Audio Engineering Society

Conference Paper

Presented at the AES 2026 International Conference on Automotive Audio
2026 July 29-31, Detroit, MI, USA

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (https:/aes.org/publications/elibrary), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Towards a Virtual Listener Panel for Car Audio System Evaluation

Roman Vlasov[1], Ivan Shishalov[1], Andrey Filimonov[1] and Mathias Johansson[1]

[1] *Harman International, Novi, MI*

Correspondence should be addressed to Roman Vlasov (roman.vlasov@harman.com)

ABSTRACT

We propose a machine-learning method that predicts trained-listener panel ratings of vehicle audio systems from in-situ microphone-array measurements. Using a dataset collected over more than 10 years from 1177 vehicle audio systems, the model is trained to predict listener-score distributions, not only mean scores, for subjective sound-quality attributes on a 1–10 scale. In this paper we focus on the timbral/spectral attribute and evaluate performance on a held-out set of 237 systems. Results show that prediction accuracy improves consistently with dataset size and that distribution-aware modelling captures not only expected score level but also listener disagreement and uncertainty. The study highlights both the feasibility of virtual listener panels and the importance of large-scale, reliable benchmarking data.

1. Introduction

Human listeners are remarkably consistent [2] in preference ranking different loudspeakers, especially when given the right conditions (not being biased by aesthetics, loudness, etc). Further, experiments reported in [2] show that subjective *preference* is closely related to subjective *accuracy* or *fidelity* of sound reproduction. Not surprisingly, people tend to prefer a sound reproduction system that, in their judgement, reproduces the original content with high fidelity. Subjective assessments are also known to be correlated with certain objective measurements, where for (bookshelf and above-sized) HiFi speaker preferences for home audio use [3], the on- and off-axis frequency responses along with low-bass extension are the key factors that predict subjective preference rankings of such speakers.

All of this suggests that it should be possible to predict the results of a subjective listener evaluation, using just a few objective measurements of the sound system.

For developers of audio systems, it is often desirable to judge not just overall preference but also how an audio system is perceived in terms of spatial performance, dynamics and power handling, distortion, etc, in addition to timbral qualities. There is thus a need for extending models that predict human listener ratings and preferences. With more detailed ratings of various sound attributes, an audio system can be improved and optimized to reach performance targets for the various subjective attributes.

Today, audio system manufacturers frequently use informal or formalized listening tests with trained or untrained listeners. To reach satisfactory statistical reliability of such listening experiments, the method needs to be designed carefully and requires a large body of listeners. Each listener must spend considerable time with the system, listening to various types of content, and they need to be properly trained to provide consistent results. Therefore, most companies do not invest the time, resource and money to evaluate and validate system performance systematically. If there existed a measurement-based, reliable, and accurate method for predicting the distribution of scores that such a listening panel yields, the developer of an audio system would be

able to judge and improve their systems iteratively during the development cycle. End customers would be able to validate claims, and the audio industry might become less dependent on subjective, error-prone, evaluation. Note that there is nothing wrong with subjective evaluation (we are after all creating audio systems for the sole purpose of subjectively pleasing listeners!), but it is difficult, time-consuming, and costly to reach the desired statistical confidence and accuracy. An automated system, a *virtual listening panel*, able to accurately and reliably predict the results of a carefully defined listening benchmark process, remains an intriguing, yet elusive, goal for the audio industry.

The quest for such a method starts with a well-designed subjective listening benchmarking process and data captured over a long time with many subjects and for many audio systems. An example of such a method, aimed at in-depth automotive audio system evaluation, and developed over several years, is described in a companion paper [1]. Based on a large dataset collected using that method, we here report initial findings from research into predicting detailed subjective ratings of various attributes, based on objective in-situ measurements.

The topic of predicting listener preferences is not new. There is a large body of work, not least by Toole, see [2] for an overview, and Olive, see e.g. [3], that studies and analyses listener preferences of loudspeakers in rooms. Methods for predicting listener preferences and certain timbral attributes of earphones have been described in e.g. [4] and [5]. A method for objective assessment evaluated on six different car audio systems is described in [6]. To the best of the authors' knowledge, however, there exists no prediction model derived from, or validated on, a large data set of vehicle audio systems. This is likely due to the significant costs and complexities associated with implementing such a program. The purpose of this paper is to present such a study.

The proposed virtual listener panel is intended as a complementary screening and decision-support tool, not as a replacement for final expert listening evaluation. Basic objective targets such as frequency response, noise floor, maximum SPL, and distortion remain useful engineering checks, but they do not by themselves estimate listener-score distributions or the degree of disagreement among listeners.

The long-term focus of our research is to accurately predict how a well-designed human listener panel would rate all relevant subjective attributes of an audio system, be it in a car, or at home, or using earphones. In this paper, we will limit our scope to prediction of timbral listener scores of car audio systems. What is novel, in addition to the large data set used to train and validate the method, is also how it predicts the distribution of ratings from a listener panel, not just the mean score. We will show examples of a predictor trained on 940 audio systems evaluated by trained listeners. The results of the predictor are evaluated on 237 audio systems. By showing performance results for training the model on subsets of the data collected, we highlight the importance of high-quality data at scale. We conclude with an outlook on future research and the complexities that need to be overcome to reach our goal of a measurement-based virtual listening panel.

2. Subjective and objective dataset

This work uses in-situ subjective evaluations of vehicle audio systems as ground truth targets for model training and evaluation. The overall in-situ evaluation method, facilities, and benchmarking workflow follow the long-running process documented in [1] The dataset has evolved over multiple years of use and covers a wide range of vehicles, configurations, and listening conditions.

A single training/evaluation sample corresponds to one vehicle audio system evaluation instance with:

- A measurement-derived representation used as model input. Objective measurements use a 6-microphone array at the driver-seat position while exciting the system with pink noise at multiple playback levels. For model training we use the average spectrum at 80 dBA. We also acquire several excitation modes that differ by the inter-channel relationship between loudspeakers: correlated, uncorrelated, and anticorrelated noise. The training pipeline that we describe in this paper uses the uncorrelated excitation mode.
- One or more listener-provided subjective scores used as targets (listener scores).
- Associated metadata required to link measurements, listeners, and evaluation context (site, vehicle under test, configuration). These are not used as input to the predictor per se.

Ground truth construction from listener scores
As listeners rate each attribute for each track listened to (see [1] for details), the dataset can represent subjective targets at different aggregation levels. In the following, the listener's score (the ground truth) of the audio system is taken to be the mean score from all tracks.

Distributional targets and mean-based evaluation
Listener ratings for the same vehicle evaluation form an empirical distribution rather than a single deterministic label. On average, each evaluation includes 10 to 15 listeners (Figure 1). In our training setup, both the model output and the ground-truth target are treated as approximated score distributions. This formulation allows the learning objective to incorporate scores uncertainty and inter-listener variability directly. The ratings used here are obtained from a standardized benchmarking protocol with trained listeners; further procedural details are given in the companion paper [1].

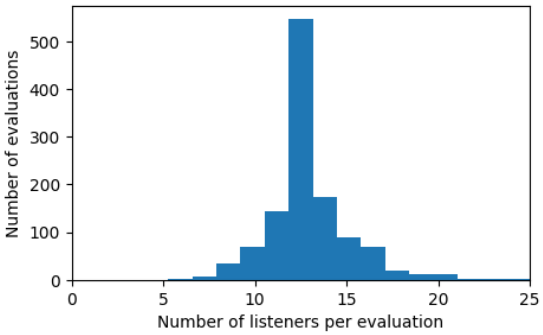


Figure 1. Distribution of number of listeners per evaluation

For many use cases, we are interested in the average listening panel score for a vehicle, and thus we often convert these distributions to a point estimate. Specifically, to report accuracies here, we use (i) the mean of the listener scores as the ground-truth reference and (ii) the mean of the model's sampled predicted distribution (an empirical estimate of the predictive expectation) as the prediction.

Figure 2 illustrates the processing flow used to connect spectral measurements to subjective scores. The pipeline starts from the input power spectral density (PSD) measured at a fixed playback level (80 dBA) (a). The set of listener scores is treated as a distribution (b, c), and the model predicts a score distribution that is compared to the ground-truth distribution (d); random samples are generated from the predicted distribution (e) to compute the predicted mean score (f), which is used for final scoring and metric evaluation. The distribution target is therefore empirical and listening-test-derived rather than a Bayesian posterior update; the model is trained to approximate this empirical listener-score distribution.

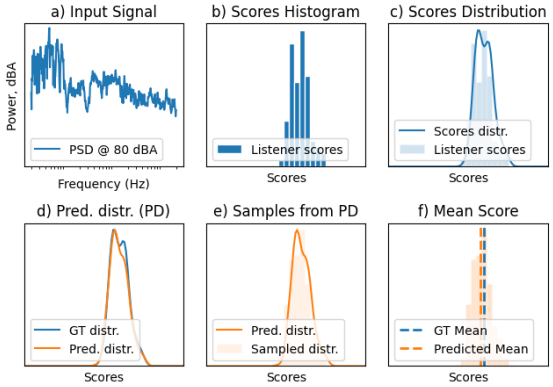


Figure 2. Processing from input spectrum to ground truth scalar value

Insufficiency of mean score. Although the mean score is a convenient scalar summary, it does not uniquely characterize perceived quality. Two systems can have the same mean listener score but substantially different score distributions, which implies different levels of listener agreement and different practical risk in decision-making.

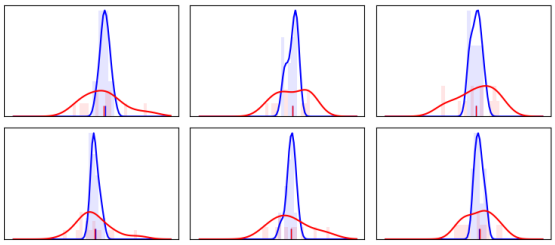


Figure 3. Examples of listener-score distributions with similar means but different shapes and dispersion.

Figure 3 illustrates this ambiguity: examples with nearly identical means can correspond to clearly different uncertainty structures and disagreement patterns. For this reason, using only mean-based targets would discard relevant information about label reliability and distribution shape.

To preserve this information, the learning objective is formulated at the distribution level. The model predicts score distributions, and these predictions are then compared with reference distributions derived from listener data. Figure 4 shows that the predicted distributions follow the main characteristics of the empirical targets, including central tendency, spread,

and asymmetry, which supports the use of distribution-aware training and evaluation.

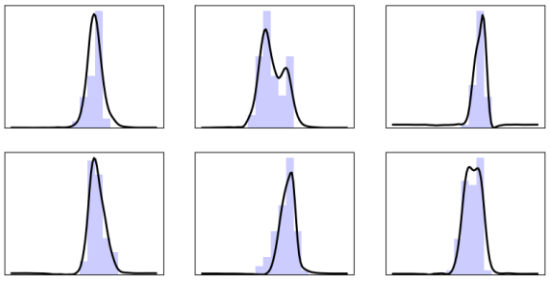


Figure 4. Comparison of model-predicted score distributions with reference listener-score distributions. The model captures not only the mean level but also the spread and distribution shape across representative cases.

Subjective score components

The listening protocol [1] captures multiple perceptual components, including spectral, spatial and dynamic dimensions. We also consider a composite defined as the arithmetic mean of the spectral, spatial, and dynamic components.

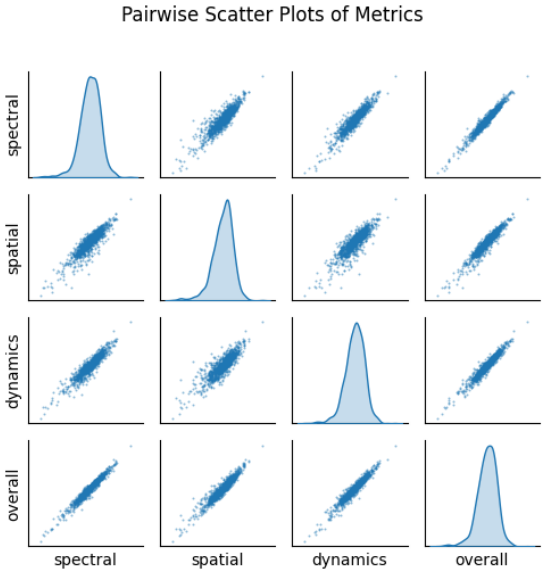


Figure 5. Scores correlation. Pairwise scatter matrix for subjective score components (spectral, spatial, dynamic) and the composite mean of spectral/spatial/dynamic.

Figure 5 shows pairwise relationships between these components. The spectral, spatial, and dynamic components exhibit strong positive, near-linear relationships with each other, indicating substantial shared variance within the dataset.

One possible explanation is that the dataset is dominated by production-intent vehicles, for which multiple perceptual attributes may have been tuned jointly rather than independently. Under this interpretation, systems with higher overall integration or quality would tend to receive higher scores across spectral, spatial, and dynamic dimensions simultaneously. This explanation, however, remains hypothetical and is not directly tested in the present study. Accordingly, the observed correlations should be interpreted as an empirical property of the dataset rather than evidence of a causal relationship among the attributes.

In this study, we use the subjective spectral sub-score as the primary training target. Under the observed correlation structure, this choice is expected to be representative of the broader "overall" subjective variation captured by the correlated components, while remaining aligned with the information content of the input features.

Dataset reliability

Listener variability implies that the dataset labels (i.e., the subjective scores) have a finite reliability, which in turn places an upper bound on achievable predictor performance of the mean score. To quantify this underlying uncertainty in the data, we show 95% confidence-interval (CI) widths of the sample means in Figure 6.

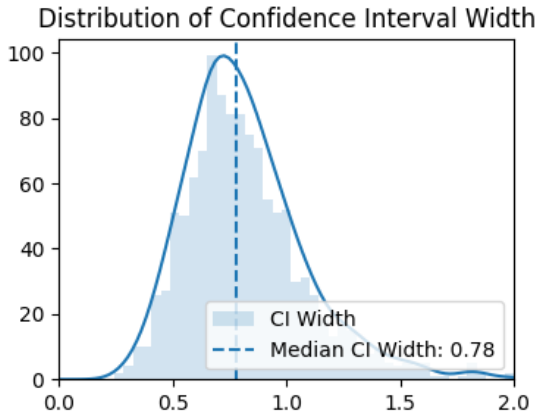


Figure 6. Overlaid distributions of per-sample uncertainty statistics computed from listener ratings: 95% confidence-interval width of the sample mean.

The CI widths were computed by treating the scores from different listeners as Gaussian independent and identically distributed random variables. In our dataset, the median CI width was found to be 0.78. The spread of individual scores on any specific car was quantified through calculating the sample standard deviation of the listener scores for each car.

The median standard deviation for all cars was found to be 0.62.

These uncertainty statistics provide a practical reference for interpreting model root mean square error (RMSE) when predicting the mean score. An RMSE on the order of the CI width can be considered relatively good, because it means that prediction error is comparable to the uncertainty of the estimated mean target itself.

3. Timbral Prediction

Model and Training Configurations.
This section describes the model architectures explored for predicting subjective quality scores from spectral measurements, summarizes how model quality is assessed during training and model selection.
Because the task is driven by subjective human ratings with non-negligible variability, we evaluate models using a combination of absolute-error metrics, order metrics, and distributional distance metrics. More on this in Section 4.

Model architectures. The architectures explored in this section are based on approaches patented in [7]. Two neural network architectures were investigated: a feed-forward (FF) regression model and a two-network autoencoder-style (AE) variant. Both approaches consume power spectral density (PSD) measurements as input and produce a predicted score distribution as output. In the current setup, the PSD input is represented as a 423-dimensional feature vector, and the score distribution is represented as a 100-dimensional vector. Figure 7 illustrates the key differences in their computational structure.
Feed-forward model (FF). The baseline configuration is a fully connected multi-layer perceptron (MLP) that directly maps the input PSD to a predicted score distribution. This architecture treats the prediction as a supervised regression task, minimizing the discrepancy between the predicted score distribution and the ground-truth score distribution. The FF model approach is characterized by its simplicity and training stability, enabling robust performance even with moderate dataset sizes.
Autoencoder model (AE): The second architecture uses two MLP blocks and an autoencoder-like training schedule. First, the forward mapping from score distribution to PSD is trained. Second, the inverse mapping from PSD to score distribution is trained. Third, a combined score-to-PSD-to-score pathway is optimized end-to-end with the inverse branch frozen. This three-stage schedule encourages internal consistency and was intended to improve distributional calibration. In our experiments, however, the added complexity did not yield sufficient gains in target metrics and evaluation results reported below are shown for the FF model, which on average demonstrated better performance than the autoencoder-style alternative across the main target metrics.

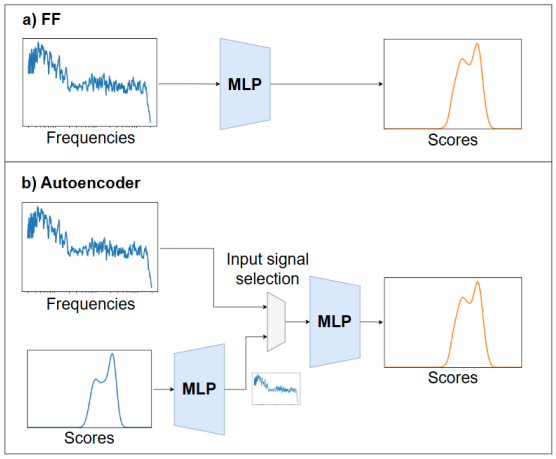


Figure 7. Model architectures comparison. (a) Feed-forward (FF) model: the input PSD is directly mapped to a predicted score distribution via a single MLP. (b) AE model: the approach uses separate processing pathways for PSD and score information, combining them to produce the predicted score distribution.

Architecture selection. The choice of these relatively simple architectures is motivated by practical dataset constraints. The complexity and cost of conducting high-quality listening panel evaluations lead to a fundamentally limited dataset size. Although the available dataset may be considered large in audio-engineering practice, it remains relatively small by contemporary machine-learning standards. This data scarcity requires architectures that generalize robustly from moderately sized training sets, rather than sophisticated approaches (e.g., large transformer models, deep residual networks, or complex ensemble systems) that typically demand orders of magnitude more labeled examples. Simpler, well-regularized models allow us to avoid overfitting and achieve stable, reproducible performance even when systematic dataset growth is slow.

4. Predictor performance

Point-estimate error on score means. Although training uses distributional targets, reporting and

many downstream decisions are often based on score means. Let y be the ground-truth mean score for an item and $\hat{y}$ be the predicted mean score (estimated from the model's predictive distribution).

- **RMSE**:

$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$. Lower is better.

Distributional similarity metrics. To quantify similarity between the full predicted score distribution and the ground-truth distribution, we additionally report distributional distances. Let p denote the ground-truth distribution and q be the predicted distribution over score bins.

- **KL divergence**: describes how much information is lost when the predicted distribution q is used to approximate the ground-truth distribution p, and it is particularly sensitive to mismatches in probability mass at specific bins. Lower is better.

$$D_{\mathrm{KL}}(p|q) = \sum_k p_k \log \frac{p_k}{q_k}$$

- **Wasserstein distance:** describes how far probability mass must be moved along the score axis to transform one distribution into the other, making it sensitive to the magnitude of score shifts. Lower is better.

$$W_1(p,q) = \sum_k |P_k - Q_k|\ \Delta x$$

where $P_k = \sum_{j \le k} p_j$ and $Q_k = \sum_{j \le k} q_j$ are cumulative distributions over bins. This metric describes how far probability mass must be moved along the score axis to transform one distribution into the other, making it sensitive to the magnitude of score shifts. Lower is better.

Pairwise and rank-based consistency. Many industrial use cases rely on *relative* comparisons (which system is better) rather than absolute score calibration. Therefore, we report the following order-preservation metrics:

- **Pairwise accuracy**: fraction of item pairs for which the ordering induced by predicted means matches the ordering induced by ground-truth means, i.e. $\mathrm{sign}(\hat{y}_i - \hat{y}_j) = \mathrm{sign}(y_i - y_j)$, (ties handled consistently). Higher is better.
- **Spearman's ρ**: rank correlation between $\hat{y}$ and y. Higher is better.
- **Pearson correlation**: linear correlation coefficient between $\hat{y}$ and y, measuring the strength of linear agreement between predicted and ground-truth mean scores. Values lie in $[-1,1]$, where 1 indicates perfect positive linear agreement. Higher is better.

Classification-oriented metrics

As an example of a typical industrial use case, we also evaluate the model's ability to perform binary classification: distinguishing systems of low quality from those of acceptable or high quality. This is motivated commercially: releasing a car with an audio system that will lead to customer complaints can be very costly. We define a binary label based on a quality threshold (e.g., mean score below 5 classified as "poor"), and measure classifier performance using receiver operating characteristic (ROC) curves (a plot that shows how a binary classifier's true positive rate changes against its false positive rate as the decision threshold varies) and ROC Area-Under-the-Curve (AUC). (AUC measures a binary classifier's ability to distinguish between classes, with values ranging from 0 to 1. An AUC of 0.5 suggests no discrimination (random guessing), while 1.0 represents perfect discrimination). In a production workflow, such a classifier could be used to flag high-risk tuning variants before committing to more expensive panel tests.

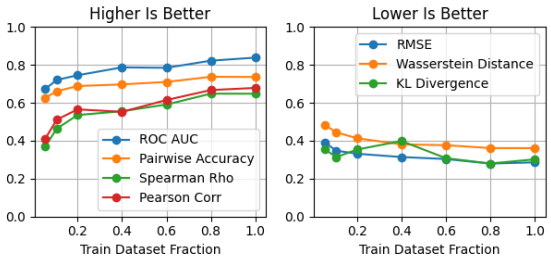


Figure 8. Test-set metrics as a function of training dataset fraction

Test-set comparison: To quantify the effect of data availability, we evaluate the same model family on a common held-out test set with 237 samples while varying the training dataset fraction from 0.05 to 1.0. Figure 8 summarizes the resulting trends.

Overall, metric behavior is consistent with improved generalization as more training data are used. Errors and distribution difference metrics decrease almost monotonically (RMSE drops from 0.39 to 0.29, KL divergence from 0.36 to 0.30, Wasserstein distance from 0.48 to 0.36), while ranking and association metrics increase (binary ROC-AUC from 0.67 to 0.84, pairwise accuracy from 0.63 to 0.74, Spearman's ρ from 0.37 to 0.65, and Pearson correlation from 0.41 to 0.68). Overall, the trajectories indicate stable gains rather than isolated

improvements at a single data point, while small non-monotonic deviations at individual training fractions can occur because each fraction is evaluated on a finite sampled subset of the training data.

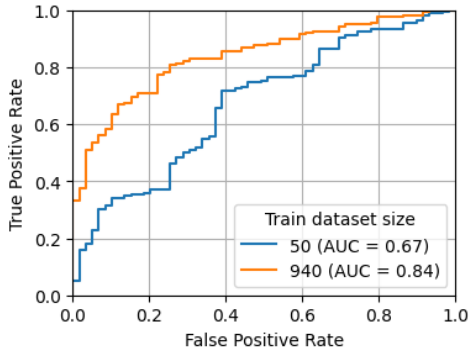


Figure 9. ROC curves for different training dataset sizes for binary classification of system quality (poor vs. acceptable/high). Blue: Model trained on 50 samples (AUC = 0.67). Orange: Model trained on 940 samples (AUC = 0.84).

Figure 9 shows ROC curves for two dataset sizes. The improvement reflects the benefit of increased training data size, enabling the model to better separate the two classes across varying operating points.

5. Conclusions

With this study, we highlight some of the inherent difficulties in predicting detailed listener ratings of car audio systems. It has been shown that it is possible – within certain fundamental limits – to predict how a listener panel would rate a sound system. However, the target of our study is intrinsically difficult: subjective audio-quality ratings are noisy; even for the same measured system the listener score distribution has non-zero variance. As a result, model errors should be interpreted relative to dataset self-consistency, not as a purely deterministic regression target. Further, this motivates an approach that focuses on predicting the likely distribution of scores from a listener panel.

Even with large data sets from a benchmarking process that is, in the authors' view, as carefully designed as one can practically hope for, there is an inherent limit to the resolution with which we can distinguish that one car audio system will be perceived as better-sounding than another. Our analysis shows that point estimates of the sound quality of car systems need to be treated with care. The predicted distribution of listener scores conveys considerably more information that needs to be accounted for in decision making. For example, our data show that two audio systems with the same mean score may still have significantly different individual scores. Perhaps, for a system with a reasonably good mean but high variance in the perceived sound, it is more important to work on system improvements that reduce the variation rather than increase the mean rating?

Our results further show that prediction performance depends strongly on dataset scale: increasing the amount of training data improves both score prediction accuracy and decision-oriented performance, including ranking consistency and detection of poor systems. Equally importantly, reliable validation requires sufficiently large and diverse test data, since average error alone can hide important performance differences across score ranges.

One of the limitations is that this study does not consider regional, demographic and other possible preference effects, such as bass-level preferences, therefore future work should include listener metadata for preference effects validation.

To summarize, our results suggest that virtual listener panels for automotive timbral assessment are feasible, but their reliability depends as much on benchmark design and dataset scale as on model architecture.

Future work will extend the approach to additional subjective attributes and incorporate more diverse training samples and objective measurements. The long-term goal is a robust virtual listener panel that can support not only assessment of automotive audio systems, but also their optimization.

References

[1] R. Kassier, D. Williamson, K. Kastrioti, B. Bazoun, J.-N. Schneider and J. Pierce, "A Practical Method for Routine In Situ Evaluation of Sound Quality in Vehicle Audio Systems", Presented at the AES 2026 International Conference on Automotive Audio, Detroit, MI, 2026 .

[2] F. E. Toole, Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms, Third Edition, Routledge (2018).

[3] S.E. Olive, "A Multiple Regression Model for Predicting Loudspeaker Preference Using

Objective Measurements: Part 2 - Development of the Model" 117th Convention, Audio Eng. Soc., Preprint 6190 (2004).

[4] S. Olive, T. Welti, O. Khonsaripour, "A Statistical Model that Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 2—Development and Validation of the Model." AES 143rd Convention, New York (2017)

[5] M. Herlufsen, N. Asp Fuglsang, B. Pedersen, S. Vase Legarth, "Establishing a Virtual Listener Panel for Audio Characterisation", AES Int. Conf. on Machine Learning and Artificial Intelligence for Audio, London (2025)

[6] M. Schäfer, "Sound Quality Assessment of Car Audio Systems" AES Int. Conf. on Automotive Audio, Dearborn (2022)

[7] Filimonov A. V., Kleshnin M. S., Kerbeneva A. Y., Lyons J., Olive S. E. "Methods and system for determining a sound quality of an audio system": patent application US20240196144A1; applicant and assignee Harman International Industries Inc. ; publ. 13.06.2024