engineering proceedings

MDPI

Proceeding Paper

Design of a Singing Evaluation System of Heyuan Hua Chao Opera Based on Mel-Frequency Cepstral Coefficients †

Shuping Sun *, Yulei Zhu and Yanhui Wang

School of Art, Xiangtan University, Xiangtan 411100, China; zhuyulei0@163.com (Y.Z.); wyh23yan@163.com (Y.W.)
* Correspondence: 18670901605@163.com
† Presented at the IEEE 5th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability, Tainan, Taiwan, 2–4 June 2023.

Abstract: Heyuan Hua Chao opera is affected by modern culture and is facing a limited number of teaching staff, little time available for teachers to be dispatched, a small scope of popularity, the inability for continuous tracking in teaching, unsystematic and coherent learning among students, and difficulty of independent learning. Therefore, the voice feature parameters required for the evaluation model were extracted through MFCC coefficient features and later input into a convolutional neural network to generate data sets and training sets. These feature-labeled sets are segmented to output singing feedback. With a comprehensive overview of the artistic characteristics of Hua Chao opera singing and the research and interviews conducted by the Hua Chao opera heritage development center and local people, a system is designed to evaluate the singing voice of Hua Chao opera singers based on the MFCC. The aim is to effectively help students participate in learning and intelligently evaluate their learning effects. The model can be applied to other opera repertoires to promote the preservation and dissemination of traditional opera culture.

Keywords: Mel-Frequency Cepstral Coefficients; Hua Chao opera; vocal evaluation; opera into the campus

Citation: Sun, S.; Zhu, Y.; Wang, Y. Design of a Singing Evaluation System of Heyuan Hua Chao Opera Based on Mel-Frequency Cepstral Coefficients. *Eng. Proc.* **2023**, *55*, 88. https://doi.org/10.3390/engproc2023055088

Academic Editors: Teen-Hang Meen, Kuei-Shu Hsu and Cheng-Fu Yang

Published: 3 January 2024

1. Introduction

Hua Chao opera is a local traditional drama in Zijin County, Heyuan City, Guangdong Province, and is one of the forms of national intangible cultural heritage. The current opera in schools is active. On this basis, Huachao Opera has built a talent training system of "elementary school–junior high school–high school–university" with a close connection and coordinated development. However, Hua Chao opera on campus still faces the problem of a limited number of teaching staff, small teaching coverage, the inability of teachers to effectively detect students' learning effects in time, and students' lack of systematic and coherent learning and difficulty in independent learning. Considering the current problem of introducing Hua Chao opera to the campus, we decided to explore new ways and ideas of Hua Chao opera culture education and inheritance using new media and design a singing-assessment system based on Mel-Frequency Cepstral Coefficients. The result is expected to improve the initiative of students' participation in learning Hua Chao opera and intelligently assess the learning effect. At the same time, we propose a method of speech classification by fusing MFCC coefficient features with convolutional neural networks, which effectively improves the recognition accuracy and noise resistance of the model and provides a reference for the study of speech processing methods to apply the model to other opera repertoires.

2. Design of the Singing Evaluation System

The design of the Heyuan Hua Chao opera singing-voice-evaluation system mainly realizes the functions of learning–grading–singing voice opinion feedback. The system is

designed to use the singing voice in Huayuan Hua Chao opera as the object of study and establish the evaluation criteria of the vocal character, clear and accurate singing words, and the stable pitch of the tune. By building a database and adopting quantitative methods, i.e., pre-emphasis, and using the fast Fourier transform method, the musical characteristics and speech spectrum information of Hua Chao opera singing are extracted [1]. The inverse spectrum-analysis method is also used to analyze the semantic features of speech signals. The large volume of singing speech signals is standardized and analyzed for objective and scientific evaluation results to achieve the noise resistance, robustness, and accuracy of the evaluation system.

2.1. Core Design

2.1.1. Design Flow of Mel-Frequency Cepstral Coefficients

1. Sample data and labels

The participants of the Hua Chao opera singing -evaluation system were students in grades 3 to 5, and the collected sample data were divided into training and testing audio sets. A total of 20,316 audio items with a chanting style were collected from the training set, while there were 4060 audio items from the training set with a fast singing style and 10,084 audio items from the training set with a traditional singing style. After the audio storage paths of the files were obtained, the labels of the speech signals, i.e., the specific feedback of the three singing voices, were imported.

2. Specific implementation of MFCC coefficient generation speech signal model

After the training set of audio sample data of Hua Chao opera was obtained, the continuous speech signals with a 16 KHz sampling frequency in 1 s were needed to improve the high-frequency part to reduce the noise output of the pre-emphasized overlapped sampling points to reduce the change in frame number in the sub-frame and increase the continuity of the left and right ends of the frame. Then, the spectral leakage of the plus window was reduced, and the time domain signal was converted into frequency domain energy to better observe the signal characteristics of the fast Fourier transform. The speech signal envelope and spectral details were extracted to obtain the sound properties of the cepstrum analysis using five processing methods. Finally, the speech signal model was generated to be used in the Heyuan Hua Chao opera singing-evaluation system (Figure 1).

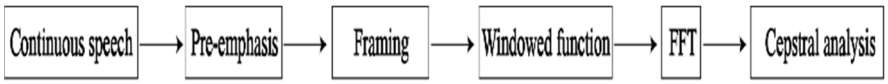

Figure 1. Specific implementation steps of the speech signal model.

The sound interval of Hua Chao opera, such as more step-in, less jump-in, regular rhythm, and high-frequency signal transmission, can easily weaken [2]. Based on the audio sample collection of vocal skills, volume, environment, and radio effect, the signal frequency varies, and a large number of speech signals are piled up in the low-frequency and high-frequency parts. The low-frequency part of the value gap is the large spectrum tilt (Spectral Tilt). The speech signal was taken through a high-pass filter method, a differential first-order signal value to reduce the high-frequency part (high-frequency change position) differential value and increase the low-frequency part (low-frequency change position) differential value so that the speech spectrum maintained a stable state (Figure 2, Equation (1)).

$$y(t) = x(t) - \alpha x(t-1),\ 0.95 < \alpha < 0.99 \quad (1)$$

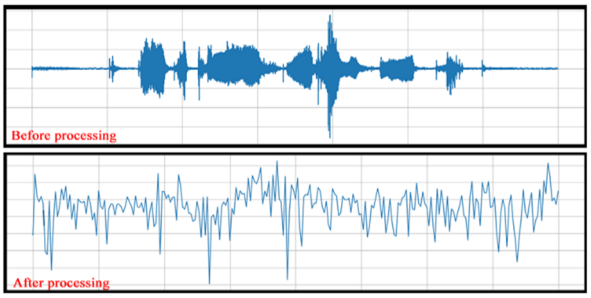

Figure 2. Spectrogram of changes before and after the pre-emphasis method.

The voice in Hua Chao opera is simple and healthy, the lyrics are catchy, and the voice signal changes slowly. The frequency in the voice signal after pre-emphasis becomes unstable with the change in time, so it needs to be processed in frames. However, the length of frames must not be too long or short. An excessive length reduces the time resolution, while a short length aggravates the cost of operation. As the speech signal sampling frequency was 16 kHz, the standard resolution was 25 ms, and the frame length of 16 kHz signal had 0.025 × 16,000 = 400 samples (the chart unit is N). As the frameshift in the speech signal is usually 10 ms, there were 0.01 × 16,000 = 160 samples in the frame [3]. The difference between frame length and frameshift in the overlapping area in the speech signal, the so-called "frame iteration", accounted for about 1/3 of each frame, that is, 400 − 160 = 240 samples (the chart unit is M). The main role of the 240 samples was to maintain the frame number difference. The starting value of the speech frame with 400 sample points was 0, the starting value of the second speech frame was 160, and the starting value of the third speech frame was 320 (Figure 3). According to this rule, the frames were divided until the end of the speech signal.

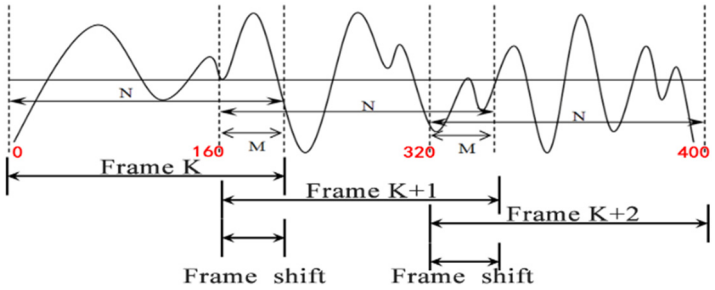

Figure 3. Schematic diagram of frame splitting.

The speech signal of Hua Chao opera is a time series, but the middle domain values are disconnected when carrying out the frame splitting process, which causes the spectral leakage phenomenon. Considering the long-time fluctuation of the speech signal and the disconnection of the endpoints, it is necessary to process its fixed characteristics. Each frame number is substituted into the Hamming Windows function (Hamming Windows), which truncates the original speech sample signal with amplitude–frequency to achieve the finite speech signal and then supplements the continuity of the left end and right end of the frame. By formulating the out-of-window value of the Hamming Windows function as 0 (based on Equation (2)), one takes a generic 0.46 and applies the function to each frame number, the sharp angle in the truncated finite speech signal is blunted, and the waveform amplitude slowly tapers to 0, thus reducing the truncation effect of the speech frames (Figure 4).

$$w(n,\alpha) = (1-\alpha) - \alpha cos\frac{2\pi n}{N-1}, (0 \le n \le N-1) \quad (2)$$

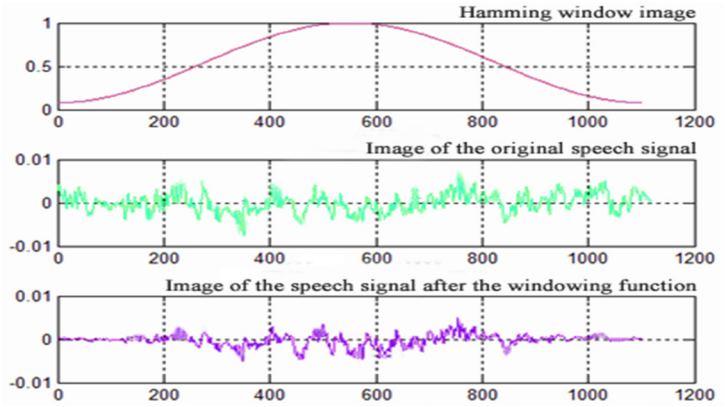

Figure 4. Schematic diagram of the Hamming Window function.

After the speech signal is substituted into the Hamming window function, the individual frames in the signal need to be fast-Fourier-transformed, because the melodic direction of the Hua Dynasty opera tune is rich and variable. It is more difficult to capture its dynamic information when the speech signal belongs to the one-dimensional signal. Thus, the time-domain information cannot be seen as frequency-domain information. Therefore, the speech signal must be processed and output into a 500 × 20 two-dimensional matrix model group as the MFCC matrix model generation code (Figure 5). After that, the speech signal is arranged according to frame, and each frame of speech corresponds to a spectrum using coordinates. In the horizontal coordinate, the duration and vertical coordinate become the decibels of the sound, indicating the relationship between frequency and energy. The amplitude of the speech signal is mapped to a gray-level embodiment, and the amplitude value is proportional to the corresponding gray area. The larger the amplitude value, the darker the corresponding area. The display duration of the speech spectrum is increased to obtain a change over time depicting the sound spectrum of the speech signal (Figures 6 and 7). By doing this, its energy distribution on the spectrum is obtained, and the influence of frequency points higher than the sampled signal is removed to make static and dynamic information parameters more intuitive for the improved robustness of speech signal data acquisition.

```
def wav2mfcc(path, max_pad_size=11):
    y, sr = librosa.load(path=path, sr=None, mono=False)
    if len(y) == 2:
        y = y[0, :]
    audio_mac = librosa.feature.mfcc(y=y, sr=16000)
    audio_mac = cv2.resize(audio_mac, (500,20))
    y_shape = audio_mac.shape[1]
    if y_shape < max_pad_size:
        pad_size = max_pad_size - y_shape
        audio_mac = np.pad(audio_mac, ((0, 0), (0, pad_size)), mode='constant')
    else:
        audio_mac = audio_mac[:, :max_pad_size]
    return audio_mac
```

Figure 5. MFCC matrix model generation code.

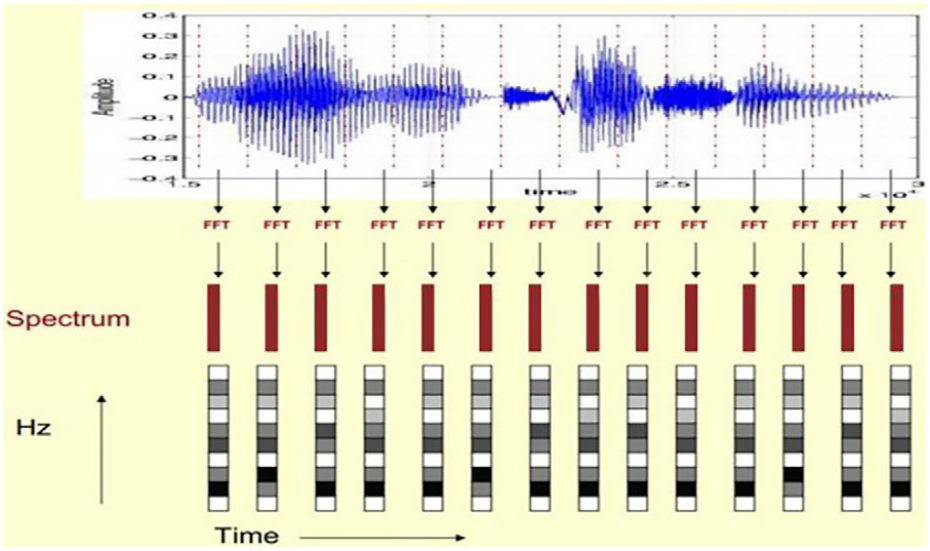

Figure 6. Fast-Fourier-transform process.

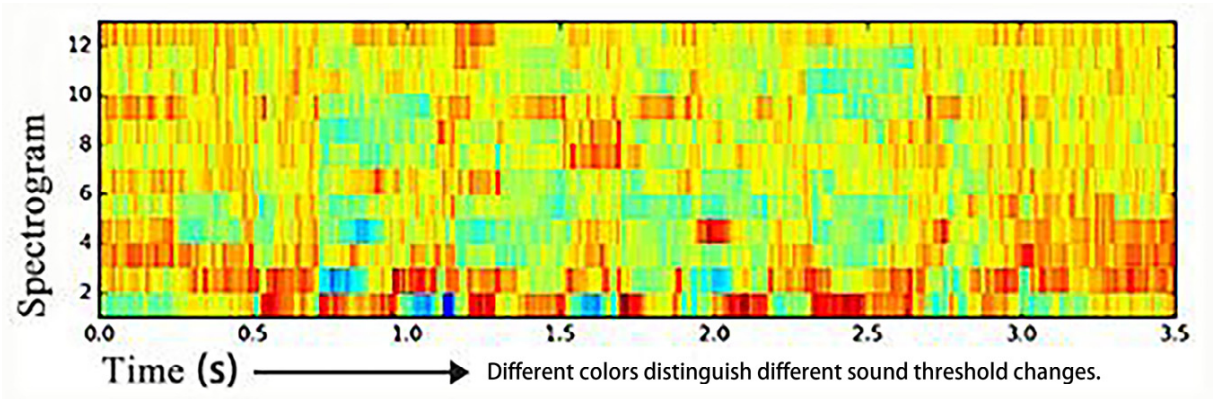

Figure 7. Sound spectrum after conversion.

Since Hua Chao opera is sung in the Zijin dialect and its transcription recognition is poor, it is important to perform cepstral analysis on speech signals to extract speech components. The peak resonance peaks on the speech spectrogram carry important sound-recognition properties, while the smooth curve connecting the resonance peaks is called the envelope [4]. To better identify the acoustic properties of the speech signal, the envelope needs to be separated from the peaks using inverse spectral analysis. Considering the peak as a detail of the spectrum, the horizontal axis is considered as the envelope E of the low-frequency component, which is considered as a sinusoidal signal with four cycles per second, giving a peak at 4 Hz on the coordinate axis. The vertical axis is the spectral detail H[k] of the high-frequency component, which is considered as a sinusoidal signal with 100 cycles. A peak is assigned to the position of 4 Hz on the pseudo-coordinate axis. The vertical axis is the spectral detail of the high-frequency component, H[k], viewed as a sinusoidal signal with 100 cycles per second, and a peak is assigned at 100 Hz on the coordinate axis (Figure 8).

```
def train(model, device, train_loader, optimizer, epoch):
    model.train()
    running_loss = 0.0
    cnt = 0

    loss_sum = 0
    loss_cnt = 0

    for batch_idx, (inputs, labels, file) in enumerate(train_loader, 0):
        inputs, labels = inputs.to(device), labels.to(device)
        # warp them in Variable
        inputs, labels = Variable(inputs), Variable(labels)
        # zero the parameter gradients
        optimizer.zero_grad()
        # forward
        outputs = net(inputs)
        # loss
        loss = criterion(outputs, labels)
        # backward
        loss.backward()
        # update weights
        optimizer.step()

        running_loss += loss.item()
        cnt += 1

        loss_sum += loss.item()
        loss_cnt += 1
        # print("batch_idx   :  ", batch_idx)
        if (batch_idx + 1) % 100 == 0:
            #print(1.0 *  batch_idx/(len(train_loader)))
            print('Train Epoch: {} [{}/{} ({:.0f}%)]\tLoss: {:.6f}'.format(
                epoch, batch_idx * len(inputs), len(train_loader.dataset),
                100.0 *  batch_idx / (len(train_loader)), running_loss/cnt ))
            running_loss = 0.0
            cnt = 0
    return loss_sum/loss_cnt
```

Figure 8. MFCC training function process.

Firstly, the multiplicative signal (x[k]) is turned into a multiplicative signal via the convolution of the two superpositions; then, the multiplicative signal is transformed into an additive signal by taking the logarithm. Finally, an inverse transformation is performed, and the additive signal is restored to a convolutive signal, which is returned to the matrix model to generate a new speech signal model (Figure 9).

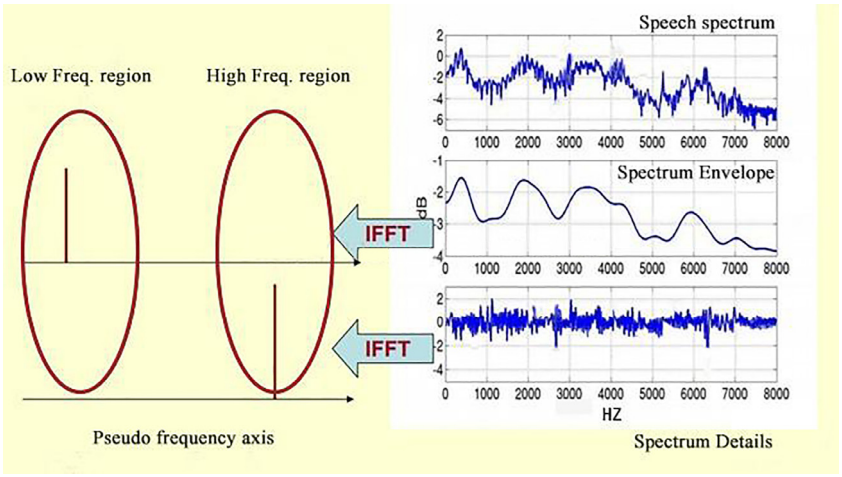

Figure 9. Inverse spectrum analysis process.

3. Accuracy testing of speech signal training models

The training model is built on python, and the training set of audio sample data of the three singing voices collected in the early stage is imported into the model. There are 458 test audio samples of chanting-style singing, 124 test audio samples of fast-singing-style singing, and 301 test audio samples of traditional-singing-style singing. Then, the processed training set audio signals are imported into the test model together with the relevant speech parameters, and the accuracy of a test result is generated in the model through constant comparison of the two parameters. If the deviation between the two data is large, the model reverses the transmission to adjust the parameters of the speech layer until the accuracy rate rises, and the accuracy of the tested speech signal model is 65.1% (Figure 10 shows the test function and the resulting accuracy results).

```
def test(model, device, test_loader):
    model.eval()
    test_loss = 0
    correct = 0
    with torch.no_grad():
        for inputs, labels, files in test_loader:
            inputs, labels = inputs.to(device), labels.to(device)
            output = model(inputs)
            test_loss += criterion(output, labels)  #
            pred = output.max(1, keepdim=True)[1]  #
            correct += pred.eq(labels.view_as(pred)).sum().item()

            # for i in range(len(pred)):
            #     print(output[i]," :  ",pred[i]," : ", labels[i])
            # break

    test_loss /= len(test_loader.dataset)
    print("\nTest set: Average loss: {:.4f}, Accuracy: {}/{} ({:.4f}%) \n".format(
        test_loss, correct, len(test_loader.dataset),
        100. * correct / len(test_loader.dataset)
    ))
    temp = 100. * correct / len(test_loader.dataset)
    return temp, test_loss
```

Figure 10. Test function and accuracy results.

2.1.2. Convolutional Neural Network for Classification Processing

In the two-dimensional matrix processing of the speech signal in the training model and the frequency-domain information processing of the speech signal, the classification method of the convolutional neural network is integrated by using convolutional fixity to suppress the multiplicity of the speech signal [5]. The main principle is to input the corresponding speech signal parameters, read the parameters through a layer of convolution, process the convolved information using pooling, and then perform the same operation as in the previous step. The secondary processed information is passed into the two neural layers (fully connected), which form an unusual two-layer neural network layer. Finally, a classifier is used to connect the layers for classification.

The convolutional neural network is not limited by having to process the input information of each pixel. Based on its characteristic of having a batch filter that continuously rolls over the two-dimensional graph, the information is collected in the graph. The fusion of MFCC coefficients is used to extract feature parameters and implement classification processing for each small block of the pixel area in the two-dimensional matrix model. The fusion enhances the continuity of vocal information in the speech signal such that the neural network can acquire graphical information, not a single pixel point, and deepens the neural network's knowledge of graphical information, allowing the classification and collation to have an actual presentation.

2.2. Back-End Data Acquisition and Processing Design

2.2.1. Overall Construction of the Back-End Framework

4. Implementation of algorithm scoring function and overall algorithm deployment

The singing evaluation system is based on the testing algorithm of speech recognition to complete the communication connection between the scoring system and the back-end port. The standardized speech and the test speech signal are compared and identified based on melody and number of words. The difference in degree between the two is obtained by calculating the ratio of the two. The smaller the value of the different degree, the higher the matching degree. The test data shows that nearly 70% of the test speech signals are higher than 60 points, of which the average score is 67 and the highest score is 90.

The choice of back-end server is a Linux system, with the use of FTP terminal simulation software (v5.0.1221). With nohup background running services, the unit is used in the form of numbers instead of singing feedback comments. Different values represent different singing feedback comments. These parameters are requested for the audio path of the server to access the relevant score and numerical parameters, and finally, the prediction results are returned to the WeChat applet via HTTP to achieve data communication.

5. Overall back-end deployment

The core part of the back-end development is to open the communication with the algorithm related to singing assessment, the processing of the audio reception path, and the collection of different types of singing learning data and answering data from local users to link the system-affiliated functions. The specific implementation process includes three core modules: the data-collection module, the data-analysis module, and the data feedback module. The data-collection module collects users' audio data and historical behavior record information, including scoring parameters of singing voice learning, feedback parameters, question quantity, and accuracy rate for questions. The data-analysis module starts the MFCC algorithm deployed on the server to analyze the audio and answer data set output by the data-collection module. The data feedback module analyzes the information dataset traversed by the module, stores it in the MySQL database twice, and returns the response results to the application-side display on the WeChat applet through the path. At the same time, the API interface is called to obtain the user login name, school, and other identifying information to build the user experience record dataset.

2.2.2. Front-End and Back-End Data Communication Implementation

The whole front-end and back-end communication-implementation method is mainly divided into three steps. The front-end logic is to build the official API in the applet: wx. request () interface, direct calls js logic function. Then, the back-end server receives and processes the request passed by the front end. In the process of returning the response data, it runs the web framework program derived from Flask under python to handle network reception and sending and other behaviors. The background interface accesses the database and the database-related data, which are returned to the front-end through JSON format and the final front-end HTML for the display of feedback results.

3. Singing Voice Evaluation System Test

The carrier of the Heyuan Hua Chao opera singing-evaluation system is the WeChat applet, and the hardware environment to test the system is Android and iOS. The test included functional and performance tests of the applet. The functional test of each module meets the business requirements and the expected results. The system's performance test is carried out for the applet's response time, concurrent users, throughput rate, business success rate, error rate, thinking time, and system resources. The performance test report shows that the maximum concurrent access results in 1 s include an access-success rate of 100% for 80 users, and the system is running stably. CPU memory occupation is not more than 50%, and the number of errors reported in a 1 s response time is 0. This shows that the Heyuan Hua Chao opera singing-assessment system runs normally and stably.

4. Conclusions

At this stage, the Heyuan Hua Chao opera singing assessment system has been put into use normally, named "Zhaohua opera", from the harmonic sound "Dawn Blossoms Plucked at Dusk". The main interface comprises a practice, competition, list, and four functional modules (Figure 11). Through the interaction logic of the video-learning–user-recording–audio-playback–system-operation–feedback, it aims to convey the core educational functions of system singing-voice selection, singing-voice experience, and singing-voice feedback. It has the advantageous features of noise resistance, robustness, and accuracy, and users conduct independent learning and receive real-time feedback, which is practical and agile. Traditional opera is a content provider and a content disseminator. By using new media such as WeChat applets, we can expand the scope of the audience and the group base of Hua Chao opera for its survival. At the same time, support for future generations with a database of audio samples of Hua Chao opera is required to inspire more scholars to think about the path for the preservation and dissemination of traditional opera culture.

Figure 11. Core UI interface of "Zhaohua opera" (the language is set to Chinese).

Author Contributions: Conceptualization, Y.Z. and S.S.; methodology, Y.Z.; software, Y.W.; validation, Y.Z., Y.W. and S.S.; formal analysis, S.S.; investigation, Y.W.; resources, Y.Z.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, S.S.; visualization, Y.Z.; supervision, S.S.; project administration, S.S.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: There are no ethical issues involved in this article, as we did not study any human or animal subjects, nor did we collect any personal information or sensitive data.

Informed Consent Statement: Informed consent was obtained from all participants prior to registration for publication.

Data Availability Statement: The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, K.; Zhang, T. A preliminary study on the feature engineering of Mei Lanfang's singing voice using artificial intelligence technology. *Mei Lanfang J.* **2020**, 137.
2. Wu, G. *Selected Essays on Hakka Gupi Studies*; Propaganda Department of the CPC Heyuan Municipal Committee: Heyuan, China, 2010; Volume 191.
3. Li, S.; Wang, X.; Zhang, Y.; Li, H.; Xiang, H. Road rage emotion recognition based on improved MFCC fusion features and FA-PNN. *Comput. Eng. Appl.* **2021**, *9*, 306–313.
4. Wang, X.G.; Zhu, J.W.; Zhang, A.X. A vocal pattern identity identification method based on MFCC features. *Comput. Sci.* **2021**, *12*, 343–348.
5. Long, H.; Zhang, L.P.; Shao, Y.B.; Du, Q.Z. Speaker feature-constrained speech enhancement for multi-task convolutional networks. *Small Microcomput. Syst.* **2021**, *10*, 2178–2183.