Statistical Classification of Vehicle Interior Sound Through Upsampling-Based Augmentation and Correction Using 1D CNN and LSTM

JINYOUNG KIM and JONGSOO LEE
Department of Mechanical Engineering, Yonsei University, Seoul 03722, South Korea
Corresponding author: Jongsoo Lee (jleej@yonsei.ac.kr)

This work was supported by the National Research Foundation of Korea, Domain Adaptation-Based Performance Prediction and Fault Diagnosis with Designable Data Augmentation and Transfer Learning, under Grant 2022R1A2C2011034.

ABSTRACT
To quantitatively classify the results provided by engineers, we built a sound quality (SQ) classification model using neural networks. The data used in this study were recorded wav files obtained from various vehicle specifications while driving under wide open throttle conditions. However, the data lengths were not constant. The upsampling and interpolation scheme (USIS) was used to achieve constant data lengths. After the USIS was applied to the data, dynamic time warping was used to verify that there was no change in the data characteristics. The verified dataset was transformed into Mel-spectrogram to confirm the characteristics, and dimensionality reduction was applied by using a high-pass filter. Clarifying the differences between clusters improves the model performance. The classification models of 1D convolutional neural network and long short-term memory exhibited training accuracies of about 94.9% (64 or 65 out of 68 classified) and test accuracies of about 87.5% (7 out of 8 classified). For additional undefined label classification, the quantitative evaluation and statistical classification of undefined sound quality labels are successfully identified in the present study. Both neural networks produced effective results that can be used by sound design engineers to quantitatively examine the SQ of vehicle interior noise.

INDEX TERMS
Upsampling and interpolation, dynamic time warping, mel-spectrogram, sound quality, classification.

I. INTRODUCTION
All vehicles have unique interior noises, and the characteristics of these noises are represented by sound quality (SQ). Nowadays, SQ has become a purchasing factor that vehicle buyers consider [1], [2]. Hence, vehicle manufacturers use active sound design (ASD) to highlight the SQ characteristics of the vehicle. A few studies have been conducted on SQ evaluation previously. Zwicker [3] suggested major SQ evaluation factors, such as loudness, sharpness, booming, roughness, fluctuation strength, and tonality. These factors made it possible to represent the vehicle sound quantitatively. SQ factors were extracted from the data to study the SQ evaluation prediction model of the user [4], [5], [6], [7].

Hanato and Hashimoto [8] proposed a new method based on power summation of the weighted 1/3 octave to find the frequency band that affects the quantitative booming index of SQ. Genuit [9] described how to collect SQ data accurately. Additionally, there were studies that classified vehicle interior noise using neural networks. Lee [10] and Wang et al. [11] conducted a study to predict the preferred SQ by a user through artificial neural networks (ANN). Tan et al. [12] created a prediction model for SQ evaluation using four SQ factors, user evaluation, and back-propagation neural network. Huang et al. [13] performed data denoising using a discrete wavelet transform and predicted user evaluation via a deep belief network. Lee et al. [14] created a model to design and generate a pleasant and powerful vehicle interior noise using SQ factors. Furthermore, various studies have used other features as indicators instead of SQ.
The major problems encountered in this study were inconsistent data lengths and insufficient data as per vehicle specifications. To solve the problem of inconsistent data lengths, the upsampling and interpolation scheme (USIS) was used. To verify whether the data characteristics were maintained, the data before and after conversion were compared along with the frequency components. The data were then converted to a Mel-spectrogram, and the mean-standard deviation distribution before and after the conversion was expressed as a scatter plot and compared with the scatter plot results obtained using the K-means algorithm. In this study, a 1D CNN and LSTM were used to perform stochastic classification of vehicle interior noise. To train a neural network with less data, the number of parameters in the neural network was reduced, as summarized in Fig. 1. The distinct characteristics of this study are as follows:

1) Upsampling and interpolation were used to match different data lengths. When the extracted wav data were digitized, the data length was adjusted by increasing the sampling point per second using USIS.

2) After employing USIS, the retention of data characteristics was verified using Euclidean distance comparison and dynamic time warping, major frequency domain comparison using cross-correlation, and comparison of mean-standard deviation distribution of Mel-spectrogram.

In this paper, Section II describes the data measurement method in detail. Section III describes the USIS and Mel-spectrogram conversion process. Section IV presents the structure of the two neural network models, method of outputting the results, and construction of training and test datasets. Section V details the probabilistic SQ classification results of the neural network according to the dataset and the related discussions thereon. Section VI highlights our conclusions.

II. MEASURED DATA

In total, 121 vehicle interior noises produced by various vehicles were measured. These vehicles consisted of sedans with different engines and vehicle grades. For generalization of vehicle interior noise measurement, the experiment was performed with limited driving conditions, such as wide open throttle (WOT). WOT was used because it is a condition where the characteristics of noise, vibration, and harshness (NVH) of the vehicle are revealed. WOT conditions were measured from 1000 rpm or less in a three-gear state until

![Flowchart of proposed USIS system.](image-url)
the accelerator pedal was fully depressed and shifted to the upper gear. However, it was difficult to extract noise data for WOT conditions below 1500 rpm. Hence, the data recorded were limited to 1500 – 5500 rpm. Vehicle interior noise was recorded using microphones (1/2 or 1/4 inch) installed on the left and right sides of the headset, and stored in the form of wav files through the frontend device.

Table 1 shows the classification of 112 vehicles based on the engine cylinder number. Table 2 shows the classification of 97 vehicles based on SQ. The obtained data had different lengths because the time required to reach 5500 rpm varies depending on the specifications of each vehicle. Therefore, to train the neural network, it is necessary to have a constant data length. Moreover, the data characteristics must be maintained.

III. DATA PREPROCESSING

A. UPSAMPLING AND INTERPOLATION SCHEME (USIS)

USIS were used to correct the data length. According to the Nyquist theory, the wav file was digitized by setting the sampling rate (SR) to 44,100 Hz (SR of 44,100 Hz means 44,100 samplings per second) [31]. Fig. 2 shows that the data lengths were different. If data A in Fig. 2(a) is approximately 11 s long, it consists of 11[s] · SR = 485,100 sample points. In Fig. 2(b), if data C is approximately 19.5 s long, it consists of 859,950 sample points. If data A has twice the sample points with minimal correction in length, the same 859,950 sample points can be obtained. Similarly, data B in Fig. 2(c) can also be length-corrected, and therefore data A, B, and C can have the same number of sample points. This process is called upsampling (US). The sample point per second can be increased by zero-padding and interpolation [32]. The procedure for USIS is shown in Fig. 3. First, zero-padding was done successively between existing sample points to double the length, as shown in Fig. 3(a). The padded points were derived using interpolation, as shown in Fig. 3(b). Thus, the data can be increased N times, and this is called the upsampling ratio (UR). Interpolation uses the firwin function and convolve of the SciPy module, a Python library. Scheme is applied according to the following process.

1. Filter data above audible frequency using firwin as a band-pass filter.
2. Zero-padding between original data points. (Fig. 3(a))
3. Based on the padded-zero value, create a window (length: NumTaps) and extract a point series (padded series, PS).
4. Extract the point series (original series, OS) to which zero-padding is not applied in the same section.
5. Convolution the two series and extract only the average value. And map the average value to zero. (Fig. 3(b))
6. Repeat the algorithm for all data sections.

The main parameters of the interpolation function firwin are NumTaps and Cutoff. NumTaps denotes the length of the filter, and it is a parameter that determines the number of points that need to be interpolated by combining them back and forth. Cutoff represents the frequency of the filter and has a range of [0, SR/2]. Both parameters can be expressed using (2) to (4) as follows:

\[ UR = \frac{\text{Length of the USIS signal}}{\text{Length of original signal}}, \]  
\[ \text{NumTaps} = 2 \times \left\lceil UR \right\rceil + 3, \]  
\[ \text{Cutoff} = \frac{\text{SR}}{(2 \times \text{SR})}, \]

The length of data A was doubled to obtain a length of 986,390 sample points. To match the augmented length of data A to the length of data C, 123,541 sample points must be removed from data A. As this is approximately 12% of the sample points in data A, the data loss is large. Therefore, the least common multiple of data length was calculated to minimize data loss. For example, if the length required is N times to match the length of 2,400,000 sample points, the value of N for data A is 5 times (2,465,975), 6 times for data B (2,530,308), and 3 times for data C (2,588,547). The cut out
To observe the Euclidean distance (ED) between two signals, the original and interpolated signals were divided and compared, as shown in Fig. 4(a) [33], [34]. Fig. 4(b) shows the original (red line) and interpolated (blue line) signals superimposed over each other, and it is evident that they are almost similar. Fig. 4(c) shows the ED between the two signals, which is less than 3% based on the original signal amplitude. Therefore, considering the interpolation error, the two signals are almost similar. The data used in the Fig. 4(c) is a value calculated according to the Euclidean distance formula of the original data and interpolated data, and ED can be expressed as follows:

$$ED = \sqrt{\text{original data} - \text{interpolated data}}^2,$$  \hspace{1cm} (5)

Second, dynamic time warping (DTW) was used to compare two signals. DTW is a method for measuring the similarity between two time series sequences [35], [36], [37]. Once the original and interpolated signals were similar each other, there is no difference between two signals so that the warping path should be the diagonal of matrix as shown in Fig. 4(d). The red line is the wrapping path that represents the perfect match of two signals whereas the black line is the wrapping path of the two signals, which has a quite similar tendency to the red line.

Third, the frequency component was compared using cross-correlation. Fig. 4(e) shows the frequency components of the original (red line) and interpolated (blue line) signals. For comparing the frequency components, the main region from 0 – 1200 Hz was enlarged and compared. Fig. 4(f) shows absolute value of the difference in magnitude between the main frequency components. From the figure, it was confirmed that the error was less than 3% based on the original signal frequency magnitude. An error of less than about 3% is a small. When the largest magnitude error is about 400, the error is 2.5%. Fig. 4(f) shows that most of the component errors show less than 150. This is a 0.94% error. Therefore, the frequency components were almost similar considering the interpolation error.

Next, the mean-standard deviation scatter plots of scale-converted Mel-spectrograms were compared. If the data characteristics are maintained after conversion, the scatterplot of the Mel-spectrogram mean – standard deviation should be similar. However, because of the interpolation error, the scatter plot is not perfect. The third verification will be described further using Mel-spectrograms in Section III.C.

### B. VERIFICATION OF USIS DATA

It is necessary to verify if the data obtained from USIS retained the original signal characteristics. Hence, the three methods were compared to determine the homogeneity of the data before and after conversion. First, the difference in signal amplitude was compared followed by comparisons of the frequency components and mean-standard deviation scatter plots of the data converted into Mel-spectrograms. If the data had the same characteristics after USIS, these three methods should not provide any significant difference considering the error.

First, the difference in signal amplitude was considered. To observe the Euclidean distance (ED) between two signals,
factors affecting SQ are represented by the change in dB at a specific frequency, and these characteristics are reflected in the data. As shown in Fig. 5, the Mel-spectrogram showed a prominent difference according to the SQ label. The Fig. 5 (a) showed low frequency components with high dB in all sections. However, the Figs. 5(b) and 5(c) showed high dB in all sections. From Fig. 5, it can be observed that the average dB of the entire section is high in the order of Figs. 5(b), 5(c), and 5(a), and the standard deviation of dB is also shown in the order of Figs. 5(b), 5(c), and 5(a).

The SR and N_FFT parameters were considered for creating the Mel-spectrograms. The Mel-spectrogram parameter SR is expressed as Mel_SR, which represents the number of samples per second. N_FFT denotes the length of the fast Fourier transform window. The original signal was expressed as a Mel-spectrogram, i.e., Mel_SR: 44,100 Hz, N_FFT: 22,050. However, after USIS, the parameter values changed, and were adjusted using (6) and (7).

\[
\text{Mel}_\text{SR} = 44,100 \times \lfloor UR \rfloor, \tag{6}
\]

\[
\text{N}_\text{FFT} = 22,050 \times \lfloor UR \rfloor \times \lfloor Const \rfloor, \tag{7}
\]

\(\cdots \text{N}_\text{FFT must be an integer}\)

The frequency range limitation, as shown in Fig. 6(a), limits the frequency range from 20 – 20,000 Hz to 20 – 1200 Hz.

TABLE 3. Sound quality components [3].

| SQ factors | Description |
|------------|-------------|
| Loudness   | The sound pressure level of a 1kHz tone in a plane wave and frontal incident that is as loud as the sound |
| Sharpness  | A measure of the high frequency content of a sound at a center frequency of 1kHz having a level of 60dB |
| Booming    | A measure of the low frequency content of a sound rather than high frequencies |
| Roughness  | Produced by a 1000Hz tone of 60dB which is 100% amplitude modulated at 70Hz |
| Fluctuation| quantifies subjective perception of slower (up to 20Hz) |
| Strength   | amplitude modulation of a sound |
| Tonality   | Produced by a 700Hz, Degree of pure tone |

\([Const]\) in (7) uses \{1, 1, 0.7, 0.54, 0.45, 0.37, 0.34\} according to the UR. If the UR exceeds 7, it is recommended that a lower common multiple be found. When the data were converted into a Mel-spectrogram, they were pre-processed for frequency range limitation and frequency resolution adjustment, so that the features were well revealed.

The frequency range limitation, as shown in Fig. 6(a), limits the frequency range from 20 – 20,000 Hz to 20 – 1200 Hz.
considering that the characteristics of SQ factors (Table 2) appear below 1,000 Hz. Although the SQ factors express features below 1000 Hz, the range was further increased to 1200 Hz to observe the features better. This resulted in the frequency range limitation showing a dimensional reduction effect on the data.

Fig. 6(b) shows the change in Mel-spectrogram according to the frequency resolution. The frequency resolution represents the extent to which a component value of the Mel-spectrogram expresses the frequency range. For example, in Fig. 6(b), if the frequency range of 20 – 1200 Hz was divided into 100 partitions, a frequency of approximately 11.8 Hz range (20 Hz – 31.8 Hz) would be available per partition if the 20 – 1200 Hz section was divided into more partitions, the frequency resolution would be much higher.

If the frequency range of 20 – 1200 Hz was divided into 500 partitions, i.e., a frequency of approximately 2.35 Hz per column (20 Hz – 22.35 Hz) would be expressed in dB. However, a high frequency resolution is not always appropriate. The reason for this can be observed in the scatter plot shown in Fig. 7. A mean-standard deviation cluster of data has to be formed. However, if the frequency resolution is too low or too high, no clear data cluster can be formed.

The mean-standard deviation distributions of the data divided into 200 (Fig. 7(a)), 250 (Fig. 7(b)), 350 (Fig. 7(d)), and 400 (Fig. 7(e)) partitions were ambiguous between the three classes. However, the scatter plot divided into 300 (Fig. 7(c)) partitions clearly showed boundaries between classes except for one powerful scatter.

In this study, verification of the results with the mean-standard deviation cluster was compared with the
classification results according to the mean-standard deviation using K-means. In Fig. 8(a) and 8(b), it can be seen that the clusters are similar when the scatter plots simultaneously show the mean-standard deviation distribution of the data before and after the transformation. The minor differences in the scatter plot arise due to the errors in interpolation and data editing process, as mentioned in Section 3.2. When the mean value was slightly different in the data editing process, the signal data after the high amplitude were cut off, which resulted in the change in average value. The standard deviation changed slightly due to the influence of interpolation error. Fig. 8(c) shows the classification of data using K-means. Apart from the data considered as abnormal classifications, such as 1 and 2, it was found that the classification areas are similar.

By comparing the scatter plot shown in Fig. 8, two observations were made. First, by comparing the classification made (Fig. 8(d)) by the engineer’s subject with respect to the quantitative classification result, there was a minor difference and reliability. Second, from No. 1 in Fig. 8(c), it can be confirmed that there was a difference between the subjective classification of engineers and quantitative classification obtained using K-means.

D. NORMALIZATION
The recorded data were influenced by weather conditions, resulting in minor differences. Therefore, to correct these differences, the data were subjected to rescaling considering the data characteristics. There are two types of rescaling: standardization and normalization. Of the two, normalization was used to convert the data range to $0 – 1$ to reduce the effect of relative size between the data because the magnitude of dB influences the classification. The expression for normalization is shown in Eqn. (8).

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

($x_{min}$: minimum value, $x_{max}$: maximum value)

IV. ARCHITECTURE OF NEURAL NETWORKS
In this study, a 1D CNN and LSTM were used for recognition and time series data processing, respectively. The structures of both the neural networks are shown in Figs. 9(a) and 9(b).

These neural networks are used as powerful tools for speech recognition and natural language processing. The performance of these neural networks was compared to find a suitable neural network for the SQ classification model. Because the dataset is limited and biased, the size of the neural network was designed to be small to prevent overfitting. The data has a high dimension [300, 3907], but as few classes are used, the number of layers was minimized to reduce the number of parameters. The neural networks were classified probabilistically using the Softmax function. As shown in Table 8 and IX, the output is the probability of the data being classified into each class (Luxury, Powerful, and Sporty).

V. RESULTS
A. CASE #1 – ENGINE CYLINDER CLASSIFICATION
The data structure of Case #1 is shown in Table 1. Most of the data were biased towards 4-cylinder and 6-cylinder engines. This is because most of the vehicles manufactured nowadays have either 4-cylinder or 6-cylinder engines. Table 4 and Table 5 show the classification results obtained using the 1D...
CNN and LSTM, respectively. In these Tables, the x-axis is the actual label and y-axis is the predicted label. The 1D CNN showed a high learning rate, whereas LSTM showed a relatively low learning performance of 84%. With respect to 3-cylinder and 8-cylinder, the learning performance of LSTM was low due to inadequate data, whereas the 1D CNN showed high learning performance with the same data and was able to classify the engine cylinder accurately. LSTM had poor classification performance for 8-cylinder in both training and testing. However, the classification performance of LSTM for 4-cylinder and 6-cylinder with a large amount of data was 90.91%, which was higher than that of the 1D CNN (72.73%). The following conclusions can be drawn from Case #1:
TABLE 8. Result of case #3 classification using 1D CNN.

| Data   | Luxury | Powerful | Sporty | Label  | Classification |
|--------|--------|----------|--------|--------|----------------|
| Data A | 0.995  | 0.0002   | 0.0048 | Luxury | Luxury         |
| Data B | 0.9939 | 0.0001   | 0.006  | Luxury | Luxury         |
| Data C | 0.9948 | 0.0003   | 0.0049 | Luxury | Luxury         |
| Data D | 0.0013 | 0.824    | 0.1747 | Powerful| Powerful       |
| Data E | 0.0026 | 0.0059   | 0.9914 | Sporty  | Sporty         |
| Data F | 0.0026 | 0.0059   | 0.9914 | Sporty  | Sporty         |
| Data G | 0.0024 | 0.2943   | 0.7034 | Sporty/Powerful | Sporty |
| Data H | 0.0007 | 0.9823   | 0.017  | Powerful| Powerful       |
| Data I | 0.4677 | 0.0033   | 0.529  | Luxury/Sporty | Sporty |
| Data J | 0.2138 | 0.0033   | 0.7829 | Luxury  | Sporty         |
| Data K | 0.0003 | 0.9881   | 0.0115 | Powerful| Powerful       |
| Data L | 0.0004 | 0.9875   | 0.012  | Powerful| Powerful       |
| Data M | 0.0037 | 0.0052   | 0.9911 | Sporty  | Sporty         |
| Data N | 0.995  | 0.0003   | 0.0048 | Luxury  | Luxury         |
| Data O | 0.0049 | 0.0044   | 0.9907 | Sporty  | Sporty         |
| Data P | 0.0118 | 0.0028   | 0.9853 | Sporty  | Sporty         |
| Data Q | 0.9926 | 0.0001   | 0.0073 | Luxury  | Luxury         |
| Data R | 0.0016 | 0.0095   | 0.9889 | Powerful| Sporty         |
| Data T | 0.0017 | 0.0084   | 0.9899 | Sporty/Powerful | Sporty |
| Data U | 0.0002 | 0.9906   | 0.0092 | Powerful| Powerful       |

1) 1D CNN has good overall classification performance even with less data.

2) LSTM tends to underperform in learning with less data. However, it shows high classification performance when a sufficient amount of data is available.

B. CASE #2 – ENGINE EMOTION CLASSIFICATION

The data structure of Case #2 is shown in Table 2. The composition of the data was distributed evenly among the three classes. Table 6 and Table 7 show the result of SQ classification using the 1D CNN and LSTM, respectively. Unlike Case #1, LSTM showed a high learning performance.

However, both neural networks showed some differences in their learning performances. In the 1D CNN, the learning performance with respect to Luxury and Sporty vehicles was poor, and the classification of sporty vehicles in the test was somewhat poor. However, both neural networks showed some differences in their learning performances. In the 1D CNN, the learning performance with respect to Luxury and Sporty vehicles was poor, and the classification of sporty vehicles in the test was somewhat poor. However, for LSTM, the learning performance with respect to Powerful and Sporty vehicles was poor, and learning of Sporty vehicles was poor in the test.

The conclusions drawn from Case #2 are as follows:

1) 1D CNN and LSTM had low classification performance at the neighboring classification boundary, as shown in Fig. 8(c).

2) There was no confusion between Luxury and Powerful vehicles where data clusters were not close. This shows that the mean-standard deviation plot of the Mel-spectrogram after USIS is valid.

The learning performance of the 1D CNN and LSTM was 94.12% and 95.59%. However, the classification performance of both networks was the same at 87.5%.

C. CASE #3 – UNDEFINED SQ LABEL CLASSIFICATION

The data of Case #3 are listed as Undefined in Table 2. Fig. 10 shows a result in which undefined data (black) are projected on the scatter plot of mean-standard deviation of the Mel-spectrogram (Fig. 8(c)). Undefined datasets were classified using the two neural networks, as described in section V.B. Table 8 and Table 9 show the 1D CNN and LSTM results, respectively. In the 1D CNN results, J was closer to the distribution of Luxury but was classified as Sporty. Additionally, M, S, and T were in the Powerful and Sporty boundaries, and classified as Sporty by the K-means algorithm, as shown in Fig. 8(d). In the LSTM results, J was classified as Sporty although the adjacent scatter points were classified as Luxury. M and P were on the classification boundary between Sporty and Powerful, but they were misclassified as Luxury. Furthermore, LSTM classified T as Sporty. Among M, S, and T, which caused confusion in 1D CNN, M was incorrectly classified as Luxury and S as Powerful, unlike 1D CNN.

However, T was classified as Sporty. The conclusion drawn from Case #3 are as follows:

1) Neural network data at the boundary of each class do not have the same accurate definitions as the ASD experts. However, the present the classification probability for each class.

2) Although there was a slight difference in learning performance, the results of undefined data showed similar performance in both neural networks. In Fig. 10, the two neural networks are misclassified at the interface. If both had poor classification performance, both neural networks would have misclassified data far from the classification boundary.

VI. DISCUSSION

This study provides ASD engineers with a probabilistic evaluation method for SQ classification of vehicle interior noise.
This model enables ASD engineers to quantitatively evaluate the criteria for SQ. The critical aspect of this study was the use of USIS to produce data in a suitable form that enabled neural networks to learn. The preprocessed data were verified by comparing the ED, the DTW between signals and cross-correlation coefficients. The verified datasets were converted by adjusting the Mel-spectrogram parameter according to the UR in the USIS process. Additionally, the data were preprocessed to limit the frequency domain and adjust the frequency resolution to show the data cluster clearly. This cluster was compared with the classification results using K-means. Except for No. 1, and a few scatters in Fig. 8(d), most of the clusters formed were similar. Two datasets were provided based on engine cylinder number and SQ. The SQ dataset was not a numerical classification of SQ factors, such as loudness. However, it was classified by a veteran ASD engineer. In this study, the 1D CNN showed better classification performance for classes with less data, e.g., Case #1 (3-cylinder and 8-cylinder). However, LSTM had better classification performance for classes with a large amount of data (4-cylinder and 6-cylinder). In Case #2, there was no difference in performance between the 1D CNN and LSTM. The numbers of data points for 3-cylinder and 8-cylinder were 13 and 11, respectively, but that of Case #2 exceeded 22. Compared to Case #1, the number of data for Case #2 was about 2 times more. If 20 data points per class are used, it is expected that performance will be improved. We believe that verification will be possible if additional training data of 3-cylinder and 8-cylinder can be obtained in a future study. Case #3 revealed two things. First, as shown in Tables 8 and 9, the classification was incorrect or unclear due to the class boundary. A neural network classification model that quantitatively classifies data also had difficulty in classifying SQ data accurately at the class boundary. Second, LSTM had a completely opposite classification compared to that of the 1D CNN for cases like Figs. 11(a) and 11(b). In the case of Figs. 11(c) and (d), the classification was not clear because these cases were at the class boundary. However, in the case of the 1D CNN, the classification was not wrong. There was a probabilistic unclear result at the class boundary. The 1D CNN classified Fig. 11(a) and S as Sporty, as shown in Table 8, and LSTM classified Figs. 11(a) and (b) as Luxury, as shown in Table 9. Fig. 5 is a representative Mel-spectrogram of Luxury, Powerful, and Sporty. Figs. 11(b), 11(c), and 11(d) show the Mel-spectrogram of Figs. 11(a), 11(b), 11(e), respectively. All three data points were at the border between Powerful and Sporty. Additionally, it is difficult to classify Luxury and Sporty, as shown in Fig. 5. The Sporty class was characterized by having a slightly lighter line than that of the Powerful class, and the 1D CNN seemed to classify Sporty based on this feature. LSTM, which specializes in time series data, appeared to classify based on the characteristics of changes in frequency components over time. From Fig. 11(b) and 11(c), it can be observed that the frequency components gradually increased with time.
VII. CONCLUSION

The purpose of this study was to propose a valid data preprocessing and verification method for data transformed using USIS to obtain a common data length for the recorded data to make them suitable for neural network learning. The results of this study prove that the USIS can be used as a tool for studies related to vehicle interior noise SQ classification.

The classification performance of the 1D CNN and LSTM were different depending on the amount of data. However, LSTM showed better classification performance with a large amount of data. The 1D CNN tends to incline towards Sporty, and LSTM towards Luxury. Therefore, the classification result of boundary data should be judged by considering data characteristics. The overall neural network performance was good with some exceptions. The learning and classification performances of the 1D CNN in Case #1 (Engine Cylinder) were 93.20% and 75%, and 94.12% and 87.50% in Case #2 (Sound Quality), respectively. The learning and classification performances of LSTM in Case #1 were 81.56% and 68.75%, and 95.59% and 87.50% in Case #2, respectively. Based on the neural network performance results in Case #1 and Case #2, the quantitative evaluation and classification of undefined sound quality labels are successfully identified in the present study.

For the further direction, the classification performance is to be conducted in the frequency domain as well as the time domain. Additionally, the more active study on the sampling techniques to handle the various lengths of sound data in the context of active sound design (ASD) engineering.

ACKNOWLEDGMENT

This study was conducted with the support from the Sound Design Research Laboratory, Research and Development Division of Hyundai-Kia Motors.

REFERENCES

[1] B. H. Vrkljan and D. Anaby, “What vehicle features are considered important when buying an automobile? An examination of driver preferences by age and gender,” J. Saf. Res., vol. 42, no. 1, pp. 61–65, 2011.
[2] R. J. Hafner, I. Walker, and B. Verplanken, “Image, not environmentalism: A qualitative exploration of factors influencing vehicle purchasing decisions,” Transp. Res. A, Policy Pract., vol. 97, pp. 89–105, Mar. 2017.
[3] E. Zwicker and H. Fastl, Psychoacoustics: Facts and Models, vol. 22. Heidelberg, Germany: Springer, 2013.
[4] R. Bisping, “Car interior sound quality: Experimental analysis by synthesis,” Acta Acustica United Acustica, vol. 83, no. 5, pp. 813–818, 1997.
[5] G. Kwon, H. Jo, and Y. J. Kang, “Model of psychoacoustic sportiness for vehicle interior sound: Excluding loudness,” Appl. Acoust., vol. 36, pp. 16–25, Jul. 2018.
[6] H. B. Huang, J. H. Wu, X. R. Huang, W. P. Ding, and M. L. Yang, “A novel interval analysis method to identify and reduce pure electric vehicle structure-borne noise,” J. Sound Vib., vol. 475, no. 9, 2020, Art. no. 115258.
[7] L. Liang, S. Chen, and P. Li, “The evaluation of vehicle interior impact noise inducing by speed bumps based on multi-features combination and support vector machine,” Appl. Acoust., vol. 163, Jun. 2020, Art. no. 107212.
[8] S. Hanato and T. Hashimoto, “Booming index as a measure for evaluating booming sensation,” in Proc. Inter-Noise, vol. 233, 2000, pp. 1–5.
[9] K. Genuit, “The sound quality of vehicle interior noise: A challenge for the NVH-engineers,” Int. J. Vehicle Noise Vib., vol. 1, nos. 1–2, pp. 158–168, 2000.
[10] S. K. Lee, “Objective evaluation of interior sound quality in passenger cars during acceleration,” J. Sound Vib., vol. 310, nos. 1–2, pp. 149–168, 2008.
[11] Y. S. Wang, G. Q. Shen, and Y. F. Xing, “A sound quality model for objective synthesis evaluation of vehicle interior noise based on artificial neural network,” Mech. Syst. Signal Process., vol. 45, no. 1, pp. 255–266, 2014.
[12] G. P. Tan, D. F. Wang, and Q. Li, “Vehicle interior sound quality prediction based on back propagation neural network,” Proc. Environ. Sci., vol. 11, pp. 471–477, Jan. 2011.

[13] H. B. Huang, X. R. Huang, R. X. Li, T. C. Lim, and W. P. Ding, “Sound quality prediction of vehicle interior noise using deep belief networks,” Appl. Acoust., vol. 113, pp. 149–161, Dec. 2016.

[14] S. M. Lee, J. Back, K. An, and S. K. Lee, “Design and generation of a target sound to achieve the desired sound quality inside a car cabin,” Int. J. Automot. Technol., vol. 21, no. 2, pp. 385–395, 2020.

[15] A. Aljaafreh and L. Dong, “An evaluation of feature extraction methods for vehicle classification based on acoustic signals,” in Proc. Int. Conf. Netw., Sens. Control (ICNSC), Chicago, IL, USA, 2010, pp. 570–575.

[16] M. Kandpal, V. K. Kakar, and G. Verma, “Classification of ground vehicles using acoustic signal processing and neural network classifier,” in Proc. Int. Conf. Signal Process. Commun. (ICSC), Noida, India, 2013, pp. 512–518.

[17] H. B. Huang, R. X. Li, X. R. Huang, M. L. Yang, and W. P. Ding, “Sound quality evaluation of vehicle suspension shock absorber rattling noise based on the Wigner–Ville distribution,” Appl. Acoust., vol. 100, pp. 18–25, Dec. 2015.

[18] B. Selbes and M. Sert, “Multimodal vehicle type classification using convolutional neural network and statistical representations of MFCC,” in Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Survell. (AVSS), Lecce, Italy, Sep. 2017, p. 1–6.

[19] C. Daniel and L. Mary, “Fusion of audio visual cues for vehicle classification,” in Proc. Int. Conf. Next Gen. Intell. Syst. (ICNGIS), Kottayam, India, 2016, pp. 1–4.

[20] J. George, L. Mary, and K. S. Riyas, “Vehicle detection and classification from acoustic signal using ANN and KNN,” in Proc. Int. Conf. Control Commun. Comput. (ICCC), Thrivunanthapuram, India, 2013, pp. 436–439.

[21] L. Hertel, H. Phan, and A. Mertins, “Classifying variable-length audio files with all-convolutional networks and masked global pooling,” 2016, arXiv:1607.02857.

[22] V. Cevher, R. Chellappa, and J. H. McClellan, “Vehicle speed estimation using acoustic wave patterns,” IEEE Trans. Signal Process., vol. 57, no. 1, pp. 30–47, Jan. 2009.

[23] G. P. Mazarakis and J. N. Avaritsiotis, “Vehicle classification in sensor networks using time-domain signal processing and neural networks,” Microprocess. Microsyst., vol. 31, no. 6, pp. 381–392, 2007.

[24] H. Wu and J. M. Mendel, “Classifier designs for binary classifications of ground vehicles,” Proc. SPIE, vol. 5090, pp. 122–133, Sep. 2003.

[25] A. Schindler, T. Lidy, and A. Rauber, “Multi-temporal resolution convolutional neural networks for acoustic scene classification,” in Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE), Munich, Germany, Nov. 2017.

[26] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” Biomed. Signal Process. Control, vol. 47, pp. 312–323, Jan. 2019.

[27] G. Sharma, K. Uma pathy, and S. Krishnan, “Trends in audio signal feature extraction methods,” Appl. Acoust., vol. 158, 2020, Art. no. 107020.

[28] C.-H. Lee, J.-S. Iwo, H.-Y. Hsieh, and C.-S. Lin, “An intelligent system for grinding wheel condition monitoring based on machining sound and deep learning,” IEEE Access, vol. 8, pp. 58279–58289, 2020.

[29] F. Demir, D. A. Abdullah, and A. Sengur, “A new deep CNN model for environmental sound classification,” IEEE Access, vol. 8, pp. 66529–66537, 2020.

[30] T. Tran and J. Lundgren, “Drill fault diagnosis based on the scalogram and mel spectrogram of sound signals using artificial intelligence,” IEEE Access, vol. 8, pp. 203655–203666, 2020.

[31] A. J. Jerri, “The Shannon sampling theorem—Its various extensions and applications: A tutorial review,” Proc. IEEE, vol. 65, no. 11, pp. 1565–1596, 1977.

[32] L. Tan and J. Jiang, Digital Signal Processing: Fundamentals and Applications. Cambridge, MA: Academic, 2018.

[33] A. Gogolou, T. Tsandilas, T. Palpanas, and A. Bezerianos, “Comparing similarity perception in time series visualizations,” IEEE Trans. Vis. Comput. Graphics, vol. 25, no. 1, pp. 523–533, Jan. 2019.

[34] K. M. Wong, J. Zhang, J. Liang, and H. Jiang, “Mean and median of PSD matrices on a Riemannian manifold: Application to detection of narrow-band sonar signals,” IEEE Trans. Signal Process., vol. 65, no. 24, pp. 6536–6550, Dec. 2017.

[35] Y.-T. Liu, Y.-A. Zhang, and M. Zeng, “Adaptive global time sequence averaging method using dynamic time warping,” IEEE Trans. Signal Process., vol. 67, no. 8, pp. 2129–2142, Apr. 2019.

[36] S. Giraldo, A. Ortega, A. Perez, R. Ramirez, G. Waddell, and A. Williamson, “Automatic assessment of violin performance using dynamic time warping classification,” in Proc. 26th Signal Process. Commun. Appl. Conf. (SIU), May 2018, pp. 1–3.

[37] H. Li and C. Wang, “Similarity measure based on incremental warping window for time series data mining,” IEEE Access, vol. 7, pp. 3909–3917, 2018.

[38] S. Lu, P. Zhou, X. Wang, Y. Liu, F. Liu, and J. Zhao, “Condition monitoring and fault diagnosis of motor bearings using undersampled vibration signals from a wireless sensor network,” J. Sound Vib., vol. 414, pp. 81–96, Feb. 2018.

[39] L. A. L. Janssen and I. L. Artega, “Data processing and augmentation of acoustic array signals for fault detection with machine learning,” J. Sound Vib., vol. 483, Sep. 2020, Art. no. 115483.

JINYOUNG KIM received the B.S. degree in mechanical engineering from Inha University, Incheon, South Korea, in 2017, and the M.S. degree in mechanical engineering from Yonsei University, Seoul, South Korea, in 2020. His current research interests include multi-physics design optimization, data-driven prognostics, and health management.

JONGSOO LEE received the B.S. and M.S. degrees in mechanical engineering from Yonsei University, Seoul, South Korea, in 1988 and 1990, respectively, and the Ph.D. degree in mechanical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1996. He was a Research Associate at the Rensselaer Rotocraft Technology Center. He is currently a Professor of mechanical engineering at Yonsei University. His research interests include multi-physics design optimization (MDO), reliability-based robust engineering design, prognostics and health management (PHM), industrial artificial intelligence, and machine learning with applications to structures, fatigue/durability, lifetime prediction, noise, and vibration problems.