www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

ARCHIVES OF ACOUSTICS Vol. 49, No. 4, pp. 527–542 (2024), doi: 10.24425/aoa.2024.148816

Research Paper

Sound Quality Prediction Method of Dual-Phase Hy-Vo Chain Transmission System Based on MFCC-CNN and Fuzzy Generation

Jiabao LI, Lichi AN, Yabing CHENG*, Haoxiang WANG

School of Mechanical and Aerospace Engineering, Jilin University
Changchun, Jilin, China

*Corresponding Author e-mail: chengyb@jlu.edu.cn

(*received January 6, 2024; accepted July 1, 2024; published online October 21, 2024*)

The sound quality of transmission system noise significantly impacts user experience. This study aims to predict the sound quality of dual-phase Hy-Vo chain transmission system noise using a small sample size. Noise acquisition tests are conducted under various working conditions, followed by subjective evaluations using the equal interval direct one-dimensional method. Objective evaluations are performed using the Mel-frequency cepstral coefficient (MFCC). To understand the impact of the MFCC order and the frame number on prediction accuracy, MFCC feature maps of different specifications are analyzed. The dataset is expanded threefold using fuzzy generation with an appropriate membership degree. The convolutional neural network (CNN) is developed, utilizing MFCC feature maps as inputs and evaluation scores as outputs. Results indicate a positive correlation between the frame number and prediction accuracy, whereas higher MFCC orders introduce redundancy, reducing accuracy. The proposed CNN method outperforms three traditional machine learning approaches, demonstrating superior accuracy and resistance to overfitting.

Keywords: sound quality; dual-phase transmission; Hy-Vo chain; MFCC; fuzzy generation.

1. Introduction

The silent chain transmission system is widely used in automobiles, motorcycles, and forklifts because of its low noise, high reliability, and high motion accuracy. As an advanced product of the silent chain, the Hy-Vo chain transmission system reduces the polygon effect because of the rocker pin. Based on the principle of bidirectional superposition, the dual-phase Hy-Vo chain transmission system can further reduce the polygon effect, vibration, and noise. In previous studies, researchers mainly focused on the design of the dual-phase Hy-Vo chain transmission system, with emphasis on the coupling effect between size parameters and the polygon effect (CHENG *et al.*, 2015; 2016a; 2016b; 2023). So far, the noise related research of the dual-phase Hy-Vo chain transmission system has not been involved.

A lot of studies have shown that noise can seriously harm people's mental and physical health (BASNER *et al.*, 2014; DRATVA *et al.*, 2012). Therefore, consumers are also paying more attention to the use experience of low noise. Recently, there have been more and more researches on the sound quality in various fields (SONG, YANG, 2022; RUAN *et al.*, 2022; PARK *et al.*, 2020). In common sound quality prediction methods, acoustic parameters such as *A*-weighted sound pressure level (*A*-SPL), loudness, sharpness, roughness, fluctuation, and articulation index (AI) are used as inputs (WANG *et al.*, 2022; CHEN *et al.*, 2022). WANG *et al.* (2022) proposed a nonlinear sound quality modeling method that uses an extreme gradient boosting algorithm to predict the overall sound quality inside a pure electric car. CHEN *et al.* (2022) used the backpropagation neural network and support vector regression (SVR) to predict the sound quality of tractors, and used a genetic algorithm to optimize the parameters of the prediction models. To predict the sound quality using the convolutional neural network (CNN), the researchers introduced various feature maps as inputs (HUANG *et al.*, 2021; JIN *et al.*, 2021). HUANG *et al.* (2021) converted the objective parameter evaluation

into feature graphs and proposed a prediction method with an adaptive learning rate tree based on CNN. Jin *et al.* (2021) demonstrated that MFCC can distinguish noise of different sound qualities and used MFCC feature maps as inputs to predict the transmission sound quality. In the above studies on sound quality prediction, neural networks are widely used because of their strong ability to adjust to nonlinearity. However, when the number of samples is insufficient, the accuracy of a prediction model will be poor.

To predict the sound quality in the case of small samples, we have the following studies in this paper: firstly, we collected the noise of the dual-phase Hy-Vo chain transmission system under different working conditions. Random 5 s clips are taken from each noisy audio for subsequent processing. Based on the equal interval direct one-dimensional evaluation method, all noise samples are subjectively evaluated by the testers. Secondly, we calculate the MFCC for each sample. The standard MFCC only reflects the static characteristics of the noise, and the dynamic characteristics can be described by the difference of these static characteristics. To further study the influence of MFCC order and frame number on the prediction effect, we construct MFCC feature maps of different sizes as inputs of the prediction model. Thirdly, we propose a data enhancement method called fuzzy generation based on the fuzzy phenomenon in the subjective evaluation. By constructing the membership function of each noise sample, the appropriate membership degree is selected for sample generation. After the dataset is expanded, we build a CNN model for the sound quality prediction, and the prediction results show that the full-frame standard MFCC feature map has the best prediction effect when the membership degree is 0.9. The more frames, the more complete the information contained in the MFCC, and the higher the prediction accuracy. However, higher order MFCC contains more redundant information, which will damage the prediction accuracy of the model. Finally, three common sound quality prediction methods are used in this paper, including the generalized regression neural network (GRNN), SVR, and ridge regression (RR). For each noise sample, we calculate six acoustic parameters (*A*-SPL, loudness, sharpness, roughness, fluctuation, and AI) as inputs. The comparative results show that the proposed new method has the lowest prediction error and strong resistance to overfitting. The flow chart of the sound quality research in this paper is shown in Fig. 1 and the structure of this paper is as follows: Sec. 2 involves the noise acquisition test and subjective evaluation of the noise sample. After the samples are preprocessed, we organize the testers to score the noise annoyance degree and test the correlation of the subjective evaluation results. In Sec. 3, the MFCC of all noise samples is calculated as an objective evaluation. After constructing the MFCC feature maps of different dimensions, the original dataset for the sound quality prediction is obtained by combining the subjective evaluation results. To train a more accurate prediction model, we use fuzzy generation to triple the size of the original dataset. In Sec. 4, we use MFCC feature maps of different specifications as input for the sound quality prediction and compare their prediction effects. After obtaining the optimal prediction model based on CNN, we compare it with the traditional sound quality prediction method. The results show that the prediction method proposed in this paper is more advantageous. Lastly, Sec. 5 presents the study's conclusion and summary.

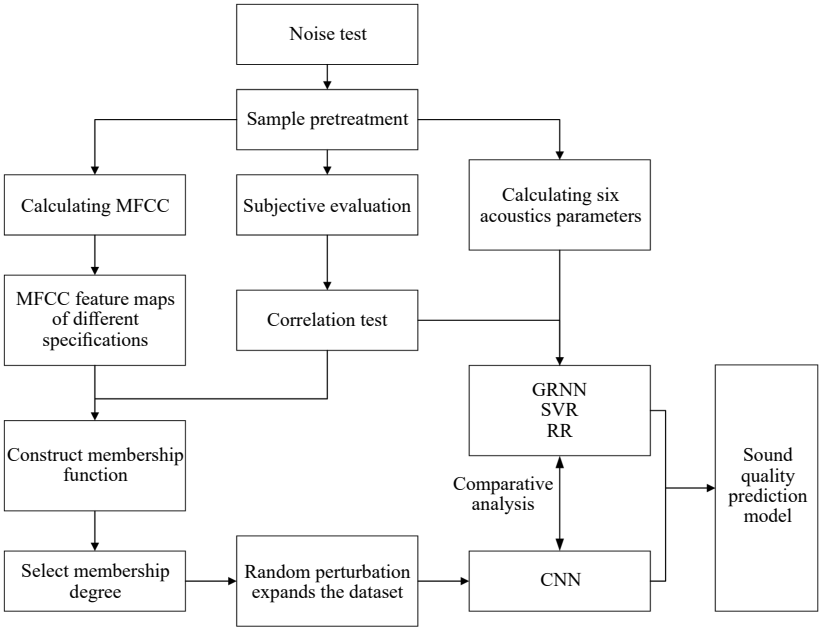


Fig. 1. Flow chart of sound quality prediction model construction.

2. Noise test and subjective evaluation

Different from single-phase transmission, the dual-phase sprocket teeth have phase difference. In our noise test, the drive sprocket tooth number is 35 with 5.14° phase difference, the driven sprocket tooth number is 37 with 4.86° phase difference, the pitch is 9.525 mm, the number of links is 84 and the chain form is 4×3.

As shown in Fig. 2, the noise test is conducted in an indoor reverberation environment. We use measurement microphone (MINIDSP UMIK-1) to collect the noise and the measurement microphone is positioned at the same height as the center of the drive sprocket. There are two measurement points we selected, the first one is at the distance of 0.5 m from the center of the drive sprocket, the second one is at the distance of 1 m from that. The minimum speed of the test is 500 rpm and the maximum speed is 4000 rpm. The test loads are 500 N, 600 N, and 750 N. Starting from 500 rpm, noise data is collected under three loads for each 500 rpm increase. There are two collection points (0.5 m and 1 m from the center of the drive sprocket), eight speeds (500 rpm, 1000 rpm, 1500 rpm, 2000 rpm, 2500 rpm, 3000 rpm, 3500 rpm, 4000 rpm), and three loads (500 N, 600 N, 750 N), so $2 \times 8 \times 3 = 48$ original noise samples can be obtained. The sampling frequency is 48 000 Hz, and the noise data is recorded using Adobe Audition 2022 software. All noise acquisition times are longer than 30 s, we randomly intercept 5 s segment for subsequent data processing. Under the same working conditions, the time-domain waveform of the single-phase and dual-phase transmissions are shown in Fig. 3. The orange line on the left represents the dual-phase transmission, and the blue line on the right represents the single-phase transmission.

As can be seen in Fig. 3, we can find that due to the principle of dual-phase superposition, the waveform of the dual-phase transmission is more uniform and denser at low speeds. When the speed is medium, the waveforms of the two transmissions are very similar. However, when running at high speed, the noise energy of the dual-phase transmission is obviously greater than that of the single-phase transmission. Therefore, the noise of the dual-phase transmission is different from that of other transmissions, and it is of great significance to study the sound quality prediction method of the dual-phase Hy-Vo chain transmission system.

After obtaining 48 noise samples, we organize twelve testers to conduct a subjective evaluation test. All of the testers are between 20 and 30 years old, and the ratio of men to women is 5:1. In addition, all testers have normal hearing and driving experience. As shown in Table 1, the subjective evaluation method is equal to the interval direct one-dimensional evaluation method (Guski, 1997). We rate the sound quality on a scale of discomfort, and there are five uncomfortable levels. A score of 0 is extremely uncomfortable level and a score of 10 is not uncomfortable level. Each of the remaining three uncomfortable levels has three scores, each score indicating the degree of discomfort in the same level. The subjective evaluation test is conducted in a quiet indoor environment, and the maximum SPL does not exceed 30 dB. The tester sits in a chair with headphones, and all the noise samples are played three times by Groove software. After listening, the tester gives the score and records it in a table.

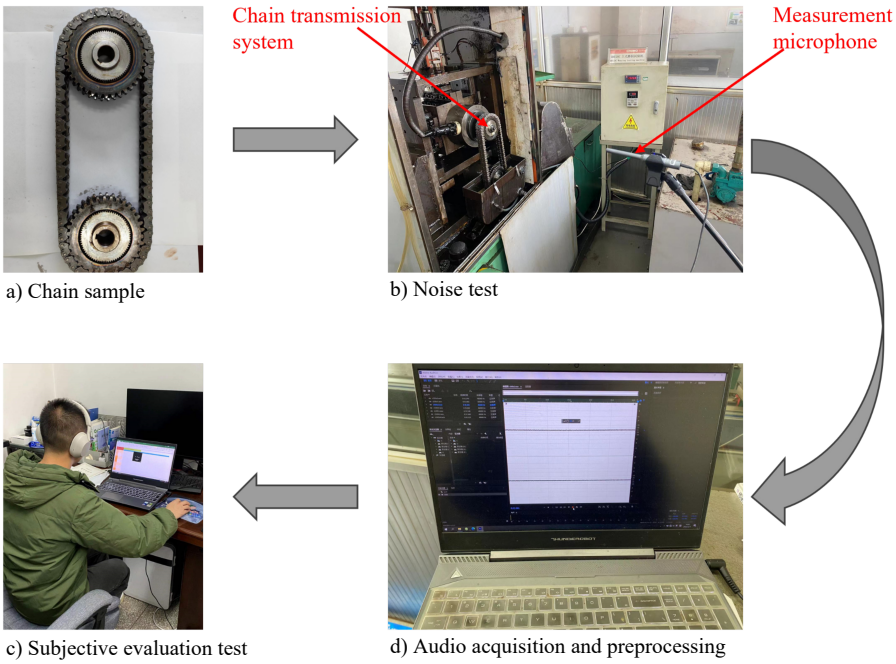


Fig. 2. Noise acquisition and data processing.

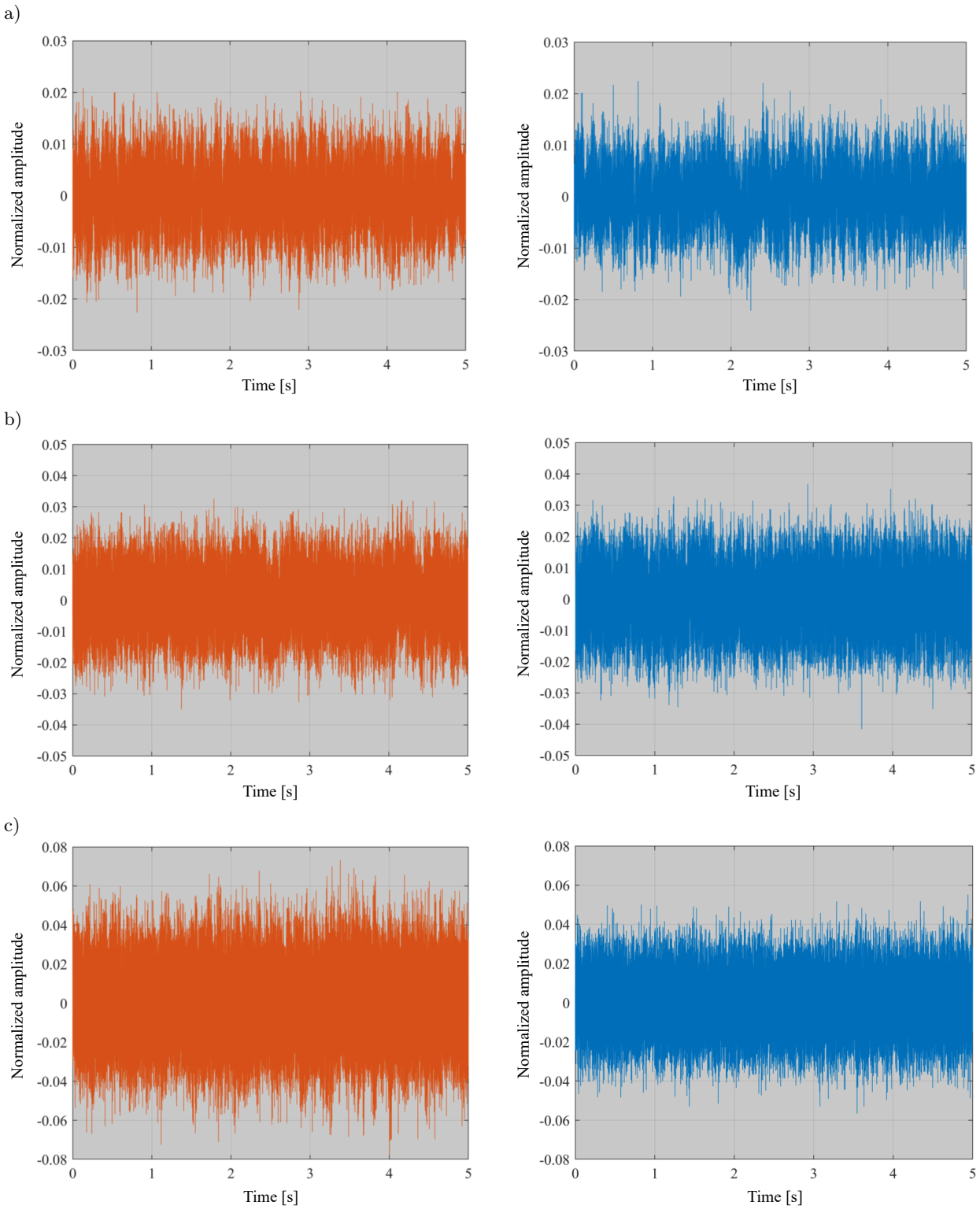


Fig. 3. Time-domain waveform comparison:
a) 1000 rpm – 0.5 m – 1000 N; b) 2500 rpm – 0.5 m – 1000 N; c) 4000 rpm – 0.5 m – 1000 N.

Table 1. Subjective evaluation scoring table.

Uncomfortable level	Extremely uncomfortable	Very uncomfortable	Moderately uncomfortable	Little uncomfortable	Not uncomfortable
Scores	0	1–3	4–6	7–9	10

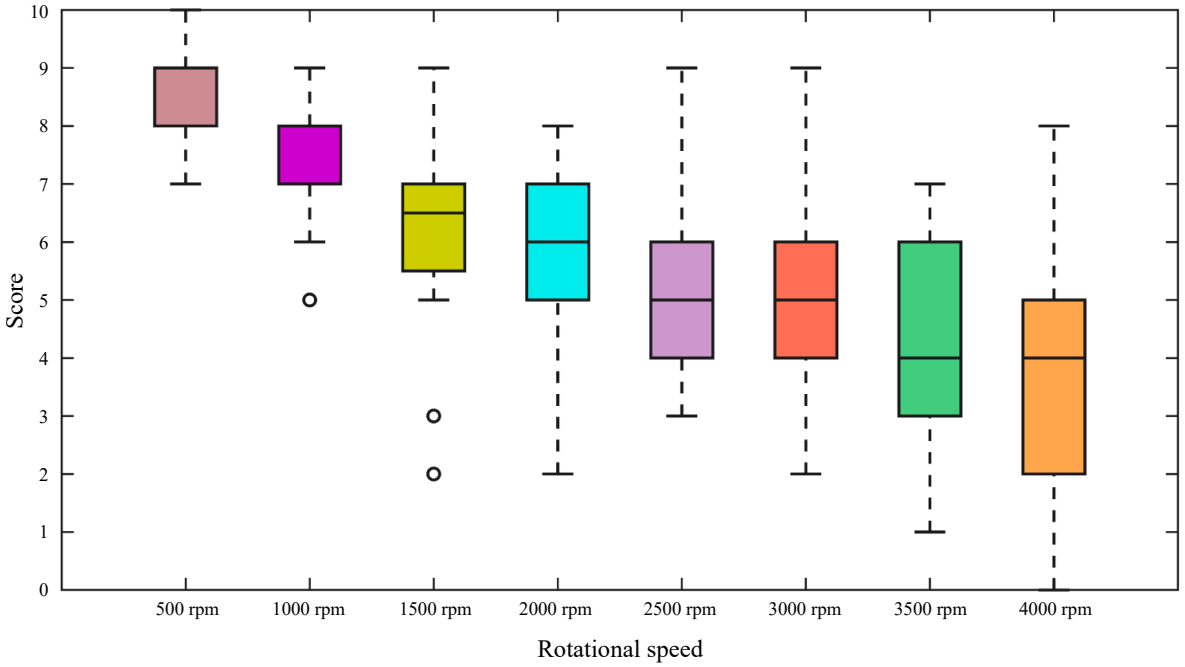


Fig. 4. Score boxplot for each speed.

All scores of each speed are presented in Fig. 4. Based on the horizontal line in the middle of the boxplot, in the range of 500 rpm–2500 rpm, it can be seen that the score decreases with the increase of the rotational speed. In this speed range, the sound quality of the chain transmission system becomes worse as the speed increases. The score of 3000 rpm remains unchanged compared to the score of 2500 rpm. However, the score continues to decline at 3500 rpm. As for the score of 4000 rpm, it is the same as the score of 3500 rpm. Therefore, in the case of medium and high speed, the sound quality of the chain transmission system shows a step-like decline trend. The length of box reflects the dispersion of scores. We can see that the scores are more dispersed at medium and high speeds, and there are even outliers at 1000 rpm and 1500 rpm. In the subjective evaluation test, we want all testers to have relatively consistent feelings about the same noise sample. The Spearman correlation analysis is performed on the scores of twelve testers and the results with poor correlation will be excluded. In the Spearman correlation analysis, the greater the coefficient R, the stronger the correlation. The equation of correlation coefficient (R) is:

$$R = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{1}$$

where x_i and y_i represent the corresponding elements of the two variables, $\overline{x}$ and $\overline{y}$ represent the average value of the corresponding variables.

Based on Eq. (1), the R between the twelve testers are calculated, as illustrated in Fig. 5. The numbers from P1 to P12 represent the twelve testers, and it can be seen that P3–P6, P3–P11, and P6–P11 have a maximum correlation of 0.96. The correlation between P5–P6 and P6–P9 are both less than 0.7, indicating a weak correlation. According to Fig. 5, we calculate the average correlation coefficient (ACC) for each tester, as shown in Table 2.

In Table 2, all testers have an ACC of more than 0.7, indicating that the scores of each tester is reasonable. Generally speaking, the average score of the twelve testers is used as the final score of each noise sample, as shown in Table 3.

Table 2. ACC of each tester.

Tester	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
ACC	0.850	0.842	0.846	0.866	0.800	0.813	0.817	0.893	0.796	0.835	0.854	0.882

Table 3. Average score of each noise sample.

Sample	1	2	3	4	5	6	7	...	42	43	44	45	46	47	48
Score	8.75	8.58	8.33	9.08	8.75	8.67	7.08	...	5.25	3.00	2.33	1.92	5.50	4.67	4.17

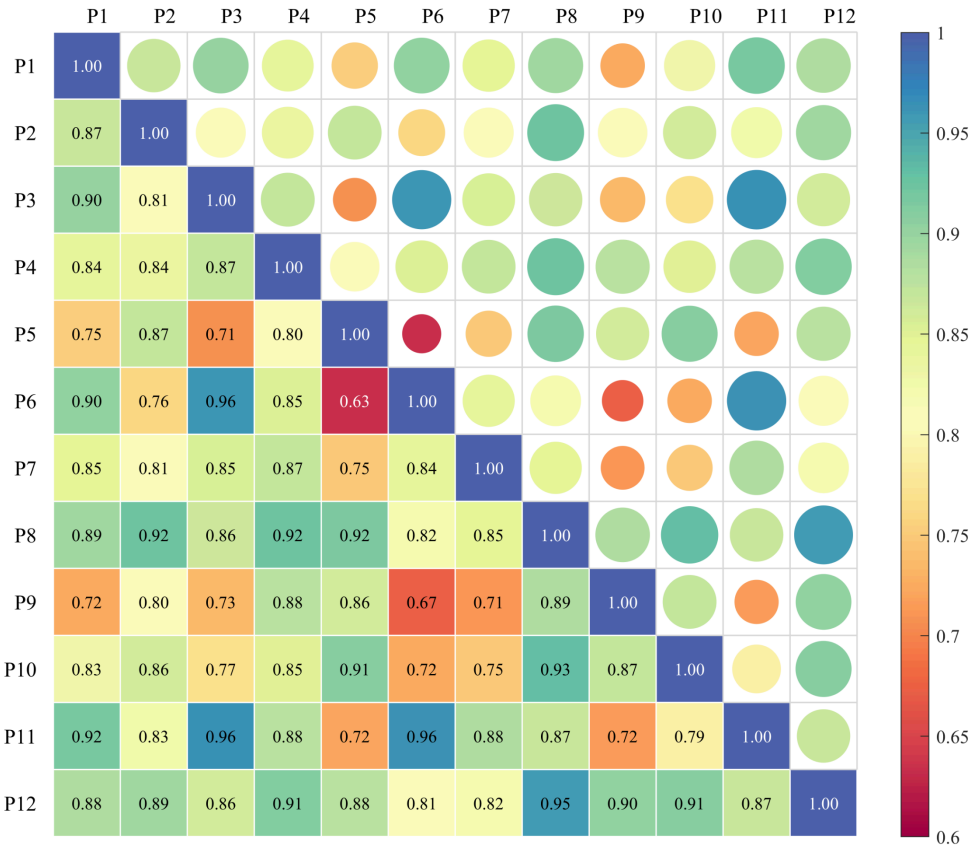


Fig. 5. Correlation heat map.

3. Objective evaluation and fuzzy generation

3.1. Construct Mel-frequency cepstral coefficient feature map

The MFCC is a feature extraction method commonly used in speech processing and audio analysis. It is based on the hearing characteristics of the human ear, by simulating the human ear's ability to perceive sounds of different frequencies, the sound signal is converted into a set of coefficients describing its characteristics. The advantage of MFCC is that they can effectively capture the main features of speech signals and have good adaptability for different speech processing tasks. However, they also have limitations, such as sensitivity to noise and possible degradation of performance in some complex environments. Therefore, in practical applications, MFCC is often used in combination with other types of feature and signal processing technologies (Abdul, Al-Talabani, 2022; Moondra, Chahal, 2023). The extraction process of MFCC is as follows:

1) Preprocessing: the sound signal is pre-weighted to increase the energy of the high frequency part:

$$y(t) = x(t) - \alpha x(t-1), \tag{2}$$

where $x(t)$ is the original signal, $y(t)$ is the pre-weighted signal, and α usually takes 0.95 or 0.97.

2) Framing: the segmentation of the sound signal into a series of short-time frames, each frame usually contains 20 ms–40 ms of data.

3) Windowing: the data of each frame is windowing processed, usually using hamming windows:

$$y(n) = x(n) \cdot \omega(n), \tag{3}$$

where $x(n)$ is the signal in a frame, $\omega(n)$ is the window function, and $y(n)$ is the signal after the window is added. The hamming window function is as follows:

$$\omega(n) = (1-a) - a \cdot \cos(2\pi n/N) \quad 1 < n < N, \tag{4}$$

where N is the number of sampling points, and different values of a will produce different hamming windows, in general, $a = 0.46$.

4) The Fourier transform: a fast Fourier transform (FFT) is performed on each frame of data to convert it into a signal in the frequency domain:

$$Y(k) = \sum_{0}^{N-1} y(n) \cdot e^{-j\frac{2\pi}{N}kn}, \tag{5}$$

where $Y(k)$ is the k-th component in the frequency domain, and N is the number of FFT points.

5) Mel filtering: the frequency domain signal is passed through a set of Mel filter banks to simulate the human ear's perception of different frequencies. Compared with the normal frequency mechanism, the Mel value is closer to the hearing mechanism of the human ear. It grows fast in the low frequency range, but it grows slowly in the high frequency range. Each frequency value corresponds to a Mel value, and the corresponding relationship is as follows:

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{6}$$

If we want to convert the Mel-frequency m to the frequency f, we can get it by sorting the above Eq. (6):

$$f = 700 \cdot (10^{m/2595} - 1). \tag{7}$$

The response $H_m(k)$ of each filter is usually defined as a triangular filter that is uniformly distributed on the Mel scale, and the output $S(m)$ is the signal energy that passes through the filter:

$$k = \frac{(1+N) \cdot f_m}{f_s}, \tag{8}$$

$$H_m(k) = \begin{cases} 0 & k < f(m-1), \\ \dfrac{2(k-f(m-1))}{a^*} & f(m-1) \le k \le f(m), \\ \dfrac{2(f(m+1)-k)}{a^*} & f(m) \le k \le f(m+1), \\ 0 & k \ge f(m+1), \end{cases} \tag{9}$$

$$S(m) = \sum_{k=0}^{K-1} |Y(k)|^2 \cdot H_m(k), \tag{10}$$

where

$$a^* = (f(m+1) - f(m-1))(f(m) - f(m-1)).$$

6) Log the output of the Mel filter bank to obtain the logarithmic energy spectrum:

$$L(m) = \log(S(m)), \tag{11}$$

where $L(m)$ is the logarithmic energy spectrum.

7) Discrete cosine transforms: perform a discrete cosine transform (DCT) on the logarithmic energy spectrum to obtain the MFCC coefficient:

$$C(n) = \sum_{m=0}^{M-1} L(m) \cdot \cos\left[\frac{\pi}{M}(m+0.5)n\right], \quad n = 1, 2, ..., L, \tag{12}$$

where $C(n)$ is the n-th cepstral coefficient, M is the number of Mel filters, and L refers to the MFCC coefficient order, usually 12–16.

From Eq. (3) to Eq. (12), we can get the standard MFCC, which only reflects the static properties of audio. The dynamic characteristics of audio can be described by the difference of these static characteristics, as follows:

$$\Delta C_t = \frac{\sum_{n=1}^{N} n(C_{t+n} - C_{t-n})}{2\sum_{n=1}^{N} n^2}, \tag{13}$$

$$\Delta\Delta C_t = \frac{\sum_{n=1}^{N} n(\Delta C_{t+n} - \Delta C_{t-n})}{2\sum_{n=1}^{N} n^2}. \tag{14}$$

Equations (13) and (14) represent the first- and second-order difference, respectfully. In this paper, we take each frame as 32 ms, the noise sample is divided into K frames and the MFCC of L order is calculated. As shown in Fig. 6, we can obtain $K \times L$ feature maps of different orders. The standard full-frame MFCC feature map is 311×13, 311×26 with only first-order differences, and 311×39 also with

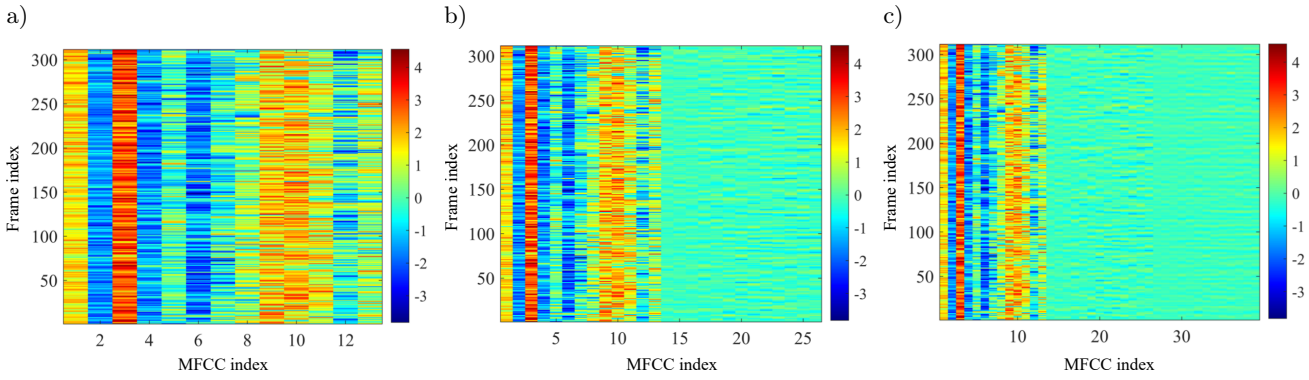


Fig. 6. MFCC feature map: a) 311×13; b) 311×26; c) 311×39.

second-order differences. We also get the feature maps of two frame numbers, 208×13 and 104×13, respectively. Finally, we can get 5 input features of different sizes.

3.2. Fuzzy generation

Fuzzy mathematics is a mathematical method to deal with uncertain information. Compared with traditional binary logic and precise mathematics, it pays more attention to the description and processing of fuzzy and uncertain phenomena in the real world. The core concept of fuzzy mathematics is a fuzzy set. Unlike traditional sets, where the elements either belong to or do not belong to the set, the degree to which an element in a fuzzy set belongs to the set is a numerical value between 0 and 1, called membership. This makes fuzzy sets more flexible in describing uncertainty and ambiguity in the real world. Membership functions are used to describe the degree to which an element belongs to a fuzzy set. The value of this function is between 0 and 1. The core strength of fuzzy mathematics is that it provides an effective way to deal with the uncertainty and ambiguity that are prevalent in the real world. By introducing fuzzy concepts, it allows for the more flexible and realistic problem solving and decision-making process (Ruan, Li, 2021; Gündoğdu, Kahraman, 2019; Bustince *et al.*, 2016).

In the previous subjective evaluation, there is a fuzzy problem. Generally speaking, for the same noise sample, researchers only calculate the average score as the final subjective evaluation score. In fact, the scores of all testers are reasonable after the correlation test. Therefore, we believe that in the range of minimum and maximum scores, the average score as a label value is when the membership degree is 1, and the fuzzy mapping is constructed as follows:

$$\begin{aligned} F : V &\to [0,\ 1], \\ m &\mapsto F(m), \end{aligned} \tag{15}$$

where V is value field [0 10], F is the fuzzy interval of V, and $F(m)$ is the membership function.

For each noise sample, we can construct its fuzzy interval and membership function. In Table 4, the average score is the core of the fuzzy interval, the minimum score is the left boundary (LB), and the maximum score is the right boundary (RB). We construct the membership function on the fuzzy interval and select the appropriate membership degree to delimit the sample generation interval. Then the label value is randomly perturbed over the sample generation interval to expand the dataset. The membership function is defined as follows:

$$\frac{F(m_d) - 0}{d - r} = \frac{1 - 0}{k - r} \Rightarrow F(m_d) = \frac{1}{r - k}(r - d), \tag{16}$$

Table 4. Fuzzy intervals.

Sample	LB	Core	RB	Sample	LB	Core	RB
1	8	8.75	10	25	4	5.17	6
2	8	8.58	9	26	3	4.33	6
3	7	8.33	9	27	3	4.17	6
4	8	9.08	10	28	5	6.83	9
5	8	8.75	10	29	4	5.83	8
6	7	8.67	10	30	4	5.75	8
7	6	7.08	8	31	3	4.33	6
8	5	6.92	8	32	2	3.75	5
9	6	7.00	8	33	2	3.83	6
10	7	8.33	9	34	4	6.33	9
11	6	8.00	9	35	3	5.83	8
12	6	7.58	9	36	3	5.50	8
13	3	5.58	8	37	2	3.67	5
14	3	5.25	7	38	1	3.25	5
15	2	5.25	7	39	1	3.25	6
16	7	7.75	9	40	3	5.50	7
17	6	7.17	8	41	3	5.25	7
18	6	6.83	8	42	3	5.25	7
19	4	5.50	7	43	1	3.00	5
20	3	5.08	7	44	0	2.33	5
21	2	4.83	7	45	0	1.92	5
22	6	7.25	8	46	3	5.50	8
23	6	6.83	8	47	2	4.67	8
24	5	6.50	7	48	1	4.17	7

$$\frac{F(m_d) - 0}{d - l} = \frac{1 - 0}{k - l} \Rightarrow F(m_d) = \frac{1}{k - l}(d - l), \tag{17}$$

$$F(m_d) = \begin{cases} 0 & (0 \le d < l), \\ \dfrac{1}{k - l}(d - l) & (l \le d < k), \\ \dfrac{1}{r - k}(r - d) & (k \le d \le r), \\ 0 & (r < d \le 10), \end{cases} \tag{18}$$

where k is the core point, l is the LB point, r is the RB point, d is a random generation point, and $F(m_d)$ is the membership of d.

As can be seen in Fig. 7, the farther away from the core point, the smaller the membership degree. For different samples, the span of their membership function is usually different. Under the same membership degree, the larger the span, the larger the sample generation interval. In the generation interval, the sample label values are randomly perturbed to expand the dataset. However, the larger the interval, the more noise the new sample points contain. In this paper, we choose four membership degrees of 0.3, 0.5, 0.7, and 0.9 for fuzzy generation. The dataset is expanded to three times its original size, including 144 samples.

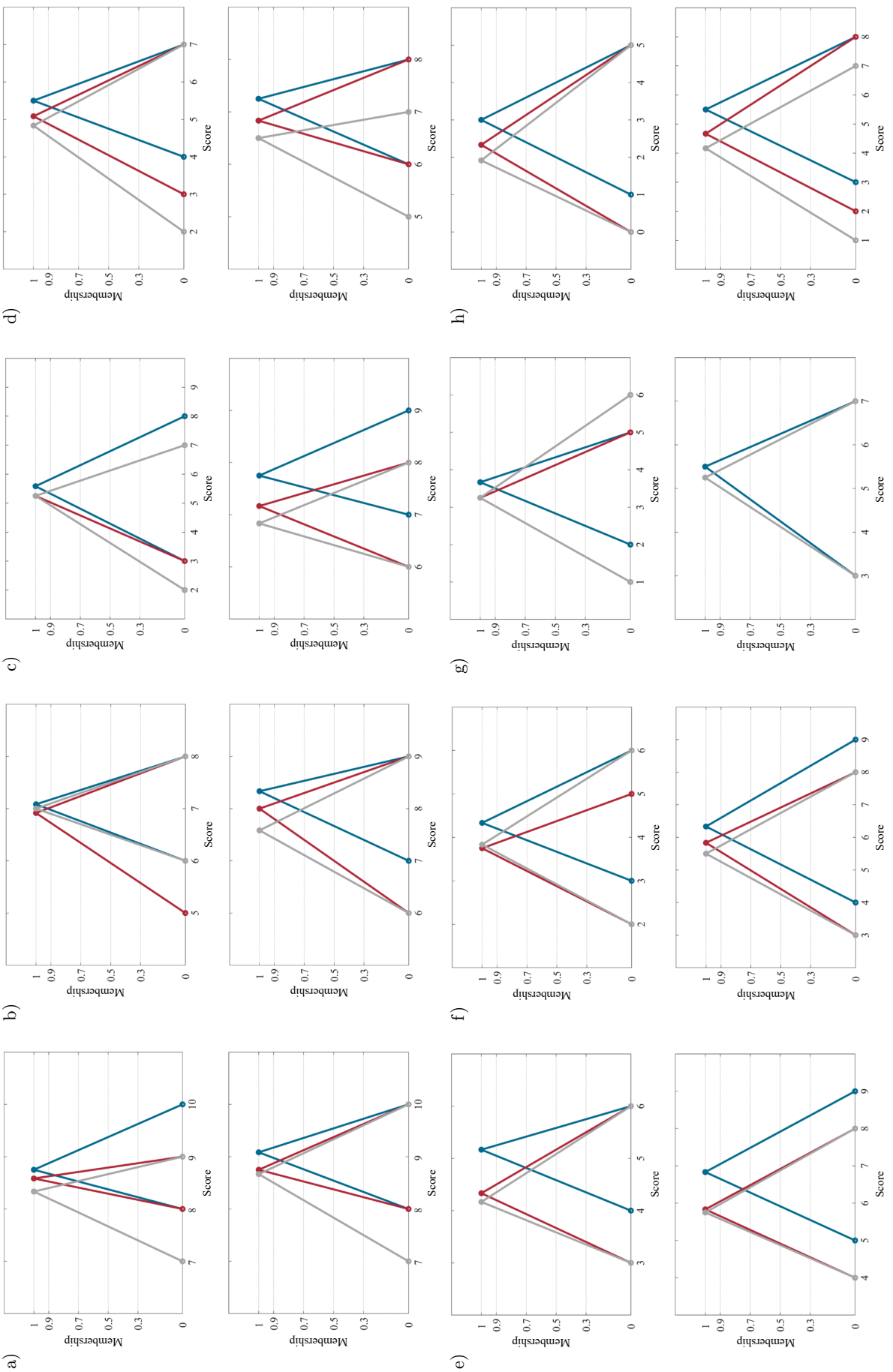
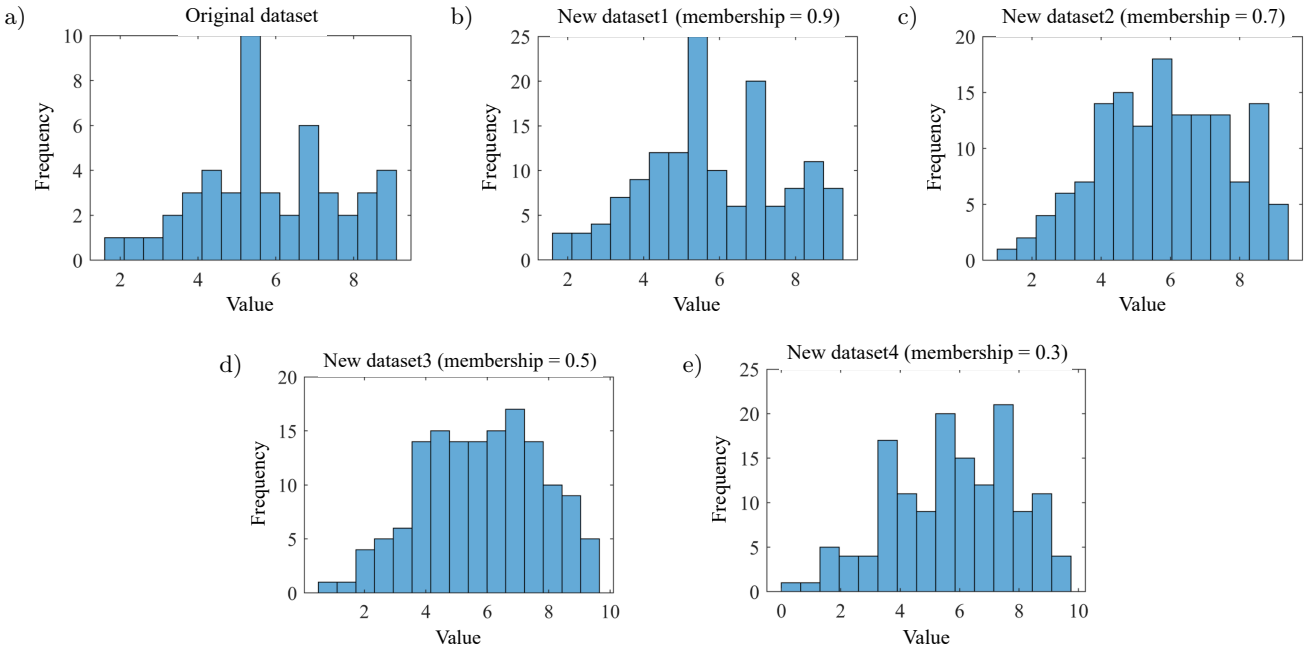


Fig. 7. All membership functions: a) 500 rpm; b) 1000 rpm; c) 1500 rpm; d) 2000 rpm; e) 2500 rpm; f) 3000 rpm; g) 3500 rpm; h) 4000 rpm. The upper graph in each subgraph represents the membership function for the samples of 0.5 m measurement point, and the lower graph in each subgraph represents the membership function for the samples of 1 m measurement point.



Type of dataset	Mean value	Standard deviation
Original dataset	5.84	1.82
New dataset1 (membership = 0.9)	5.84	1.81
New dataset2 (membership = 0.7)	5.83	1.85
New dataset3 (membership = 0.5)	5.82	1.94
New dataset4 (membership = 0.3)	5.81	2.08

Fig. 8. Histogram comparison of original dataset and new datasets.

Figure 8 shows the histogram comparison between the original dataset and the expanded new dataset. The histogram of the original dataset shows a relatively symmetric unimodal distribution with a mean of 5.84 and a standard deviation of 1.82. The new dataset1 is very similar in shape to the original dataset, with the mean remaining at 5.84 and the standard deviation slightly reduced to 1.81. The new dataset2 has a slightly changed distribution shape, with a mean of 5.83 and a standard deviation of 1.85, slightly increasing the variability. The distribution shape of the new dataset3 has a significant change, with the mean of 5.82 and the standard deviation increasing to 1.94, indicating a further increase in variability. The new dataset4 has the most significant change in distribution shape, with a mean of 5.81 and a standard deviation of 2.08, indicating the greatest variability. As the membership value decreases, the standard deviation of the new dataset gradually increases, indicating that the perturbation introduces more variability. The mean remains essentially unchanged, indicating that the new dataset is still centered around the mean of the original data. By analyzing Fig. 8, we can conclude that higher membership values (such as 0.9 and 0.7) retain the main features of the original dataset and increase the number of datasets while maintaining low variability. Lower membership values (such as 0.5 and 0.3) introduce more variability and outliers, and may introduce more noise despite increasing the diversity of the dataset.

4. Modeling and prediction

4.1. Convolutional neural network

The CNN is a kind of deep learning model that has achieved great success in image recognition, video analysis, natural language processing and other fields. CNN is particularly suited for working with data with a grid structure, such as images and time series data. The core idea of CNN is to use convolutional layers to automatically learn features of spatial hierarchy from data. These features are gradually abstracted and combined through multiple convolution layers and subsampling layers (usually pooling layers) to accomplish complex tasks. The advantage of CNN is its ability to automatically learn and extract features without the need for manual feature engineering (Bhatt *et al.*, 2021; Goumiri *et al.*, 2023; Mandouh *et al.*, 2023). CNN is generally used to solve the classification problem. To predict the sound quality, we set the output layer to have only one node, and do not use nonlinear activation function, so that the output is a linear

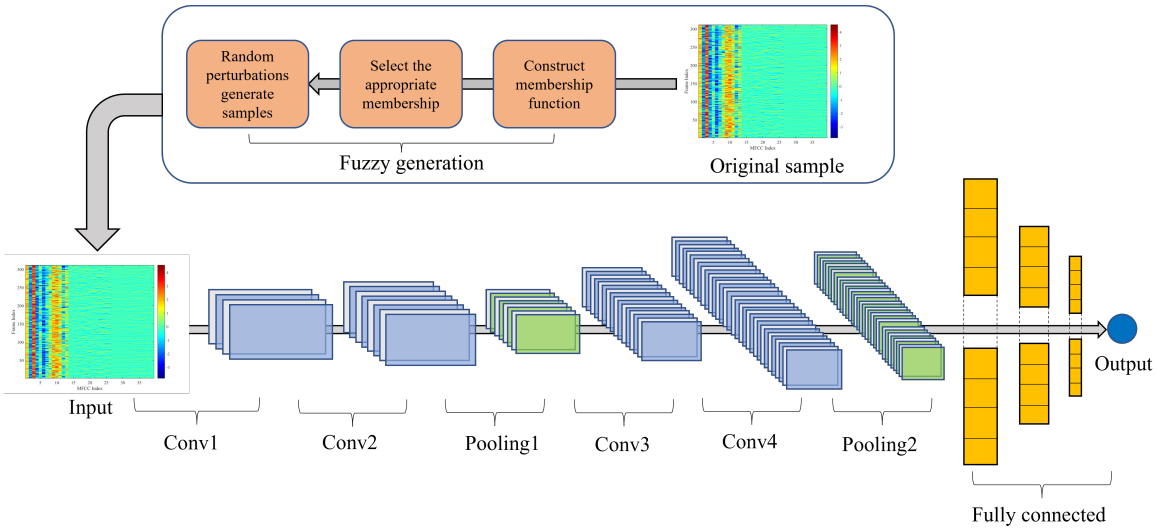


Fig. 9. Model structure.

transformation of the inputs, which can get a continuous value.

As shown in Fig. 9, the model structure consists of four convolution layers, two pooling layers, a flatten layer and three fully connected layers. In addition to the output layer, the activation function of the other layers is a rectified linear unit (ReLU). The pooling layer adopts the maximum pooling, the step size is 2 with 0 padding. The convolution layer has a step size of 1 without 0 padding. The numbers of neurons in the three fully connected layers are 1024, 128, and 1, respectively. Using dropout technology in the first fully connected layer, proceeds with the dropout rate set to 0.5. The last layer is the output layer, which outputs the evaluation score. Taking the input feature map $\mathrm{MFCC}_{311\times13}$ as an example, the model structure parameters are shown in Table 5.

Table 5. Structural parameters.

Layer type	Channels/Units
Input 311×13	3
3×3 Conv1 ReLU, stride 1	6
3×3 Conv2 ReLU, stride 1	12
2×2 Maxpooling1 ReLU, stride 2	12
3×3 Conv3 ReLU, stride 1	24
3×3 Conv4 ReLU, stride 1	48
2×2 Maxpooling2 ReLU, stride 2	48
Flatten	3600
Fully connected (1)	1024
Dropout	1024
Fully connected (2)	128
Fully connected (3)	1

The MFCC feature map is taken as input, the evaluation score is taken as output, and the ratio of training set to test set is 5:1. Using the Adam optimizer, the initial learning rate is 0.001, a root mean squared error (RMSE) is the loss function, and the epoch is set to 200. With 5 input feature maps and 4 membership degrees, the average of 5 training results is taken, and the model is trained $5\times4\times5=100$ times in total. In model training, we choose the R, the RMSE, and the mean absolute error (MAE) as evaluation indexes, and the calculation formula is as follows:

$$R=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}, \tag{19}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-y_i)^2}, \tag{20}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\,x_i-y_i\,|, \tag{21}$$

where n is the number of samples, x_i is the predicted value of the sample, and y_i is the true value of the sample.

The prediction effects of MFCC feature maps with different orders are shown in Table 6, and Δ represents the increment compared to the results in the first row. Based on the training results of standard full-frame $\mathrm{MFCC}_{311\times13}$ feature map, we can see that with the increase of the MFCC order, the three evaluation indexes are deteriorating. The high order MFCC contains too much useless information and damages the performance of the model. In addition, the influence of the MFCC frame number on the training results is shown in Table 7.

In Table 7, all three evaluation indexes get worse as the number of frames decreases. Compared with $\mathrm{MFCC}_{311\times13}$, the frame number of $\mathrm{MFCC}_{208\times13}$ decreases by 33.3 %, but RMSE and MAE increase by

Table 6. Prediction effects of different MFCC orders.

Feature maps	Training			Prediction		
	R (Δ)	RMSE (Δ)	MAE (Δ)	R (Δ)	RMSE (Δ)	MAE (Δ)
$MFCC_{311\times13}$	0.979	0.394	0.314	0.971	0.474	0.371
$MFCC_{311\times26}$	0.966 (−0.013)	0.505 (+0.111)	0.392 (+0.078)	0.953 (−0.018)	0.603 (+0.129)	0.467 (+0.096)
$MFCC_{311\times39}$	0.923 (−0.056)	0.670 (+0.276)	0.539 (+0.225)	0.902 (−0.069)	0.785 (+0.311)	0.617 (+0.246)

Table 7. Prediction effects of different MFCC frame number.

Feature maps	Training			Prediction		
	R (Δ)	RMSE (Δ)	MAE (Δ)	R (Δ)	RMSE (Δ)	MAE (Δ)
$MFCC_{311\times13}$	0.979	0.394	0.314	0.971	0.474	0.371
$MFCC_{208\times13}$	0.977 (−0.002)	0.567 (+0.173)	0.481 (+0.167)	0.970 (−0.001)	0.663 (+0.189)	0.558 (+0.187)
$MFCC_{104\times13}$	0.965 (−0.014)	1.248 (+0.854)	1.143 (+0.829)	0.958 (−0.013)	1.326 (+0.852)	1.211 (+0.840)

Table 8. Prediction effects of different membership degrees.

Membership degree	Training			Prediction		
	R (Δ)	RMSE (Δ)	MAE (Δ)	R (Δ)	RMSE (Δ)	MAE (Δ)
0.9	0.993	0.256	0.211	0.991	0.279	0.224
0.7	0.990 (−0.003)	0.319 (+0.063)	0.260 (+0.049)	0.985 (−0.006)	0.379 (+0.100)	0.305 (+0.081)
0.5	0.971 (−0.022)	0.447 (+0.191)	0.356 (+0.145)	0.966 (−0.025)	0.490 (+0.211)	0.400 (+0.176)
0.3	0.964 (−0.029)	0.552 (+0.296)	0.427 (+0.216)	0.941 (−0.050)	0.748 (+0.469)	0.554 (+0.330)

39.9 % and 50.4 %, respectively. For $MFCC_{104\times13}$, the frame number continues to decline by 33.3 %, while RMSE and MAE increase sharply by 179.7 % and 226.4 %. Therefore, the prediction error is more sensitive to the frame number. Too few frames will lead to missing key information, and the accuracy of the model will be seriously degraded.

When the standard full-frame $MFCC_{311\times13}$ feature map is used as input, different membership degrees also affect the prediction results. As can be seen from Table 8, the prediction is best when the membership degree is 0.9. The membership degree gradually decreases, and the three evaluation indexes gradually deteriorate.

Based on the above comparative experiments, we can know that the model prediction is best when the frame number is 311, the MFCC order is 13 and the membership degree is 0.9. Figure 10 shows the convergence curve under optimal conditions. In the first 21 iterations, the loss of the model decreases rapidly, but there are some fluctuations. In subsequent iterations, the model slowly converges. Finally, the RMSE of the training set is 0.142 and the RMSE of the test set is 0.153. The error is small enough to meet the scoring requirements of subjective evaluation, and the final prediction results are shown in Table 9.

Table 9. Final CNN prediction results.

Indexes	Training			Prediction		
	R	RMSE	MAE	R	RMSE	MAE
Results	0.997	0.142	0.110	0.996	0.153	0.127

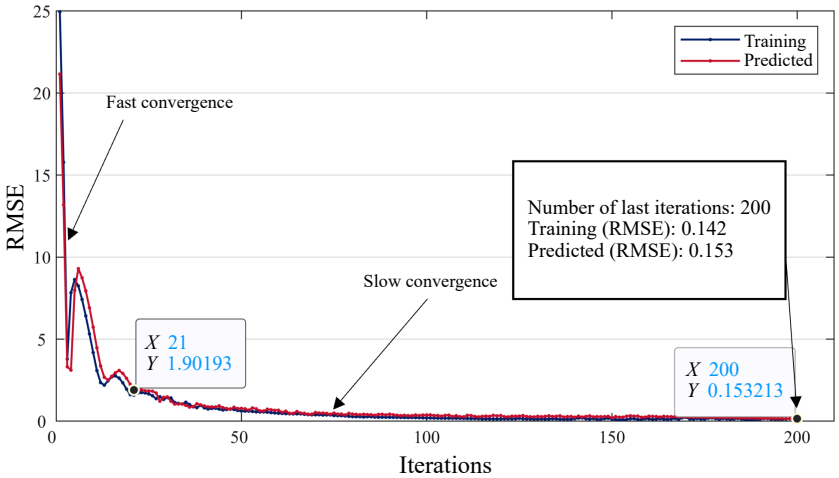


Fig. 10. Convergence curve.

To further verify the generalization ability of the proposed method in small samples, we used a five-fold cross-validation in the experiment. Five-fold cross-validation divides the dataset into five subsets, using one of the subsets as the validation set and the remaining four subsets as the training set, repeating five times to ensure that each subset is used as a single validation set. The final model performance is averaged by the results of five experiments. As shown in Table 10, training the model with the new dataset can significantly improve the model's predictive performance and decreases the MAE value. This shows that the fuzzy generation method is effective under the condition of small samples and can enhance the generalization ability of the model. When the membership value is large (such as 0.9 and 0.7), the MAE value of the model is significantly reduced. At the same time, a low standard deviation is maintained, indicating that this degree of disturbance can effectively increase the data diversity without introducing too much noise. When membership values are small (such as 0.5 and 0.3), more noise is introduced into the dataset. Although the model performance is also improved, the effect is not as good as when the membership value is larger.

Table 10. Five-fold cross-validation results.

Type of the dataset	MAE	Standard deviation
Original dataset	3.241	1.034
New dataset1 (membership = 0.9)	0.736	0.121
New dataset2 (membership = 0.7)	0.885	0.178
New dataset3 (membership = 0.5)	1.078	0.141
New dataset4 (membership = 0.3)	1.448	0.149

4.2. Comparative analysis

To compare with traditional sound quality prediction methods, generalized regression neural network, SVR and RR models are used in this paper. We first use the Audio toolbox in MATLAB to calculate six acoustic parameters (*A*-SPL, loudness, sharpness, roughness, fluctuation, and AI) for all noise samples, as shown in Fig. 11. We take the six acoustic parameters as inputs, the evaluation scores as outputs, and the ratio of training set to test the set is also 5:1.

A generalized regression neural network (GRNN) is a type of neural network based on a radial basis function, mainly used to solve regression problems. The structure of GRNN is relatively simple, including input layer, pattern layer, summation layer and output layer. GRNN has applications in many fields, especially for scenarios that require fast and accurate regression predictions (Zhu *et al.*, 2022; Yao *et al.*, 2023). In the GRNN model, only one spread parameter σ needs to be optimized. By using the particle swarm optimization algorithm, the number of particles is 30, the maximum number of iterations is 20, and the optimal parameter $\sigma = 0.12$ is found on the interval [0.01 0.8].

The SVR is a regression method based on the principles of support vector machines. The core idea of SVR is to find a function that fits the training data as best as possible within a limited error range while maintaining the generalization ability of the model. For nonlinear data, the SVR uses kernel functions to map the data into a high-dimensional space, where linear regression is performed. Common kernel functions include linear kernel, polynomial kernel, radial basis function kernel, and so on (Zhan *et al.*, 2022; Shi *et al.*, 2021). For the SVR model with radial basis function, we also use the particle swarm optimization algorithm to find the two optimal parameters (penalty parameter c and kernel parameter g). The number of particles is 30, the maximum number of iterations is 20, and the best $c = 34.83$, $g = 0.32$ are found on the interval [0.01 100].

The RR, also known as the Tikhonov regularization, is a linear regression method for dealing with multicollinearity problems. Multicollinearity refers to the fact that there is a high degree of correlation between predictor variables in a regression analysis. The RR solves this problem by introducing a regularization term, thereby improving the stability and predictive power of the model. The basic idea of RR is to add a regularization term to the loss function of ordinary least squares regression. Choosing proper regularization parameter λ is the key to applying RR (Yasin *et al.*, 2022; Dar *et al.*, 2023). In this paper, for the RR model, the 5-fold cross validation is used to find the optimal λ. The value range is $[10^{-6}, 10^{-5.76}, ..., 10^{6}]$, and the best $\lambda = 1.33$ is found when the mean square error is minimum.

Table 11 shows the three evaluation indexes of three traditional sound quality prediction methods. Compared with the CNN model, we can see that the GRNN model performs slightly better in training than

Table 11. Comparison of prediction effect on different models.

Model	Training			Prediction		
	R (Δ)	RMSE (Δ)	MAE (Δ)	R (Δ)	RMSE (Δ)	MAE (Δ)
CNN	0.997	0.142	0.110	0.996	0.153	0.127
GRNN	0.998 (+0.001)	0.105 (−0.037)	0.055 (−0.055)	0.988 (−0.008)	0.239 (+0.086)	0.210 (+0.083)
SVR	0.991 (−0.006)	0.243 (+0.101)	0.288 (+0.178)	0.964 (−0.032)	0.407 (+0.254)	0.288 (+0.161)
RR	0.977 (−0.020)	0.360 (+0.218)	0.330 (+0.220)	0.966 (−0.030)	0.389 (+0.236)	0.336 (+0.209)

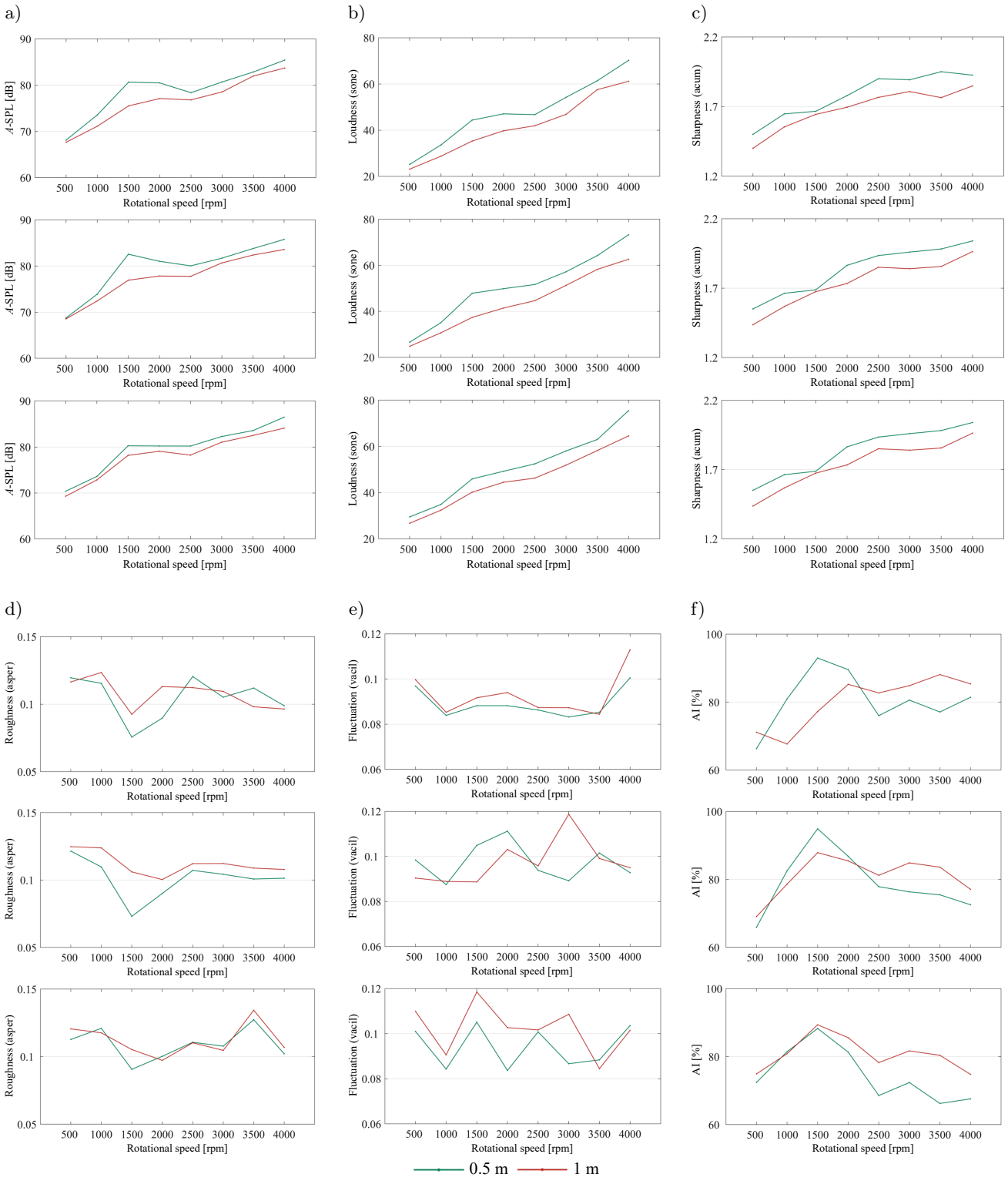


Fig. 11. Acoustic parameters: a) *A*-SPL; b) loudness; c) sharpness; d) roughness; e) fluctuation; f) AI. The top picture of each subgraph represents the parameters at 500 N, the middle picture of each subgraph represents the parameters at 600 N, and the bottom picture of each subgraph represents the parameters at 750 N.

the CNN. However, in the prediction, the R of the CNN model is the largest (0.996), while RMSE and MAE are the smallest (0.153 and 0.127, respectively). The effect of the CNN model on the test set is the least different from that on the training set, and the overfitting degree of the three traditional methods is higher. Therefore, the new sound quality prediction method proposed in this paper is superior to other three methods.

5. Conclusion

The noise of the dual-phase Hy-Vo chain transmission system is different from that of the single-phase transmission. First of all, we have carried out the noise acquisition test of the dual-phase Hy-Vo chain transmission system. Then all the noise samples are subjectively evaluated, and the results are tested for correlation. The ACC of all testers is greater than 0.7, indicating that the subjective evaluation results are reasonable.

The MFCC feature maps of all noise samples are calculated as objective evaluation. By selecting different membership degrees for fuzzy generation, the original dataset is expanded by three times. The CNN model is constructed to predict the sound quality. The comparison results show that when the membership degree is 0.9, the prediction effect of standard full-frame MFCC feature map is the best.

Compared with the traditional sound quality prediction methods (GRNN, SVR, and RR), the CNN model has the best performance on the test set. The correlation coefficient is 0.996, the root mean square error is 0.153, and the MAE is 0.127. In addition, for the CNN model, the difference between the training effect and the prediction effect is small. Therefore, the new method proposed in this paper not only has the highest accuracy, but also has a strong ability to resist overfitting.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (no. 51775222) and the Science and Technology Development Project of Jilin Province in China (no. 20200401136GX).

References

1. Abdul Z.K., Al-Talabani A.K. (2022), Mel frequency cepstral coefficient and its applications: A review, *IEEE Access*, **10**: 122136–122158, doi: 10.1109/access.2022.3223444.
2. Basner M. *et al.* (2014), Auditory and non-auditory effects of noise on health, *Lancet*, **383**(9925): 1325–1332, doi: 10.1016/s0140-6736(13)61613-x.
3. Bhatt D. *et al.* (2021), CNN variants for computer vision: History, architecture, application, challenges and future scope, *Electronics*, **10**(20): 2470, doi: 10.3390/electronics10202470.
4. Bustince H. *et al.* (2016), A historical account of types of fuzzy sets and their relationships, *IEEE Transactions on Fuzzy Systems*, **24**(1): 179–194, doi: 10.1109/tfuzz.2015.2451692.
5. Chen P., Xu L., Tang Q., Shang L., Liu W. (2022), Research on prediction model of tractor sound quality based on genetic algorithm, *Applied Acoustics*, **185**: 108411, doi: 10.1016/j.apacoust.2021.108411.
6. Cheng Y., Chen L., Ge P., Chen X., Niu J. (2023), Design and optimization of dual-phase chain transmission system based on single tooth chain plate, *Mechanics Based Design of Structures and Machines*, **51**(2): 899–913, doi: 10.1080/15397734.2020.1856683.
7. Cheng Y., Wang Y., Li L., Yin S., An L., Wang X. (2015), Design method of dual phase Hy-Vo silent chain transmission system, *Strojniški vestnik – Journal of Mechanical Engineering*, **61**(4): 237–244, doi: 10.5545/sv-jme.2014.2318.
8. Cheng Y., Wang Y., Li L., Yin S., An L., Wang X. (2016a), Multi-variation characteristic of dual phase Hy-Vo silent chain transmission system, *Mechanism and Machine Theory*, **103**: 40–50, doi: 10.1016/j.mechmachtheory.2016.04.011.
9. Cheng Y., Yin S., Wang X., An L., Liu H. (2016b), Design and analysis of double-side meshing and dual-phase driving timing silent chain system, *Strojniški vestnik – Journal of Mechanical Engineering*, **62**(2): 127–136, doi: 10.5545/sv-jme.2015.2837.
10. Dar I.S., Chand S., Shabbir M., Kibria B.M.G. (2023), Condition-index based new ridge regression estimator for linear regression model with multicollinearity, *Kuwait Journal of Science*, **50**(2): 91–96, doi: 10.1016/j.kjs.2023.02.013.
11. Dratva J. *et al.* (2012), Transportation noise and blood pressure in a population-based sample of adults, *Environmental Health Perspectives*, **120**(1): 50–55, doi: 10.1289/ehp.1103448.
12. Goumiri S., Benboudjema D., Pieczynski W. (2023), A new hybrid model of convolutional neural networks and hidden Markov chains for image classification, *Neural Computing and Applications*, **35**(24): 17987–18002, doi: 10.1007/s00521-023-08644-4.
13. Guski R. (1997), Psychological methods for evaluating sound quality and assessing acoustic information, *Acta Acustica united with Acustica*, **83**(5): 765–774.
14. Gündoğdu F.K., Kahraman C. (2019), Spherical fuzzy sets and spherical fuzzy TOPSIS method, *Journal of Intelligent & Fuzzy Systems*, **36**(1): 337–352, doi: 10.3233/jifs-181401.
15. Huang H., Wu J.H., Lim T.C., Yang M., Ding W. (2021), Pure electric vehicle nonstationary interior sound quality prediction based on deep CNNs with an adaptable learning rate tree, *Mechanical Systems and Signal Processing*, **148**: 107170, doi: 10.1016/j.ymssp.2020.107170.
16. Jin S., Wang X., Du L., He D. (2021), Evaluation and modeling of automotive transmission whine noise quality based on MFCC and CNN, *Applied Acoustics*, **172**: 107562, doi: 10.1016/j.apacoust.2020.107562.
17. Mandouh A.A., Ali M.E.N.O., Mohamed M., Taha L.G.E., Mohamed S.A. (2023), A performance analysis of point CNN and mask R-CNN for building extraction from multispectral LiDAR data, *International Journal of Advanced Computer Science and Applications*, **14**(9), doi: 10.14569/ijacsa.2023.0140948.

18. Moondra A., Chahal P. (2023), Improved speaker recognition for degraded human voice using modified-MFCC and LPC with CNN, *International Journal of Advanced Computer Science and Applications*, **14**(4), doi: 10.14569/ijacsa.2023.0140416.

19. Park J.H., Park H., Kang Y.J. (2020), A study on sound quality of vehicle engine sportiness using factor analysis, *Journal of Mechanical Science and Technology*, **34**(9): 3533–3543, doi: 10.1007/s12206-020-0805-0.

20. Ruan K., Li Y. (2021), Fuzzy mathematics model of the industrial design of human adaptive sports equipment, *Journal of Intelligent and Fuzzy Systems*, **40**(4): 6103–6112, doi: 10.3233/jifs-189449.

21. Ruan P., Zheng X., Qiu Y., Zhou H. (2022), A binaural MFCC-CNN sound quality model of high-speed train, *Applied Sciences*, **12**(23): 12151, doi: 10.3390/app122312151.

22. Shi M., Lv L., Guo Z., Sun W., Song X., Li H. (2021), High-low level support vector regression prediction approach (HL-SVR) for data modeling with input parameters of unequal sample sizes, *International Journal of Computational Methods*, **18**(08): 2150029, doi: 10.1142/s0219876221500298.

23. Song X., Yang W. (2022), Research on the sound quality evaluation method based on artificial neural network, *Scientific Programming*, **2022**: 1–8, doi: 10.1155/2022/8686785.

24. Wang Y., Zhang S., Meng D., Zhang L. (2022), Nonlinear overall annoyance level modeling and interior sound quality prediction for pure electric vehicle with extreme gradient boosting algorithm, *Applied Acoustics*, **195**: 108857, doi: 10.1016/j.apacoust.2022.108857.

25. Yao Q., Wang Y., Yang Y., Yang L. (2023), DOA estimation using GRNN for acoustic sensor arrays, *Multidimensional Systems and Signal Processing*, **34**(2): 575–594, doi: 10.1007/s11045-023-00877-9.

26. Yasin A., Amin M., Qasim M., Muse A.H., Mastor A.B.S. (2022), More on the ridge parameter estimators for the Gamma ridge regression model: Simulation and applications, *Mathematical Problems in Engineering*, **2022**: 1–18, doi: 10.1155/2022/6769421.

27. Zhan A., Du F., Chen Z., Yin G., Wang M., Zhang Y. (2022), A traffic flow forecasting method based on the GA-SVR, *Journal of High Speed Networks*, **28**(2): 97–106, doi: 10.3233/jhs-220682.

28. Zhu Z., Yin H., Liang Z. (2022), A prediction model for top-coal drawing capability in steep seams based on PCA–GRNN, *Geofluids*, **2022**: 1–9, doi: 10.1155/2022/3590764.