ICIC Express Letters
Part B: Applications
Volume **14**, Number **10**, October **2023**

ICIC International ©2023 ISSN 2185-2766
pp. 1061–1066

SPARSE MODELING OF SOUND QUALITY IN VEHICLE AUDIO

Haruka Inoba[1], Shunsuke Ishimitsu[1] and Takahiro Soshi[2]

[1]School of Information Sciences
Hiroshima City University
3-4-1, Ozuka-Higashi, Asa-Minami, Hiroshima 731-3194, Japan
haruka.inoba@studium.uni-hamburg.de; ishimitu@hiroshima-cu.ac.jp

[2]Faculty of Foreign Language Studies
Mejiro University
4-31-1, Nakaochiai, Shinjuku-ku, Tokyo 161-8539, Japan
t.soshi@mejiro.ac.jp

Received January 2023; accepted April 2023

Abstract. *Sound quality evaluation for audio instruments is often conducted by employing professional evaluators known as "golden ears". However, the correlation between the evaluation results and physical features cannot be understood clearly, and the parts are objectively disjointed from the evaluation results. Hence, this study aims to link these two aspects associated with sound design. The following evaluation system is proposed: when the car's interior acoustic features contribute significantly to the sound quality and are input to the system, the audio evaluator's evaluation is provided as the output. In this study, car interior acoustic features were incorporated as parameters, and a sparse model was designed through machine learning. Three acoustic features in the frequency domain were incorporated: frequency magnitude, frequency phase, and frequency group delay. These features were calculated on the basis of impulse responses obtained under various car conditions. Based on the auditory impression evaluation in terms of clarity, sound localization, and spatial impression, the sparseness between the features and auditory impressions was extracted and modeled by applying machine learning.*

Keywords: Car interior, Audio, Sparse modeling, Auditory impression, Golden ears

1. **Introduction.** In conventional studies based on machine learning, some sound quality prediction system is proposed for the interior noise of cars. For instance, a method is proposed based on Deep Neural Networks (DNN) such as Laplacian Score-Deep Belief Network (LS-DBN) to evaluate the interior noise of electric vehicles automatically [1]. In addition, using Support Vector Machine (SVM) and genetic algorithm, a system is also proposed to evaluate interior noise which is radiated from automotive engine [2]. Neural network is often applied to constructing sound quality prediction systems as described. However, these ways need to be clarified, which affects sound quality.

Audio instruments are generally evaluated by a professional evaluator known as "golden ears" [3]. Their evaluations are considered essential to improve the sound quality of audio devices. For this reason, the sound quality will be enhanced if it reveals what affects their evaluations efficiently. Kvist et al. presented the correlation between the subjective listening test and the sound quality model [4]. It similarly specifies the factors contributing to the sound quality of audio instruments and the efficiency of various sound designs. Sakamoto et al. evaluated some audio equipment with wavelet transform to clarify the auditory impressions [5]. As previously described, experiments are arduous, and it is crucial to have a universalized and objective evaluation of sound quality. Furthermore, the eagerly anticipated emergence of artificial intelligence's "golden ears" has been long-awaited. That is why this study aims to link the subjective and objective evaluation,

DOI: 10.24507/icicelb.14.10.1061

identifying car interior acoustic features related to auditory impression based on the Least Absolute Shrinkage and Selection Operator (lasso). The following evaluation system is proposed. First, the car interior acoustic features that contribute significantly to the sound quality are entered into the system. Then, the score assigned by "golden ears" is provided as the output. Thus, the car interior acoustic features are incorporated as parameters, and a sparse model based on machine learning is designed. The diagram depicted in Figure 1 illustrates the input-output block diagram of the model proposed. By inputting acoustic features, a sound quality evaluation identical to that of an expert "golden ear" evaluator can be conducted. Although there have been previous studies on sound classification [8] or sound separation [9], there have been few applications to sound quality evaluation. This methodology facilitates the development of the "golden ears" AI, thus introducing a more streamlined evaluation system for assessing audio quality. The paper comprises five chapters, with the subsequent chapter detailing the analytical techniques employed in this investigation. Chapter 3 delineates the experimental procedures, whereas Chapter 4 expounds on the findings. Ultimately, Chapter 5 presents a summary of the outcomes.

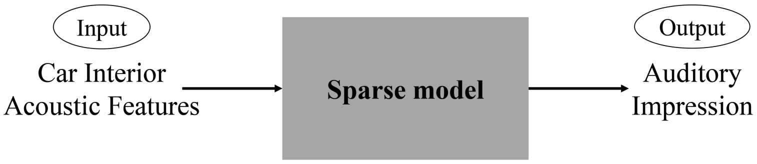

FIGURE 1. The proposed model

2. Method.

2.1. **Lasso regression.** Sparse modeling is defined as "technology to extract requisite parts from a statistical model according to given data" [10]. In this study, we applied lasso regression for model estimation to extracting car interior acoustic features. This method enables the selection of necessary variables. You can refer to lasso regression in Equation (1).

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_i^{(j)} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

Here, β is a coefficient to be estimated, $x^{(j)}$ is the j-th explanatory value and y_i is a measured value of objective value when explanatory value is $x_i^{(j)}$.

In 1996, lasso regression was initially supposed by Tibshirani [11]. Equation (1) employing the method of Lagrange multipliers, is commonly utilized [12]. Lasso regression consists of minimum mean square error and penalty term. Estimation and measured values are assigned into x_i and y_i, respectively. This part outputs β, which minimizes the mean square error between these variables. The penalty term is the most attractive point of lasso regression. This part has a role in shrinking β compared to the case without penalty term. At the same time, the whole of the equation is minimized, moreover, becoming zero in the case of unnecessary variables. In other words, the penalty term enables one to make a choice of the variables automatically. λ is called a hyperparameter, and it significantly impacts the results.

For this reason, it is essential to decide the optimal hyper parameter [13]. To estimate and evaluate this parameter, K-fold cross-validation is applied. The method is described in Section 2.2.

2.2. **K-fold cross-validation.** K-fold cross-validation is an evaluation method for model estimation. You can refer to the process of K-fold cross-validation in Figure 2.

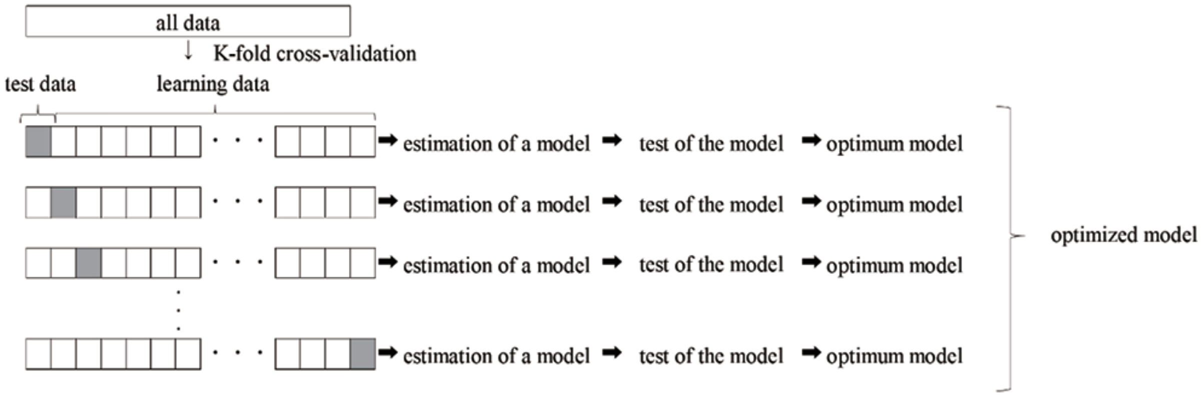

Figure 2. K-fold cross-validation

Firstly, all data is divided by K. 10 to 20 percent of the split data, and those left are used as test and learning data, respectively. As shown in Figure 2, the shaded and non-shaded cells represent the test and learning data, respectively. Subsequently, a model is estimated by learning data, and then test data evaluates the estimated model to determine whether it can predict well. Shifting the assignment of the test data, the model estimation and evaluation are conducted K times. After repeating the process, K-obtained results are averaged, and an optimized model is finally decided.

3. **Experimental Content.**

3.1. **Extraction of acoustic features.** This section describes how to extract acoustic features concretely. In Figure 3, you can refer to the process to extract acoustic features.

Firstly, "golden ears" scored sound quality in each car out of 100 according to three different auditory impression items. The vehicles are classified as "good" and "bad" depending on their scores in the evaluation. The cars obtaining high and low scores are classified as "good" and "bad". Subsequently, car interior acoustic features are calculated from impulse response in respective cars. The sample and the acoustic features from the data are set in vertical and horizontal lines, respectively. Furthermore, the individual matrix data classified as "good" and "bad" are labeled "1" and "2". These two matrix data are combined as one matrix data. The obtained matrix data are scaled in each feature, that is, along the vertical direction. This scaling aims to lose the variation among the samples and enable the extraction of acoustic features. Using this matrix data, the hyperparameter is optimized by K-fold cross-validation. In this validation, all data is divided by ten. In this study, 10 percent of the data and those of left are used as test and learning data, respectively. The most optimal hyperparameter is decided by ten times validation. Next, the labeled data is classified by lasso regression. Finally, the classification model is completed after the hyperparameter is determined when there is the slightest classification error. The model's classification accuracy is obtained from the sample rate classified according to the label.

Such a process is repeated 1000 times. After that, the number is counted as how many times each feature makes β zero. If the number is zero, the feature is considered extracted. Thus, the extracted features are related to auditory impressions. In other words, those extracted are the features affected by the classification.

3.2. **Experimental condition.** Clarity and spaciousness are known as the main factors which affect the sound quality of audio [6, 7]. For this reason, we focused on clarity, sound localization, and spatial impression in this experiment. "Golden ears" listened to sound sources in four cars and scored these out of 100 according to clarity, sound localization, and spatial impression. Based on the scores, the cars are grouped into "good" and "bad",

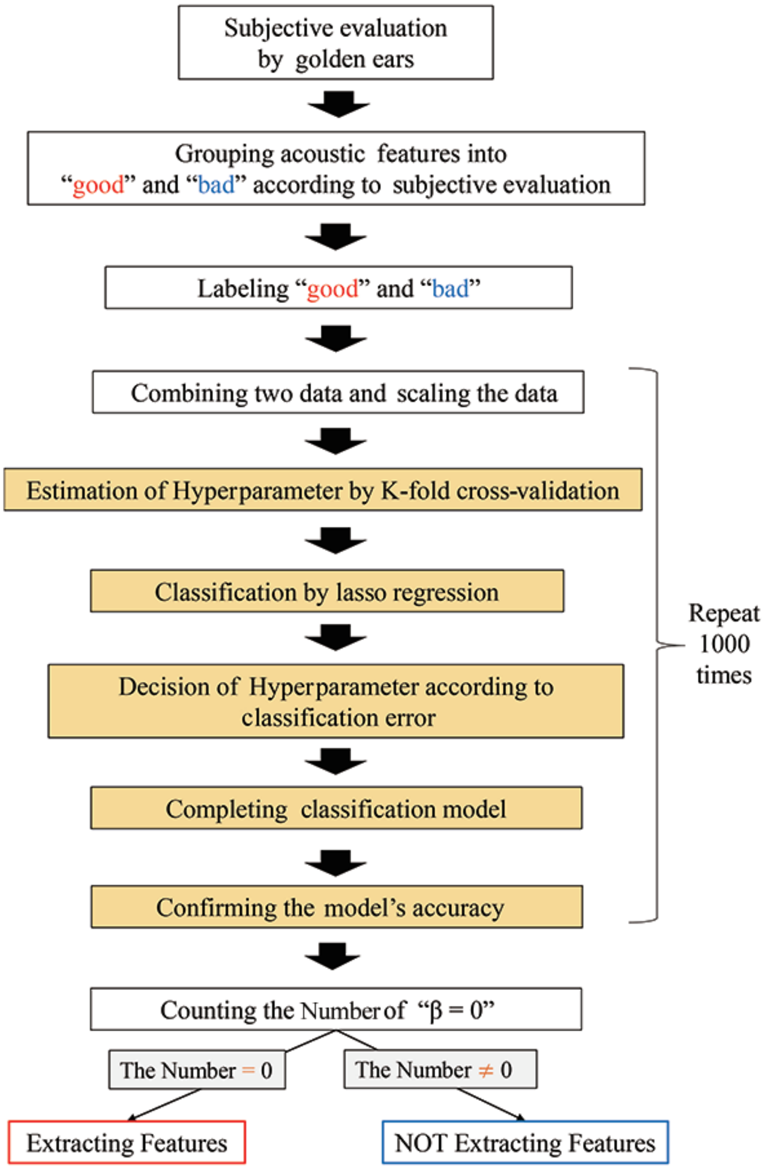
Subjective evaluation by golden ears
Grouping acoustic features into "good" and "bad" according to subjective evaluation
Labeling "good" and "bad"
Combining two data and scaling the data
Estimation of Hyperparameter by K-fold cross-validation
Classification by lasso regression
Decision of Hyperparameter according to classification error
Completing classification model
Confirming the model's accuracy
Repeat 1000 times
Counting the Number of "β = 0"
The Number = 0
The Number ≠ 0
Extracting Features
NOT Extracting Features

FIGURE 3. Process to extract acoustic features

and the data of acoustic features in each group are labeled. Although there are various acoustic features for evaluating the audio [14], we focused on frequency magnitude, phase, and group delay in this experiment. We applied the physical data calculated from impulse response since impulse response is known as a primary sound source to represent audio's characteristics [15]. Their input parameters consisted of the binaural, L-channel, and R-channel attributes, all obtained from the driver's seat. These parameters were utilized for the simultaneous playback of the front seat speakers, their L-channel-only playback, and their R-channel-only playback. Under such conditions, the experiment is conducted in 3 patterns of groups. The score of each car in different auditory impression categories is reported in Table 1. Here, the shaded cells represent scores classified as "good". The other scores are classified as "bad".

4. **Result.** The extracted acoustic features in each experiment are shown in Table 2 to Table 4. In Experiment 1, the cars were grouped based on their scores for clarity. According to Table 2, frequency magnitude was mainly extracted. With regard to the frequency phase, all features were extracted. On the other hand, the frequency group delay

Table 1. Auditory evaluation and grouping of each car in respective experiment

Experiment	Each evaluation of the cars				Auditory impression items
	Car 1	Car 2	Car 3	Car 4	
1	60	60	40	40	Clarity
2	50	40	–	–	Sound localization
	60	60	–	–	Spatial impression
3	50	40	40	30	Sound localization
	60	50	50	20	Spatial impression

Table 2. Extracted acoustic features in Experiment 1

Acoustic features	Played channel	Channel on microphones	Acoustic features	Played channel	Channel on microphones	Acoustic features	Played channel	Channel on microphones
Frequency magnitude	Both	Left	Frequency phase	Both	Left	Frequency group delay	Both	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right
	Left	Left		Left	Left		Left	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right
	Right	Left		Right	Left		Right	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right

Table 3. Extracted acoustic features in Experiment 2

Acoustic features	Played channel	Channel on microphones	Acoustic features	Played channel	Channel on microphones	Acoustic features	Played channel	Channel on microphones
Frequency magnitude	Both	Left	Frequency phase	Both	Left	Frequency group delay	Both	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right
	Left	Left		Left	Left		Left	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right
	Right	Left		Right	Left		Right	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right

Table 4. Extracted acoustic features in Experiment 3

Acoustic features	Played channel	Channel on microphones	Acoustic features	Played channel	Channel on microphones	Acoustic features	Played channel	Channel on microphones
Frequency magnitude	Both	Left	Frequency phase	Both	Left	Frequency group delay	Both	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right
	Left	Left		Left	Left		Left	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right
	Right	Left		Right	Left		Right	Left
		Right			Right			Right
		Left-Right			Left-Right			Left-Right

was not completely extracted. In Experiments 2 and 3, the cars were grouped according to sound localization and spatial impression. As you see in Table 3 and Table 4, frequency magnitude and phase were primarily extracted, while only two features were extracted regarding frequency group delay.

To sum up the above, there is the following tendency: frequency magnitude and phase were often extracted. Mainly, the frequency phase was noticeably extracted. On the other hand, frequency group delay was hardly extracted. The results indicate that frequency magnitude, particularly phase, is related to clarity, sound localization, and spatial impression. Alternatively, frequency group delays are not related to these auditory impressions.

5. **Conclusion.** To link the subjective and objective evaluation, we investigated constructing the sparse model, which outputs the evaluation by "golden ears" when acoustic features are input. As a first step, we extracted the features in the frequency domain. The results indicate that clarity, sound localization, and spatial impression depend on frequency magnitude, particularly phase, while auditory impressions are independent of frequency group delay. In further work, as a next step, focusing on the other car interior acoustic features and auditory impression items, we will extract the features as well as this study and build the system proposed at the beginning of this paper.

Acknowledgment. This work is partially supported by Mazda Motor Corporation and Pioneer Corporation. The authors gratefully acknowledge the constructive discussions with engineers working for these companies. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] B. Hai et al., The development of a deep neural network and its application to evaluating the interior sound quality of pure electric vehicles, *Mechanical System and Signal Processing*, vol.120, pp.98-116, 2019.

[2] L. Hai et al., Sound quality prediction for engine-radiated noise, *Mechanical System and Signal Processing*, vols.56-57, pp.277-287, 2014.

[3] J. Corey and D. H. Benson, *Audio Production and Critical Listening Technical Ear Traning*, 2nd Edition, Taylor & Francis, 2017.

[4] F. T. Agerkvist, P. Kvist et al., Development of a sound quality evaluation system, *The 117th Audio Engineering Society Convention*, 2004.

[5] K. Sakamoto, S. Ishimitsu, K. Sugawara, T. Yoshimi and K. Sasaki, The study of audio equipment evaluations using the sound of music, *ICIC Express Letters, Part B: Applications*, vol.2, no.3, pp.597-602, 2011.

[6] A. Gabrielson, Perceptual scaling and psychophysical relations in sound reproduction, *Proc. of the International Symposium on Subjective and Objective Evaluation of Sound*, pp.35-52, 1990.

[7] A. Gabrielson et al., Perceived sound quality of sound-reproducing systems, *The Journal of the Acoustical Society of America*, vol.65, no.4, pp.1019-1033, DOI: 10.1121/1.382579, 1979.

[8] H. R. Seresht and K. Mohammadi, Environmental sound classification with low-complexity convolutional neural network empowered by sparse salient region pooling, *IEEE Access*, vol.11, pp.849-862, 2023.

[9] M. Pezzoli, M. Cobos, F. Antonacci and A. Sarti, Sparsity-based sound field separation in the spherical harmonics domain, *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2022)*, pp.1051-1055, 2022.

[10] Y. Iba, Speed learning of model election, *Iwanami Data Science*, pp.6-18, 2017.

[11] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, vol.72, no.4, pp.417-473, 1996.

[12] S. Ikeda, Estimation using sparseness, *Iwanami Data Science*, vol.5, pp.19-35, 2017.

[13] S. Kawano, H. Matsui and K. Hirose, *Statistical Modeling via Sparse Estimation*, Kyoritsu Shuppan Co., Ltd., 2018.

[14] S. Koyano, Speaker for beginners, *Audio Engineering Society Audio Basic Seminar 2017*, pp.229-248, 2017.

[15] Y. Kaneda, Signals for impulse response measurement and measurement errors, *The Journal of the Acoustical Society of Japan*, vol.69, no.10, pp.594-554, 2013.