html (accessed on Aug. 22, 2013).

[28] B. Glasberg and B. Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *J. Audio Eng. Soc.*, vol. 50, no. 5, pp. 331–342 (2002 May).

[29] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666 (2002 Sep.).

[30] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening

Tests - A Review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451 (2008 Jun.).

[31] Research Council of Norway, "PanelCheck," http://www.panelcheck.com (accessed on Nov. 3, 2011).

[32] R. Kirk, "Practical Significance: A Concept Whose Time Has Come," *Educ. Psychol. Meas.*, vol. 56, no. 5, pp. 746–759 (1996 Oct.). https://doi.org/10.1177/0013164496056005002.

[33] M. Kramer, "R^2 Statistics for Mixed Models," in *Proceedings of the 17th Annual Kansas State University Conference on Applied Statistics in Agriculture*, pp. 148–160 (Manhattan, KS) (2005 Apr.).

[34] M. Dewhirst, "Personal Communication," (2013).

[35] P. Jackson, M. Dewhirst, R. Conetta, and S. Zielinski, "Estimates of Perceived Spatial Quality Across the Listening Area," in *Proceedings of the AES 38th International Conference: Sound Quality Evaluation* (2010 Jun.), paper 8-1.

[36] H. Abdi, "Partial Least Squares Regression and Projection on Latent Structure Regression (PLS Regression)," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 1, pp. 97–106 (2010 Jan./Feb.). https://doi.org/10.1002/wics.51.

[37] K. Esbensen, D. Guyot, F. Westad, and L. Houmøller, *Multivariate Data Analysis: In Practice: An Introduction to Multivariate Data Analysis and Experimental Design* (CAMO Process AS, Oslo, Norway, 2002).

[38] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. Roy. Statist. Soc. Ser. B*, vol. 36, no. 2, pp. 111–147 (1974 Jan.). https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

[39] R. Conetta, T. Brookes, F. Rumsey, et al., "Spatial Audio Quality Perception (Part 2): A Linear Regression Model," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 847–860 (2014 Dec.). https://doi.org/10.17743/jaes.2014.0047.

[40] D. Griesinger, "Pitch Coherence as a Measure of Apparent Distance in Performance Spaces and Muddiness in Sound Recordings," presented at the *121st Convention of the Audio Engineering Society* (2006 Oct.), paper 6917.

[41] H. Martens and M. Martens, *Multivariate Analysis of Quality* (Wiley, Chichester, UK, 2001).

[42] J. Hair, W. Black, B. Babin, and R. Anderson, *Multivariate Data Analysis: A Global Perspective* (Pearson, Upper Saddle River, NJ, 2010).

[43] B. Bowerman and R. O'Connell, *Linear Statistical Models: An Applied Approach* (Duxbury Press, Belmont, CA, 1990).

[44] R. Myers, *Classical and Modern Regression with Applications* (Duxbury Press, Belmont, CA, 1990).

A.1 RECALIBRATION

Four stages were involved to develop a modified QESTRAL model for automotive audio systems. These stages were: recalibration (Appendix A.1); existing metric replacement (Appendix A.2); large metric set (Appendix A.3); and new metric replacement (Appendix A.4).

The original QESTRAL model [Eq. (1)] was recalibrated using the listening test results from SEC. 1.5 to explore whether the original metrics with different weightings could improve prediction accuracy. Compared with the performance of the original model, the performance of the recalibrated model improved; calibration R^2 increased by 0.02 and calibration RMSE decreased by 15.91 (Table 9). However, the recalibrated model had poor cross-validation performance ($R^2 = 0.51$ and RMSE $= 19.59\%$), which suggests that the model may not generalize to other automotive audio systems.

A.2 EXISTING METRIC REPLACEMENT

The robustness of the metrics in the original model [Eq. (1)] was evaluated to reveal any metrics that were unduly influenced by extreme values. The iacc_9band, front_angle_diff, and mean_entropy metrics are based on calculating averages, hence they should be more robust to extreme values compared with metrics based on minimum or maximum values. The std_spectral_rolloff metric is based on calculating a standard deviation, hence it should also be more robust to extreme values. However, the max_rms_diff metric is based on calculating a maximum value, and therefore could be unduly influenced by extreme values.

The max_rms_diff metric was replaced by a more robust level-based metric called mean_rms_diff that calculates—for an array of 1-s pink-noise probe signals panned from 0° to 360° in 10° increments on the horizontal plane—the mean value of the level differences between the reference system and a device under test (DUT) [39]. This was implemented in the model, and the results showed that this improved both calibration and cross-validation performance compared with the recalibrated model in Appendix A.1; calibration R^2 increased by 0.06 and calibration RMSE decreased by 1.75, while cross-validation R^2 increased by 0.12 and cross-validation RMSE decreased by 2.48 (Table 9).

A.3 LARGE METRIC SET

The results so far indicated that the revised model still failed to accurately predict the spatial quality of some automotive audio systems, suggesting that additional metrics were needed to account for features unique to automotive audio systems. To account for this, metrics pooled from those employed to develop the QESTRAL model [25] and those from the current literature [40] that were surmised to be relevant to the spatial quality of automotive audio systems were implemented. An iterative process was undertaken to test and select the most relevant metrics, over four stages (Appendixes A.3.1 to A.3.4).

A.3.1 16 Metrics/Three PLS Components

There were 16 metrics employed initially, all of which were based on binaural signals from artificial-head measurements (binaural signals were considered to be potentially more perceptually relevant, especially compared with first-order microphone signals, due to the inherent head-related filtering). The mean_spectral_rolloff metric is the mean of the high-frequency spectral rolloff over the total number of time frames in the measured binaural signals [35]. The std_spectral_rolloff metric, which was included in the original model [Eq. (1)], was replaced with

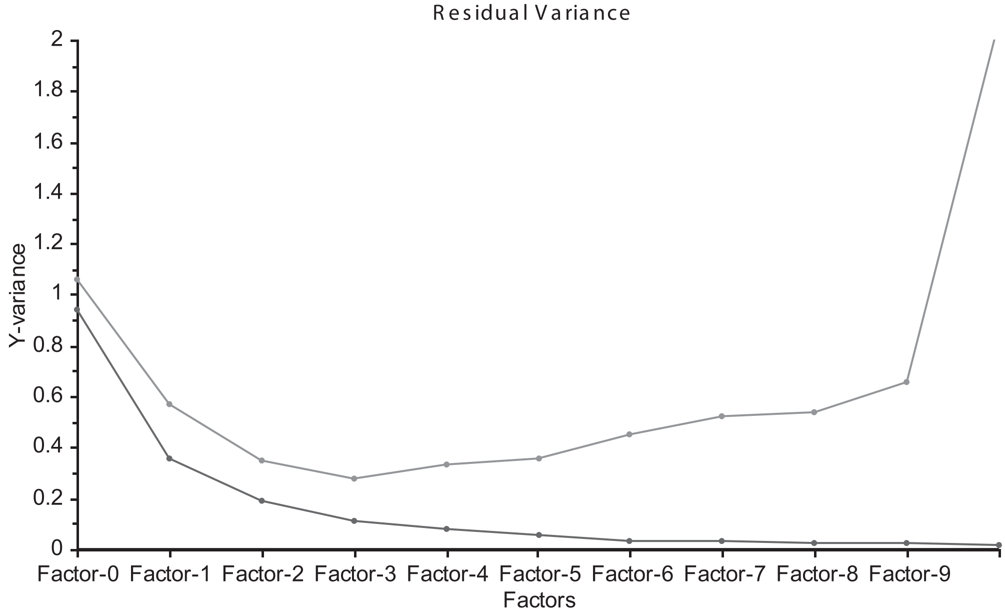


Fig. 8. Residual Y-variance as a function of PLS component number for the iteration that employed 16 metrics and 3 PLS components. The bottom data refer to calibration results, while the top data refer to cross-validation results. Although residual Y-variance was calculated for 16 PLS components, only 10 PLS components are shown for clarity.

mean_spectral_rolloff because the latter metric was believed to be more perceptually relevant to perceived distance and perceived envelopment in automotive audio systems [25], had higher Pearson's correlation to perceived overall spatial quality in automotive audio systems, and was found to contribute toward a more potentially generalizable perceptual model by employing fewer PLS components.

The residual cross-validation variance of the response variable (e.g., perceived overall spatial quality) can be interpreted as the error resulting from predicting new datasets [37]. The number of PLS components that coincide with the minimum residual cross-validation variance results in a model that is optimal regarding future prediction accuracy [41]. A lower number of PLS components can lead to an underfitted model, whereas a higher one can lead to an overfitted model. Three PLS components were chosen, as this number coincided with the minimum residual cross-validation variance (Fig. 8). This iteration showed acceptable performance for calibration, but not for cross-validation as the RMSE was a little high at 14.46% (Table 9).

Standardized coefficient and variance inflation factor (VIF) values of the metrics were evaluated. Standardized coefficients of a PLS regression model enable direct comparison of the relative effect of each independent variable (e.g., each metric) on the dependent variable (e.g., predicted overall spatial quality). The standardized coefficients are calculated from standardized data, which are created by subtracting the mean of the metric values from each metric value and then dividing by the standard deviation [42]. VIF—which is a measure of multicollinearity where two or more metrics have a strong linear relationship—was employed to remove metrics that were redundant toward predicting overall spatial quality. The values of VIF employed to determine multicollinearity were a mean VIF criterion of substantially greater than 1 [43], and a maximum VIF criterion of greater than 10 [44].

The number of metrics was reduced because the maximum VIF criterion was far exceeded. The max_angle_diff, max_iacc, max_iacc_9bands, 1/(1 − max_iacc), and max_rms_diff [25] metrics were removed because metrics based on maximum values are not robust to extreme values. The mean_iacc and mean_iacc_9bands metrics were removed because the iacc_9band metric contributed the most toward predicting overall spatial quality (i.e., the standardized coefficient values were 0.029, 0.126, and −0.156, respectively). The mean_angle_diff metric was removed because it demonstrated collinearity with front_angle_diff ($R = -0.976$, $p < 0.001$) and contributed less toward predicting overall spatial quality (i.e., its standardized coefficient value was 0.179 compared with −0.244).

A.3.2 Eight Metrics/Two PLS Components

The eight metrics that remained after the reduction procedure in the previous iteration were employed for a new iteration. Although the global minimum of the residual cross-validation variance was located at four PLS components (i.e., at 0.221), two PLS components were chosen as this number coincided with the first local minimum of the residual cross-validation variance (i.e., at 0.223). Esbensen et al. [37] recommend that the first local minimum is employed because choosing fewer PLS components leads to a more robust model that is less sensitive to noise and errors. This iteration displayed acceptable performance for calibration but not for cross-validation due to its slightly high RMSE of 12.91% (Table 9).

The standardized coefficients and VIF values of the metrics were evaluated to remove any metrics that contributed little toward predicting overall spatial quality. The hull metric—which could be considered as a measure of spa-

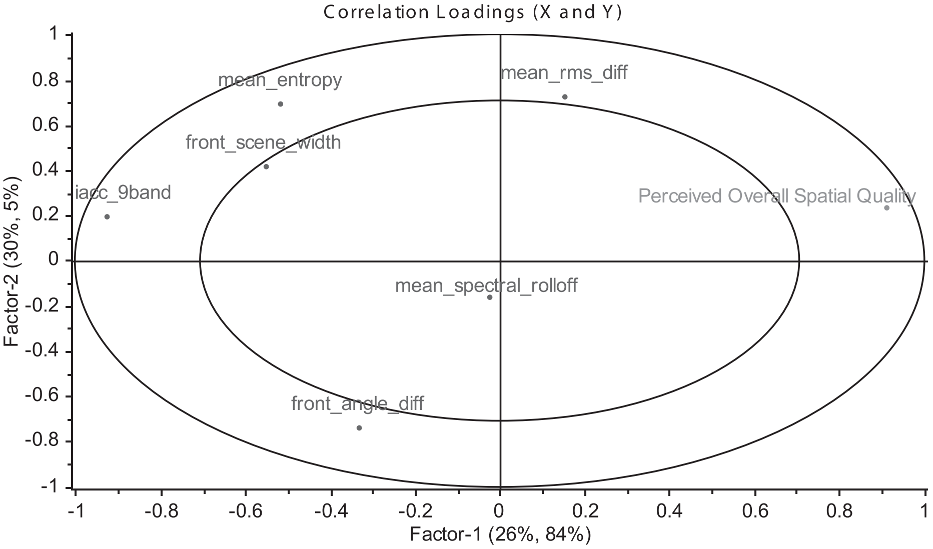


Fig. 9. Correlation loadings for the iteration that employed six metrics and two PLS components. The outer ellipse represents 100% explained variance, the inner ellipse represents 50% explained variance, and the middle of the ellipses represents no variance explained by either the first or second PLS component. The numbers in the parentheses refer to the total variance each PLS component explains. The first number refers to the *X*-variance (i.e., the metrics), and the second number refers to the *Y*-variance (i.e., perceived overall spatial quality).

tial scene width [11]—was removed because it contributed the least toward predicting overall spatial quality; its standardized coefficient value was the smallest at −0.058. The removal reduced the chance that this iteration of the model overfitted the calibration data.

A.3.3 Seven Metrics/Two PLS Components

The seven metrics that remained after the reduction procedure in the previous iteration were employed for a new iteration. Two PLS components were chosen because a break from a monotonic decrease was observed in the residual cross-validation variance. Esbensen et al. [37] mention this criterion—in addition to choosing the number of PLS components based on the minimum residual cross-validation variance—to choose the optimal number of PLS components for a potentially generalizable model. This iteration demonstrated acceptable performance for calibration but not for cross-validation due to its slightly high RMSE of 12.69% (Table 9). The standardized coefficients and VIF values of the metrics were evaluated to remove any metrics that contributed little toward predicting overall spatial quality. The 1/(1 − mean_iacc) [25] and iacc_9band metrics demonstrated the highest VIF values (i.e., 4.384 and 5.937, respectively). Of the two, 1/(1 − mean_iacc) was removed because it contributed less toward predicting overall spatial quality; its standardized coefficient had a lower value than that of iacc_9band (i.e., 0.192 and −0.401, respectively).

A.3.4 Six Metrics/Two PLS Components

The six metrics that remained after the reduction procedure in the previous iteration were employed for a new iteration. Two PLS components were chosen, as this number coincided with the minimum residual cross-validation variance. This iteration demonstrated acceptable performance for calibration but not for cross-validation as the RMSE was still slightly high at 12.19% (Table 9).

The standardized coefficients and VIF values of the metrics were evaluated to remove any metrics that contributed little toward predicting overall spatial quality. To examine this iteration of the model in more detail, a correlation loadings plot was analyzed, which shows how the metrics and perceived overall spatial quality are correlated with the first two PLS components (Fig. 9). The mean_spectral_rolloff metric is near the center of the plot, and had the lowest standardized coefficient of −0.130, which means that it contributed the least toward predicting overall spatial quality. However, upon analyzing a further iteration without the metric (i.e., five metrics and two PLS components), its presence revealed improved calibration and cross-validation performance, decreased the mean VIF from 1.875 to 1.809, and improved the accuracy of the predicted score for BRIR 13 (the VW Golf) for calibration. This BRIR displayed early roll-off of high frequencies, which the metric could account for. Hence, mean_spectral_rolloff was retained and the iterations were terminated at six metrics and two PLS components.

A.4 NEW METRIC REPLACEMENT

The model developed by iteratively narrowing down a large set of metrics exhibited cross-validation RMSE that was slightly high. To further improve the potential generalizability of the model, five new metrics were created and their performance was assessed (Appendixes A.4.1 to A.4.5).

A.4.1 Localization-Related Metric - front_angle_raw_diff

Regression analysis indicated that localization-related metrics were highly correlated to the second PLS component, which accounted for 22% of the variance in the data, so additional localization metrics were evaluated. Informal audition of automotive audio systems considered to be outliers [i.e., BRIRs 8 and 13 in Fig. 4(a)] revealed differences in spatial scene skew. The front_angle_raw_diff metric was created to investigate whether—compared with the front_angle_diff metric—a different approach to averaging multiple localization angles could model spatial scene skew more accurately for these outliers. The metric is a modified version of front_angle_diff that is based on the mean of the raw differences of the angles which can account for skew direction.

An exploratory model was created to assess whether the new metric can further improve the performance of the six metric/two PLS component model developed in Appendix A.3.4. When front_angle_raw_diff replaced front_angle_diff, both calibration and cross-validation performance worsened. For calibration, R^2 decreased by 0.18 and RMSE increased by 5.44, while for cross-validation, R^2 decreased by 0.30 and RMSE increased by 7.56 (Table 11); the prediction improved for the outliers but at the expense of other BRIRs. Therefore, this metric was not retained.

A.4.2 Localization-Related Metric - front_angle_std

An alternate predictor of spatial scene skew was evaluated: front_angle_std calculates the standard deviation of the differences in localization angles in the frontal audio scene (i.e., ±30°) between the reference system and a DUT. When front_angle_std replaced front_angle_diff, both calibration and cross-validation performance again worsened. Calibration R^2 decreased by 0.08 and cross-validation R^2 decreased by 0.11 (Table 11). Calibration RMSE increased by 2.69 and cross-validation RMSE increased by 3.16. This metric had a mixed effect on the outliers mentioned above: the calibration and cross-validation of BRIR 8 improved while those of BRIR 13 worsened. This metric was also not retained.

A.4.3 Localization-Related Metric - surround_angle_diff

The locations of the left-surround and right-surround loudspeakers in five-channel automotive audio systems are likely to have shifted from those in a standardized five-channel surround system [21]. These shifts—along with the shifts of the front-left, front-right, and center loudspeakers—could affect perceived scene skew, particularly that of the two outliers [i.e., BRIRs 8 and 13 in Fig. 4(a)]. The ±110° angular range—which coincides with the locations of the surround loudspeakers—could also account more accurately for the scene skew in certain domestic audio degradations based on downmixes (e.g., the low anchor, which reproduced a 1/0-channel downmix through the left-surround loudspeaker). To better account for these situations of scene skew, a localization-related metric called surround_angle_diff was created. The metric is calculated similarly to the front_angle_diff metric but with an extended angular range of ±110°.

When surround_angle_diff replaced front_angle_diff, calibration performance slightly worsened where R^2 decreased by 0.01 and RMSE increased by 0.53, though cross-validation performance improved (Table 11). The replacement improved the prediction accuracy for the two outliers along with many other BRIRs. However, this was at the expense of substantial overprediction of the high anchor and substantial underprediction of the 3/2-Channel to 3/1-Channel Downmix, so this metric was not retained.

A.4.4 Timbre-Related Metric - log_rolloff_slope

Fig. 10 shows the score and loading plots for the six metric/two PLS component model. The first and third PLS components are shown in the plots. In the scores plot, four automotive audio systems lie along the third PLS component (i.e., BRIRs 8, 5, 4, and 13). BRIR 13 in the scores plot is similar in direction—along the third PLS component—to the mean_spectral_rolloff metric in the loadings plot, and hence the third PLS component could be interpreted as changes in timbre [37], or more specifically, changes in the high-frequency roll-off point of each BRIR. Informal audition of the four automotive audio systems revealed that the magnitude of the high-frequency content matched the order along the third PLS component: BRIR 8 was perceived to have the most high-frequency content, followed by BRIRs 5, 4, and 13.

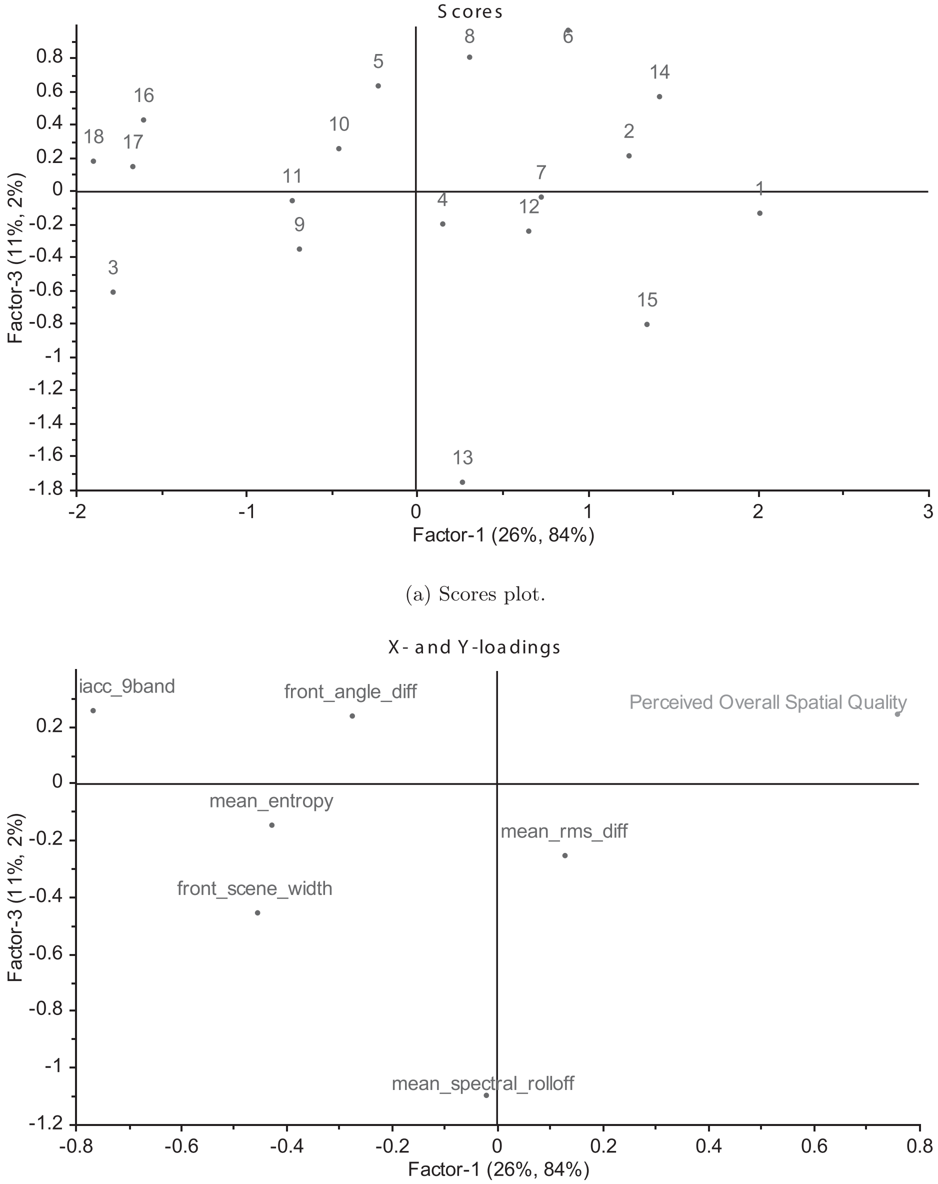
The values of the high-frequency roll-off point (i.e., the values of the mean_spectral_rolloff metric) for the four automotive audio systems were compared. The ordering of the automotive audio systems along the third PLS component did not agree with that of the values of the high-frequency roll-off point, so a more accurate metric of the high-frequency variation was sought.

Table 12 lists the slopes of high-frequency roll-off for the four automotive audio systems (i.e., BRIRs 8, 5, 4, and 13). The monotonically increasing slope rates were consistent with the order of the automotive audio systems along the third PLS component. Based on these results, a timbre-related metric called log_rolloff_slope was created based on the slope of the high-frequency roll-off between 5 and 10 kHz on a logarithmic frequency scale. This frequency range was chosen because the slopes varied the most.

When log_rolloff_slope replaced mean_spectral_rolloff, the predicted overall spatial quality scores for calibration and cross-validation of BRIR 8 were virtually unchanged, while those of BRIR 13 became less accurate (i.e., the error increased by 10.10 points for calibration and 3.97 points for cross-validation). The replacement also caused mean VIF to increase from 1.809 to 2.363, which suggests multicollinearity. Finally, the replacement decreased model performance (Table 11). This metric was therefore not retained.

Table 11. Performance of new metrics when incorporated in six metric/two PLS component model.

		Calibration		Cross-Validation	
Type	Metric	R^2	RMSE (%)	R^2	RMSE (%)
Localization	front_angle_raw_diff	0.71	14.29	0.51	19.75
Localization	front_angle_std	0.81	11.54	0.70	15.35
Localization	surround_angle_diff	0.88	9.38	0.83	11.68
Timbre	log_rolloff_slope	0.86	9.88	0.78	13.06
Scene Width	front_hemisphere_scene_width	0.91	8.10	0.85	11.03



(a) Scores plot.

(b) Loadings plot.

Fig. 10. Plots that aid interpreting the BRIRs that lie along the third PLS component for the six metric/two PLS component model. The numbers in the parentheses refer to the total variance each PLS component explains. The first number refers to the X-variance (i.e., the metrics), and the second number refers to the Y-variance (i.e., perceived overall spatial quality). Refer to Fig. 3 to identify BRIRs.

Table 12. Slopes of high-frequency roll-off and BRIR distribution order from the top to bottom of the third PLS component.

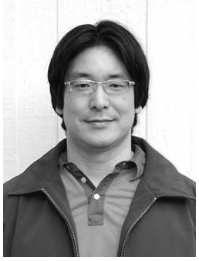
BRIR Number	BRIR Name	Slope (dB/octave)	Distribution Order Along Third PLS Component
8	Tech Car, Time-Alignment Bypassed	−10.5	8
5	Audi A8, Rear Seats	−11.1	5
4	Audi A8, Front Seats	−12.2	4
13	VW Golf	−16.4	13

A.4.5 Width-Related Metric - front_hemisphere_scene_width

Informal audition revealed differences in perceived scene width between the outliers [i.e., BRIRs 8 and 13 in Fig. 4(a)]. A width-related metric called front_hemisphere_scene_width was created to investigate whether a wider scene width compared with the one determined by the front_scene_width metric could account more accurately for the perceived scene width of these outliers. The front_hemisphere_scene_width metric calculates the largest angle spanned by a spatial scene in front of the listener between ±90° as opposed to ±30° for front_scene_width [25].

When front_hemisphere_scene_width replaced front_scene_width, calibration and cross-validation performance improved (Table 11). Both calibration and cross-validation R^2 increased (i.e., by 0.02 and 0.04, respectively), and both calibration and cross-validation RMSE decreased (i.e., by 0.75 and 1.16, respectively). The replacement improved the predicted overall spatial quality scores of the two outliers for both calibration and cross-validation, without a substantial increase in error for the other stimuli. Hence, this metric was retained.

THE AUTHORS



Daisuke Koya

Russell Mason

Martin Dewhirst

Søren Bech

Daisuke Koya received a B.S. degree in electrical engineering from the University of California at Los Angeles, Los Angeles, CA; an M.S. degree in music engineering from the University of Miami, Coral Gables, FL; an MRes degree in audio engineering from the University of Essex, Colchester, UK; and an MPhil degree in sound recording from the University of Surrey, Guildford, UK. His MPhil research was funded by Bang & Olufsen, and it investigated the modeling of spatial quality in automotive audio systems. His research interests include psychoacoustics and loudspeakers. He has had internships at Harman International and Apple Inc., and he has worked as a loudspeaker engineer.

•

Russell Mason was awarded a Ph.D. in audio engineering and psychoacoustics from the University of Surrey in 2002 and is currently a senior lecturer in the Institute of Sound Recording, University of Surrey, with over 100 published journal and conference papers. His research interests are focused on psychoacoustic engineering, and he has led the development of subjective evaluation methods and computational models of aspects of auditory perception, for application in spatial audio, evaluation of timbre, source separation, and personal sound zones.

•

Martin Dewhirst studied mathematics for his Master's degree at UMIST in Manchester, UK, and received his Ph.D. in modeling spatial aspects of psychoacoustics from the University of Surrey, UK. Following six years lecturing at the Institute of Sound Recording at the University of Surrey, he is now a senior software and firmware engineer at Focusrite working on digital audio interfaces and digital signal processing.

•

Dr. Søren Bech is Director of Research at Bang & Olufsen and Professor of Audio Perception at Aalborg University, Section AI and Sound, The Technical Faculty of IT and Design. Dr. Bech is also Adjunct Professor at Surrey University (GB) and McGill University (CAN). He received an M.Sc. and Ph.D. from the Department of Acoustic Technology (AT) of the Technical University of Denmark. From 1982–1992, he was Research Fellow at AT studying perception and evaluation of reproduced sound in small rooms. Dr. Bech has authored 50 peer-reviewed journal papers and more than 90 conference contributions. He has been PI of 19 funded (EU and national funding bodies) international collaborative research projects. He is Fellow of the Acoustical Society of America and the AES, is past Governor and Vice-President of the AES, and now serves as associate technical editor of the Journal of the AES. Dr. Bech has been vice-chair of the International Telecommunication Union working group 10/3. His research interest includes psychoacoustics and in particular human perception of reproduced sound in small and medium-sized rooms. Other interests include experimental procedures and statistical analysis of data from sensory analysis of audio and video quality.