International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 3256-3269

https://doi.org/10.70517/ijhsa463274

Innovations in Artistic Expression by Combining Vocal Technique and AI Technology on the Opera Stage

Yu Wang[1,*]
[1] School of Music, Shanghai Normal University, Shanghai, 200233, China
Corresponding authors: (e-mail: 13816749976@163.com).

Abstract The life of opera works lies in the performance, and the performers on the opera stage can give new artistic expression to the artistic performance of the opera works through the expression of effective vocal skills and the integration of them with AI technology. On the basis of combing the vocal singing skills on the opera stage, the article discusses the relevant training methods of vocal singing skills. Then the vocal singing voice data is collected with 10 different types of opera works, and the signal preprocessing is carried out through the methods of pre-emphasis, frame-splitting, endpoint detection, etc., and then MFCC is used to extract the emotional features of the vocal singing voice. As the input of CLDNN network, different emotional features are fused with dimensionality reduction and weighting through linear layer, and SAGAN model is introduced to mine the temporal relationship in emotional features, and then realize the recognition and classification of emotional features of vocal singing speech. The results show that the CLDNN-ASGAN model's classification accuracy of vocal singing emotional features reaches 88.75%, and the average score for different types of opera vocal singing speech reaches 91.99, with a difference of only 0.18 points from the score of professional listeners. Utilizing AI technology to assist vocal performers in clarifying the emotional performance of their works during the performance process can promote the enhancement of the artistic expression of opera works.

Index Terms AI technology, signal preprocessing, MFCC, CLDNN network, SAGAN model, vocal singing technique

I. Introduction

Opera, as a comprehensive art that integrates music, theater, fine arts and dance, has attracted countless audiences with its unique artistic charm since its birth. In the performance of opera, vocal stage performance is not only the core of its artistic expession, but also an important bridge connecting the work and the audience's emotions [1]. The use of vocal skills is not only the basis of the singer's performance of musical rhythm and dramatic depth, but also directly related to the artistic charm and infectious force of the whole work [2]-[4]. In addition, modern opera singing also incorporates new vocal techniques and performance forms to adapt to the ever-changing artistic requirements, such as the singing style that combines elements of popular music, aiming to be close to the aesthetics of modern audiences [5]-[7]. The application of these vocal techniques enables opera singers to deeply excavate the artistic value of the work, truly convey the emotions of the characters, and bring the audience emotionally rich, audio-visual three-dimensional artistic enjoyment [8], [9].

With the development of the times, the performing art of opera continues to evolve, from the initial religious theater performances to the later court entertainment, and then to the modern public theater performances, the expressive power of opera is constantly being expanded and deepened [10]. Artificial intelligence (AI) technology has been widely used in artistic creation, and immersive sound design has become an important driving force for innovation in musical theater [11], [12]. Especially in the field of sound generation and emotional expression, AI not only improves the efficiency of audio generation, but also makes the interaction between sound and emotion reach an unprecedented degree of precision [13], [14]. Through the deep integration of multimodal data, AI-driven sound design is no longer limited to the generation of traditional sound effects, but promotes the comprehensive innovation of the musical theater art form through the combination of emotional computing and auditory perception [15]-[17].

In the process of opera stage performance, AI technology is used to optimize the performer's vocal skills as a way to enhance the artistic expression of the opera stage and reveal the unique emotional communication and cultural value of opera stage art. The article firstly discusses the importance and classification of vocal skills on the opera stage, and sorts out the importance of vocal singing skills training and related methods. Secondly, ten different types of opera works are selected, and 20 professional opera performers are invited to collect vocal singing voice

data, and extract the voice emotional features in vocal singing through preprocessing and features. Then, CLDNN network is used to fuse the multi-source emotional features, and SAGAN model is chosen to mine the temporal information of the emotional features, and then the emotional features are output and the vocal singing scoring is carried out. Finally, the effectiveness of the model is verified by the comparison of emotion recognition effectiveness and scoring.

II. Vocal singing techniques on the opera stage

Opera stage performance is a kind of stage art, which can make the audience have strong resonance and emotional experience through the performer's singing to express the storyline and character image. Performers need to give life to the work through their own understanding of the emotion and content of the work, combined with their own emotions and body language expression and vocal singing skills to mobilize the audience's emotions, so that the audience and the performer have emotional resonance. Vocal singing skills and opera stage performance have an important connection, and there are similarities between the two in terms of artistic expression. Being able to effectively integrate vocal singing skills into opera stage performance is of great significance in enhancing the effect and artistic expression of opera stage performance.

II. A.Importance and Classification of Vocal Technique

II. A. 1) Techniques related to vocal singing

(1) Breathing techniques. In terms of opera performance, lyrical slow song passages predominate. The singer needs to convey the emotion of the lyrics to the listener, and at the same time cooperate with the singing and breathing to achieve seamless integration of phrases and make the music more beautiful. Breathing control and distribution should be done well when singing, you can take a light closing of the lips, the body slightly tilted, the use of exhalation to drive the vibration of the lips, so as to train the coordination of pitch and breathing. Or inhale deeply until the body expands, focus the sound on the edge of the teeth, gently hiss. During exhalation, the ribs naturally return from the expanded state to the normal state. This technique helps to maintain sustained breathing, which is beneficial for vocal exercises and actual opera performances.

(2) Soprano Control Technique. In the art of vocal performance in opera, the mastery of the high register is not only a difficulty for the artist to show his/her skills, but also a manifestation of his/her artistic attainments. In order to ensure the fullness and strength of the singing voice, regular breathing rhythm is the foundation. Before mounting the opera stage, the singer should have a deep understanding of the structure and physiological function of the larynx. With a thorough understanding of the vocal mechanism, the efficiency of breathing and soprano training will be significantly improved, enabling the singer to accurately locate the vocal parts. Soprano singing requires the rear of the vocal folds to be in a taut state, and also involves synchronized control of the front of the body, which is directly related to the precision of vocal fold manipulation. During the training process, singers should focus on opening the pharyngeal cavity, learn to distinguish the difference between vocal fold vibration and throat articulation, and strengthen the vocal folds while effectively controlling the nasal sound [18].

II. A. 1) Importance of vocal technique

Opera works in the creation of its fully embodied comprehensive and artistic, opera elements, music elements, literary elements and dance elements in the opera works are fully embodied, in which the vocal performance in the opera works singing occupies an important position, and opera works singing in the different application of vocal skills is to make the opera works present a sense of the picture and sense of hierarchy of the key. In order to make the opera works present close to perfect performance, the vocalists need to fully understand the whole opera works, and at the same time to master the mature singing skills, through the application of different skills to sing, in order to present the connotation of the corresponding works in front of the audience. On the basis of a full understanding of the content expressed in the work, choose to apply different techniques to interpretation, so as to highlight the connotation of the opera works, with different vocal techniques to present the main line of the opera works and the organization of the opera works, so that the whole work presents a good sense of hierarchy and sense of the picture, the audience's emotions into the appreciation of the opera works, so that they can produce emotional resonance, to get a better viewing experience.

During the interpretation of the opera, the performers will place themselves in the situation of the opera, and present the image and character traits of the characters in the opera with the corresponding interpretation skills and methods. And in the whole context of the opera works during the creation of the singing skills of the vocalist has a very important role, whether the opera works of the theme interpretation, the atmosphere created by the test of the

different use of vocal skills, the quality of its creation determines the audience to listen to whether they can get a good sense of light experience. In addition, the use of different vocal techniques can present a strong expressive power, which contains the content of the opera lyrics effectively presented, thus enhancing the infectious force of the opera works.

II. B.Training Methods for Vocal Singing Skills

II. B. 1) Importance of vocal training

Vocal training plays an extremely important role in the singing of opera works. In vocal music teaching, vocal technique is an important training content, which includes breathing, pitch, timbre, volume and other aspects. In the process of opera singing, in order to let the audience better understand the plot and the role, it is necessary to use the voice vocalization skills to express the character's emotions and inner world. After vocal training, the stability and controllability of the actor's voice are greatly improved. In addition, vocal training can also expand the actor's range, so that he or she can better cope with the high and low voice requirements of the opera and show richer expressiveness [19].

In the process of opera performance, the use of vocal technique has a great influence on the singing effect of the actor. In vocal training, the actor masters the correct use of breath, controls voice strength and stability, and makes opera singing more comfortable. In addition, vocal training can also help singers master the correct vocal posture, improve the coordination of the laryngeal muscles and reduce the risk of vocal cord injury, thus achieving the purpose of protecting the health of the voice. Vocal training is an indispensable part of opera singing, which not only improves the singing skills, but also ensures the health of the voice, so that it can show the charm of the role and infect the audience's heart.

II. B. 1) Vocal Technique Training Methods

(1) Flower-smelling breathing exercise and sighing breathing exercise. For flower-smelling breathing, first of all, keep a relaxed posture, lift up the chest, let the shoulders drop naturally, the waist and the abdomen should be slightly developed, adjust the posture, and then take a deep breath. During the inhalation, let the chest and both shoulders relax, do not sway from breathing, and do not change the height randomly. When the singer does the deep inhalation exercise, the whole person should be relaxed, the chest should not move, and by sniffing the flowers, slowly inhale the airflow into the abdomen, letting the belly expand until the belly can't hold a breath, and hold it for a few seconds. Then slowly exhale, breathing should be smooth, do not rush, breathing maintained for a long time to have better results.

(2) Humming training. Humming is the root of resonance, the advantage of humming is that it can relax the pharyngeal cavity and avoid the production of guttural sounds. In the humming practice, should be taken down three degrees, and then transition to five degrees, it is recommended that the humming is placed in the second project of the daily vocal training, after practicing the humming, the humming part to join the vowel training.

(3) Skipping tone training. Breathing should be done in a relaxed manner and vocal exercises for the diaphragm should be used. Skipping is a proven training method for training the diaphragmatic fulcrum because each syllable requires a significant pressure of movement of the diaphragm in order for the air to be blown across the vocal cords in a passive state.

(4) Chest-pounding training for mutes. The “chest pounding” is performed without accompaniment. While producing the long “a” sound, the chest is rhythmically slapped to artificially force the vocal cords to produce the sound, stretching them tightly and making them flutter rhythmically. At this stage, we add breath-holding underneath the sound, and this is the high note we are aiming for.

III. Emotional feature recognition for vocal singing combined with AI

Throughout history, just like all art varieties, Chinese opera has gradually matured in a long and difficult development process. From the content of the opera to the form of expression, from the rich national style to the superb creative skills, from the excellent singing skills to the perfect stage performance, all of them show the distinctive national color, and its regional, historical and period characteristics, with immeasurable artistic value and strong artistic vitality. In the rapid development of new technology today, how to realize the effective integration of vocal skills and AI technology on the opera stage, in order to enhance the vocal singing skills of the singers, has become an inevitable choice to further enhance the artistic expression of opera.

III. A. Voice Acquisition and Processing for Vocal Singing

III. A. 1) Voice Acquisition for Vocal Singing

For the purpose and requirements of this experiment, a dataset of vocal singing in a quiet recording environment is needed to complete the recognition of emotional features of vocal singing, while the dataset of vocal singing that can be found on the Internet is relatively small and of uneven quality, which can not meet the experimental requirements, so it is necessary to manually record the vocal singing voice library. In order to record conveniently, the client software with interface display is firstly written in C# language, and the software contains the voice recording and saving function, which is developed based on the PortAudio library.PortAudio is an open-source audio I/O library, which can be operated on various platforms such as Windows, Linux, and so on, and the development and implementation of audio programs can be simplified by using it.

Ten different types of opera repertoire were provided in the recording software, with a total of 20 opera performers participating in the recording, and each opera performer was required to sing each type of opera repertoire 10 times each. The recordings were made in a laboratory environment using a Takstar TS-6700 microphone, and a total of 2000 opera singing voices were recorded as the source data set for the emotional feature recognition of vocal singing techniques.

The recorded voices need to be preprocessed before they can be used as input to the network. Firstly, all the folders are traversed by depth-first search, and the naming of the files is unified according to the format of “name/filename.wav”. Noisy data are then removed, including incomplete recordings, corrupted data that cannot be read, and data with a recording length greater than or less than 20 seconds, etc. However, removing only these noisy data may jeopardize the balance of the data. Therefore, for a certain noise data, recordings of the same opera surface from other recorders were randomly used to replace it, to avoid the sample imbalance problem affecting the experimental results.

III. A. 1) Voice processing for vocal singing

In order to process the information carried in the speech signal of non-smooth time-varying signals, the input speech signal first needs to undergo speech signal preprocessing, which generally includes sampling, pre-emphasis, frame-splitting, windowing and other operations [20]. Sampling refers to the selection of a fixed time interval to detect the analog value of the analog signal, followed by quantization operations.

(1) Pre-Emphasis and Framing

Pre-emphasis operation is a compensation method that utilizes the information of the vocal tract and reduces the energy loss of the high-frequency portion of the speech, thus improving the speech quality.

Let the i th sampling point of the input speech signal be $w[i]$, and the sampling point of the output speech signal be $y[i]$, the pre-emphasis operation formula is as follows:

$$y[i] = w[i] - aw[i-1] \tag{1}$$

where a is the pre-emphasis coefficient, $a \in [0.9,1]$, generally $a = 0.97$.

Segmentation refers to the segmentation of the speech signal for short-time analysis, and the frame length is generally 10-30 ms. Windowing refers to the addition of a window function before and after the framing of the signal to inhibit the inter-frame time edge effect and ensure the continuity of the speech signal between each frame. Commonly used windowing methods include rectangular window, Hamming window and Hanning window. Among them, the Hamming window can better retain the frequency characteristics of the speech signal and is most widely used, and the Hamming window is defined as follows:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\dfrac{2\pi n}{N-1}) & 0 \le n \le N-1 \\ 0 & \text{other} \end{cases} \tag{2}$$

where $w(n)$ denotes the window function and N denotes the window size.

(2) Speech endpoint detection

In the actual speech signal, it usually contains many silent regions (e.g., silence, noise, ambient sound, etc.), and in practical applications we are usually only interested in the part of it that contains speech. The purpose of endpoint detection is to remove the silent regions and retain the useful part of speech, thus reducing the amount of computation for subsequent processing and improving the efficiency and performance of the system. The energy-based approach is one of the simplest and most commonly used endpoint detection methods, which determines the presence of speech activity by calculating the energy of the speech signal, which can be regarded as the

beginning of speech activity when the energy of the speech signal exceeds a predetermined threshold. When the energy drops below another threshold, it can be regarded as the end of speech activity. Common energy calculation methods include short-time energy, short-time over-zero rate and so on.

Assuming that $x_n(m)$ is the audio signal obtained from the nth frame after windowing and framing, then the mathematical expression of $x_n(m)$ can be expressed as:

$$x_n(m) = w(m)x(n+m), 0 \le m \le N-1 \tag{3}$$

where $w(m)$ represents the window function, $n = 0, T, 2T, ...N$ denotes the frame length of the audio signal, and T represents the frame shift.

Assuming that E_n is the short-time energy of $x_n(m)$, we have:

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \tag{4}$$

To minimize the arithmetic error, the short-time average amplitude is substituted for the energy of the audio signal, viz:

$$M_n = \sum_{m=0}^{N-1} | x_n(m) | \tag{5}$$

The short-time zero crossing rate Z_n indicates how often the signal crosses the zero point in a short period of time. It is the number of times the signal crosses the horizontal zero line in a short time window, so the short time zero crossing rate of the audio signal $x_n(m)$ can be expressed as:

$$Z_n = \frac{1}{2} \sum_{m=0}^{N-1} \mathrm{sgn}[x_n(m)] | - \mathrm{sgn}[x_n(m-1)] \tag{6}$$

where $\mathrm{sgn}[]$ is the sign function with the expression:

$$\mathrm{sgn}[x] = \begin{cases} 1, x \ge 0 \\ -1, x < 0 \end{cases} \tag{7}$$

The short-time-over-zero rate provides some degree of information about the sound change and can therefore be used to detect where the speech signal begins as well as where it ends.

III. A. 1) Vocal singing feature extraction

(1) Time-domain characteristics of vocal singing speech

The time-domain characteristics of vocal singing speech refer to the time-domain parameters of each post calculated from the music signal, which are mainly represented by short-time average amplitude and short-time autocorrelation function in this paper. The details are as follows:

Short-time average amplitude. As the calculation of short-time energy all need square operation, which enlarges the difference between large and small amplitude, thus can not accurately reflect the characteristics of the signal short-time energy change with time. For this reason, another feature to characterize the time-varying characteristics of the signal energy is proposed - the short-time average amplitude, which is defined as:

$$M_n = \sum_{m=n-(N-1)}^{n} | x(m)w(n-m) | = \sum_{m=n-(N-1)}^{n} | x(m) | w(n-m) \tag{8}$$

The short-time autocorrelation function is defined as:

$$R_n(k)=\sum_{m=n-(N-1)}^{n}[x(m)w(n-m)]\{x(m+k)w[n-(m+k)]\}\underline{\underline{n-m=m}}\sum_{m=0}^{N-1-k}[x(m+n)w(m)[x(m+n+k)w(m+k)] \tag{9}$$

where k is the autocorrelation lag time. The above equation shows that $R_n(k)$ of each frame of the signal is a sequence with lag time k as the independent variable.

The short-time autocorrelation function can clearly reflect the periodicity of the signal, so its curve can be utilized to discriminate between turbid and clear tones, as well as to estimate the fundamental frequency.

(2) Frequency domain characteristics of vocal singing voice

The frequency domain characteristic of vocal singing voice refers to the time domain signal of vocal singing voice is transformed into frequency domain signal by Fourier transform, and then the frequency domain parameter of each frame is obtained from the frequency domain signal. The short-time Fourier transform of the vocal singing speech signal can be expressed as:

$$X_n(e^{jw})=\sum_{m=n-(N-1)}^{n}x(m)w(n-m)e^{-jwm} \tag{10}$$

That is, it is the Fourier transform of a frame of signal selected by a moving window $w(n-m)$, where N is the width of the window function. The short-time Fourier transform $X_n(e^{jw})$ is a function of frequency w and time n, which is able to express the slow change of the spectrum of the music signal over time.

(3) Cepstrum characteristics of vocal singing speech

In this paper, we mainly extract the cepstrum features of vocal singing speech by Mel frequency cepstrum coefficient (MFCC) [21]. Assuming that $x(n)$ denotes the input music signal, the complete solution process of MFCC features can be realized by the following steps:

Step1 Perform a series of pre-processing on the vocal singing speech signal, including pre-filtering, quantization, pre-emphasis, end-point detection, etc.

Step2 Perform a short-time Fast Fourier Transform on the input vocal singing speech signal to transform it from the time domain to the frequency domain to obtain the spectrum with the following equation:

$$X(m)=\sum_{n=0}^{F-1}x(n)\times w(n)\times e^{-j2n\pi\frac{m}{F}} \tag{11}$$

where $m=0,1,\cdots,F-1$, F is the frame size, and $w(n)$ is the Hamming window function.

Step3 Find the spectrum squared, i.e., the energy spectrum. I.e:

$$Y(m)=|X(m)|^2 \tag{12}$$

Step4 Define M Mel filters, here triangular filters are used and calculate the energy output from each Mel filter. I.e:

$$S[k]=\sum_{m=0}^{\frac{F}{2}-1}w_k(m)\times Y(m) \tag{13}$$

where $1\le k\le M$, M is the number of Mel filters, and $w_k(m)$ is the triangular weighting function associated with the kth Mel filter.

Step5 Take the logarithm of the output $S[k]$ of each of the M filters and perform a discrete cosine transform to find the L MFCC coefficients. That is:

$$c[n]=\sum_{k=1}^{M}\log(S[k])\times\cos\left[n\times(k+0.5r)\times\frac{\pi}{M}\right],n=1,2,\cdots,L \tag{14}$$

III. B. CLDNN-SAGAN Emotion Recognition

In the process of opera stage performance, how to effectively ensure that the performer's vocal singing skills can better meet the artistic performance effect of the opera stage is the key to enhance the effect of opera performance. Based on this, this paper proposes a vocal singing emotion feature recognition model that combines CLDNN network and SAGAN model, and constructs a vocal singing scoring module, so as to help performers better master the emotional changes of the opera. Figure 1 shows the basic framework of the CLDNN-SAGAN vocal singing emotion feature recognition model.The CLDNN network was first proposed to generate for audio signals, mainly to solve the problem of the original RNN module's excessive computational volume, and then the number of convolutional layers used is small, the size of the convolutional kernel becomes large, and the choice of inflated causal convolution, so that this way of processing can indeed reduce the network's complexity. However, this approach does reduce the complexity of the network, but also brings certain problems, the features extracted and learned are limited to low-dimensional rough feature vectors. Some researchers have confirmed that the superposition network model with complex neural networks can be helpful for feature extraction, so this paper combines the two modules of CLDNN network and SAGAN.

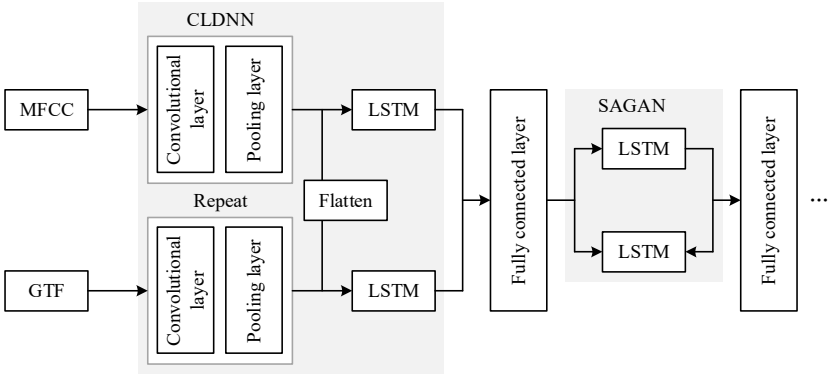

Figure 1: CLDNN-SAGAN Vocal performance Emotional Feature recognition

After the vocal singing emotion features have passed through the CLDNN network, they will be inputted into the SAGAN network for the final regression prediction, and this paper chooses the direct isotropic connection.The most important feature of the CLDNN-SAGAN network is that the two parts of the model are differently divided, and emotion extraction is put in the CLDNN network, and regression prediction is put in the SAGAN. As the artificial neural network of continuous emotion model, firstly, the inflated RNN layer in the CLDNN network is able to design the repetition module according to the actual number of input features, the dimensionality reduction and weighted fusion of the two emotion features is responsible for by the linear layer, the mining of temporal relationship in the emotion features is responsible for by the SAGAN, and normalization and prediction processing of the regression data is completed by the final fully connected layer. Finally, the classification results are input into the vocal singing scoring module to assist performers in understanding the defects in the vocal singing process, so as to better improve their vocal skills and enhance their artistic expression on the opera stage.

III. B. 1) CLDNN Network Architecture

The network structure of CLDNN is shown in Fig. 2, which shows that CLDNN is equivalent to a combined network of CNN, LSTM and DNN.

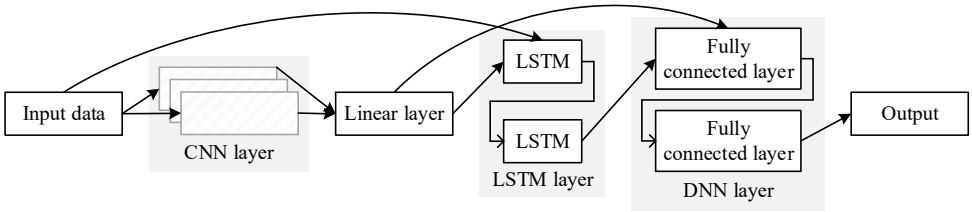

Figure 2: Structural schematic diagram of CLDNN

The CNN layer, into which the preprocessed data first enters, extracts the training features that are correlated across the different data, thus eliminating frequency domain variations in the feature input. Since the extracted features are often too large, the linear layer can be downscaled for this. The subsequent LSTM layer combined with the original preprocessed data can summarize the before and after information of the previous inputs more efficiently, thus making the output more predictable. The final DNN layer increases the depth of the previous hidden layer and the subsequent output layer, fuses the features processed by the CNN and LSTM separately, and converts the deep information into a feature space to make it easy to classify.

III. B. 1) SAGAN model

In this section, the SAGAN module is constructed based on the traditional GAN network with the addition of the self-attention mechanism module, because according to recent studies, effective regulation of the generator can affect the performance of the GAN network, so the spectral regularization is added to the generator of the GAN network in order to get better results [22]. Self-attention module in the statistical efficiency, computational efficiency and the ability to process long sequential inputs can go and a good balance, so this paper constructs self-attention generative adversarial network (SAGAN). The specific computational process is as follows:

In this paper, notation $W_g \in R^{\tau * c}, W_f \in R^{\tau * c}, W_h \in R^{c * c}$ are the learned weight matrices, all of which are convolved by 1*1 to denote by $\beta_{j,i}$ the j th region when the The degree of influence of the model on the i th position is given by the following formula:

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, s_{ij} = f(x_i)^T g(x_j) \tag{15}$$

Output $o = (o_1, o_2, \ldots o_j, \ldots o_N) \in R^{C*N}$ of the concern layer, where:

$$o_j = \sum_{i=1}^{N} \beta_{j,i} h(x_i), h(x_i) = W_h x_i \tag{16}$$

In addition to this, in the process of outputting o, the scale parameter γ needs to be multiplied and added to the input, and the overall output is shown in the following equation:

$$y_i = \gamma o_i + x_i \tag{17}$$

First γ is initialized to 0, after which progressively more weights are assigned to non-local features.

III. B. 1) Loss Functions and Fusion Strategies

(1) Loss function design

In this paper, the maximum soft loss function and the center loss function are chosen to jointly achieve the final classification. The maximum soft loss function can separate the deep features between each class, which are distributed in a narrow band. The center loss can learn a center for the deep features of each class and penalize the distance between the deep features and their corresponding class centers, which results in a compact distribution of deep features within the class. Specifically, the maximum soft loss function has to ensure the maximum distance between classes, while the center loss function has to reduce the distance between samples within classes, and they have to reach a balance between them.

The maximum soft loss function and center loss function formulas are denoted as respectively:

$$\zeta_S = -\sum_{i=1}^{\gamma} \log \frac{e^{W_{y_i}^T X_i + b_{y_i}}}{\sum_{j=1}^{\eta} e^{W_{y_j}^T X_i + b_j}} \tag{18}$$

$$\zeta_C = \frac{1}{2} \sum_{i=1}^{\gamma} \| X_i - c_{y_i} \|_2^2 \tag{19}$$

where γ denotes the number of training samples in each batch and η denotes the number of classes. c_{y_i} is the center of all y_i class samples corresponding to the i th sample. In order to make c_{y_i} update in real time with the features to be learned and to avoid misclassification, this paper adds a parameter λ to limit the value of the center loss function to achieve a balance between the two loss functions. ζ is the final function of loss in this paper. Namely:

$$\zeta = \zeta_S + \lambda \zeta_C \tag{20}$$

The larger the hyperparameter of λ set, the smaller the spacing within the class.

(2) Emotional feature fusion strategy for vocal singing

In this paper, feature level fusion is utilized to learn the interaction information between the higher order features extracted by the feature encoding network. In feature-level fusion, feature extraction methods are used to generate derived representations of lower-level features, which are then combined. The fused feature vectors have to be combined in such a pattern that they enhance the recognition of individuals through the combination of multiple features. First, a set of different low-order features such as timbre, tempo, migration features, Mel spectrograms, harmonic spectrograms, impact spectrograms, scattering transform spectrograms, etc. are extracted from the original audio signal. Each low-order feature is then fed into a separate feature encoding network to extract higher-order features.

If these higher order features $f_\tau^1, f_\tau^2, ..., f_\tau^v$ are to be combined by concatenation, where τ is the dimension size and υ denotes the number of these higher-order features, then the f_{fused} features will be written as Eq. (21):

$$f_{fused} = f_{u \times \tau} \tag{21}$$

In order to learn the interaction information between the higher-order features extracted by the feature encoding network, a global CNN classifier needs to be built to predict the final classification of these higher-order features. Since the input higher-order features are a one-dimensional vector that conforms more to the temporal structure than to the spectral structure, one-dimensional convolution is applied to the CNN classifier in this paper. In addition, the self-attention mechanism is used to learn the dependencies between heterogeneous high-level features.

IV. Vocal singing emotional characteristics identification validation analysis

Vocal singing plays a significant role among opera stage performers, as it is a key indicator of the maturity of the performer, but also an element that ensures that the work effectively conveys emotion. Evaluation of the effectiveness of opera singing generally focuses on two dimensions. One is the vocal technique of singing, and the other is whether the content of the drama and music can be accurately expressed. Therefore, the artistic treatment of vocal works in the opera is very critical, and only by fully grasping it, can a high-quality opera be presented in front of the audience, giving them a real appreciation of the feeling. Thus, this paper uses AI technology to assist in realizing the recognition and analysis of emotional characteristics of vocal singing, and combines the corresponding evaluation to illustrate the feasibility of its application.

IV. A. Validity of Emotion Recognition Models

IV. A. 1) Performance of different sample lengths

In order to verify the classification effect of the CLDNN-SAGAN model established in this paper on vocal singing emotion features under different sample lengths, this paper establishes a global CNN classifier in the model, which combines the maximum soft loss function and the center loss function with the model for fine-tuning training. Wav2Vec-C, COLA, TRILL and other models are selected to conduct comparison experiments on the vocal singing speech dataset established in this paper to verify the effect.

The model is trained on the dataset for 20 rounds each with an initial learning rate of 0.001 and a batch size of 512. The number of hidden layers for both the complex spectral feature extractor and the angular phase feature extractor is 32, the feature fusion is configured as 5 blocks, the convolution kernel size is 3×3 with a step size of 2×2, and the LSTM is configured as 128×128. The mask filter has a convolution kernel size of 2 and a step size of 1. The 3D convolution layer has a convolution kernel size of 3×3×3 and a step size of 3. Accuracy (ACC) is chosen as the evaluation index and the comprehensive performance of the model is examined by intercepting different sample lengths. Figure 3 shows the comprehensive performance of the model under different sample lengths.

In terms of classification accuracy, the CLDNN-SAGAN model proposed in this paper performs better in vocal

singing speech classification, with a classification accuracy of around 88.75%, in comparison, the classification accuracy of each comparison model is lower than that of this paper's model. In addition, from the point of view of the comprehensive performance of different sample lengths, with the increase of sample length, the performance of each model is a substantial increase in the accuracy rate, until the sample length of vocal singing speech is more than 55s, the accuracy rate gradually tends to stabilize. This indicates that even if the sample features are incomplete, the model can still extract vocal singing speech features. However, there is still a difference in the recognition effect of local features. The model proposed in this chapter can still learn the difference of vocal singing speech features (with an accuracy of 42.84%) when more than 80% of the samples are missing (with a sample length of 35s), which suggests that the model is able to pull the distance between different samples in the coding space. The other models perform poorly when more samples are missing, and even the BYOL model based on comparison pre-training, because it does not use negative samples to increase the training difficulty, its classification performance is still far from that of the CLDNN-SAGAN model proposed in this paper (with an accuracy of only 15.48%) when faced with a more difficult experiment (with a sample length of 35s). Taken together, the CLDNN-SAGAN model proposed in this paper can effectively extract emotional features in vocal singing speech, and achieve superior performance even in the case of missing samples. This lays a solid foundation for the optimization and enhancement of performers' vocal singing skills, and aids in improving vocal singing performance on the opera stage.

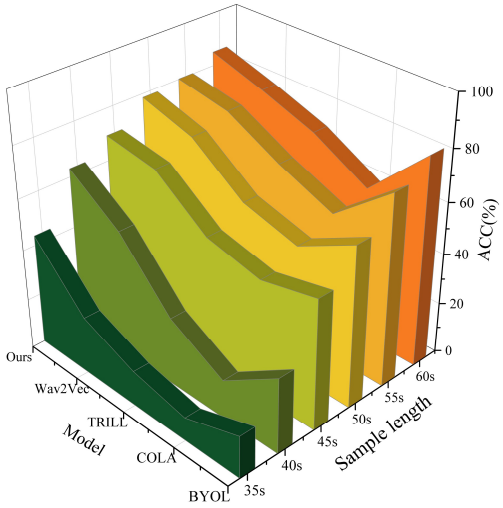

Figure 3: Performance of different sample lengths

IV. A. 1) Analysis of model ablation experiments

In order to verify the necessity of each module in the vocal singing speech emotion feature recognition model designed based on the CLDNN-SAGAN model in this paper, a set of ablation experiments are designed in this paper.The CLDNN-SAGAN model mainly consists of four different modules, namely, the CLDNN network (A), the SAGAN model (B), the loss function (C), and the fusion strategy (D), and for the SAGAN further decomposed into GAN model (B-1) and self-attention mechanism (B-2) to explore the effectiveness of the self-attention mechanism. Unweighted average recall (UAR) and weighted average recall (WAR) are chosen as the evaluation indexes of the model, and five different vocal singing emotion feature recognition tasks (Task1~ Task5) are selected for validation, and the comparison results of UAR and WAR of the individual models in the ablation experiments are shown in Table 1 and Table 2, respectively.

Based on the results in the table, it can be seen that the modules proposed in this paper bring obvious support to the vocal singing emotion feature recognition task, and the average UAR and WAR of the proposed model in the five sets of emotion feature recognition tasks are 55.23% and 55.27%, respectively, which are ahead of the rest of the models in the ablation experiments by 3.92%~18.94% and 3.59%~19.96%. For the ordinary CLDNN model, the limited feature extraction capability is often difficult to adapt to the vocal singing speech emotion recognition task with different feature distributions, while the self-attention network provides effective guidance for the original CLDNN model and enhances its feature extraction capability, resulting in improved model performance. At the same time, the GAN model can help the model to learn the features of vocal singing speech through the adversarial

generative encoder, which significantly improves the recognition performance and stability of the original CLDNN model. When the SAGAN model, loss function and fusion strategy are added into the CLDNN network, the model performance is optimized. The experimental results show that the performance of the feature extractor has an impact on the generalization ability of the model, and an efficient feature extractor ensures that the model extracts the complex higher-order features in the samples while avoiding over-fitting the data. Moreover, the global CNN classifier in the fusion strategy can pre-train the feature extractor and guide the feature extractor to extract more robust invariant features, which further improves the model's vocal singing speech emotion recognition.

Table 1: UAR of each model in the ablation experiment (%)

Method	Task1	Task2	Task3	Task4	Task5	AVG
A	38.42	35.06	38.51	39.14	30.34	36.29
A+B-1	41.06	32.78	39.42	43.47	32.18	37.78
A+B-2	42.35	36.95	42.11	45.79	33.06	40.05
A+B	51.61	41.54	46.93	54.42	39.65	46.83
A+B+C	58.46	45.79	50.35	59.38	42.57	51.31
A+B+C+D	64.27	50.34	52.18	62.52	46.83	55.23

Table 2: WAR of each model in the ablation experiment (%)

Method	Task1	Task2	Task3	Task4	Task5	AVG
A	38.42	30.27	35.28	43.02	29.57	35.31
A+B-1	42.63	33.91	39.16	45.37	31.28	38.47
A+B-2	44.46	36.45	41.54	50.88	33.54	41.37
A+B	53.18	42.38	48.69	54.23	36.78	47.05
A+B+C	58.72	47.49	51.72	57.95	42.54	51.68
A+B+C+D	63.75	50.31	55.38	60.94	45.95	55.27

IV. B. Application of emotion recognition models

IV. B. 1) Emotion Recognition Confusion Matrix

For the emotion feature recognition of vocal singing speech, this paper adopts semantic annotation software for emotion annotation of vocal singing speech on the basis of the collected vocal singing speech, and annotates 2000 vocal singing speech into five kinds of emotion features, namely, happiness, sadness, anger, fear, and neutrality, respectively. For the effectiveness of this paper's model for emotion recognition classification of vocal singing speech, this paper is further verified by the confusion matrix. Figure 4 shows the results of the confusion matrix for emotion recognition.

As can be seen from the figure, the numbers in the confusion matrix represent the number of different pronunciation features identified, the darker the color represents the number of pronunciation features, the number on the diagonal represents the number of pronunciation features predicted accurately by the model, and the numbers in the other cells represent the number of incorrectly predicted by the model, and the number in the upper left corner is 1115, which has the darkest color, because the number of happy emotion features expressed by vocal singing in opera performances has the largest proportion. Overall, the number on the diagonal of the confusion matrix is higher, and the CLDNN-SAGAN model established in this paper has a higher accuracy in the prediction of vocal singing speech emotion features.

In addition, for vocal singing on the opera stage, there are differences in vocal singing techniques for different tunes, so this paper recognizes three different types of vocal tones, namely, high, middle and low tones, to further illustrate the effectiveness of this paper's model in distinguishing between different vocal tones. Figure 5 shows the confusion matrix for recognizing different tones. As can be seen from the figure, the darkest colored number in the upper left corner is 1206, indicating that in opera performance, the number of high vocal tones in vocal singing is higher, and the values on the diagonal line are all the maximum values in the row or column, which indicates that the CLDNN-SAGAN model designed in this paper has a high accuracy in vocal tone distinction and recognition.

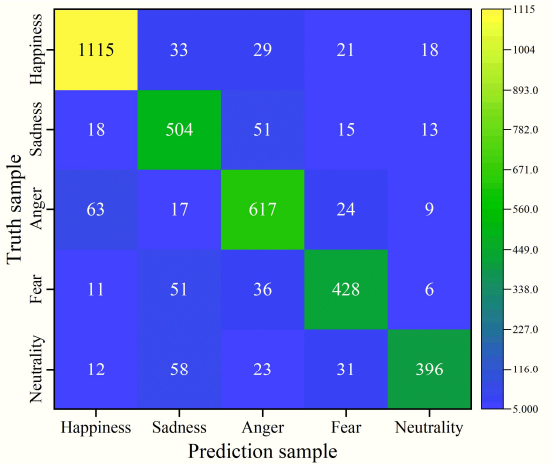

Figure 4: The result of the confusion matrix of the emotional recognition

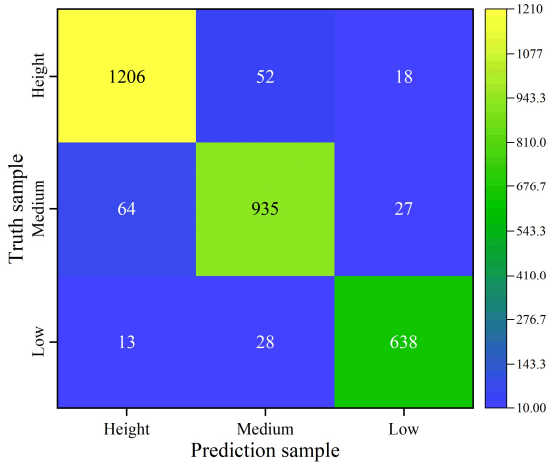

Figure 5: Confusion matrix of different tones

Combining the results of the above analysis, it can be seen that the CLDNN-SAGAN model designed in this paper can effectively recognize the emotional performance and vocal expression during vocal singing in opera performance, which is conducive to better assisting the opera performers to distinguish the expression of different opera works, so as to further enhance the artistic expression of opera works.

IV. B. 1) Evaluation of vocal singing skills

Evaluating an opera production is a subjective process because everyone has different feelings and preferences about vocal singing. However, the quality of vocal singing can be evaluated by some criteria that can help people better understand and appreciate opera. In the CLDNN-SAGAN model proposed in this paper, after categorizing the vocal singing features, a vocal scoring module is also designed, which aims at assisting vocal performers to better master the emotional expression of opera works and further adjust the vocal performance skills, so as to enhance the artistic expression of opera works.

In order to verify the feasibility of the scoring module, 15 professional listeners (with or engaged in opera performance experience) and 15 ordinary listeners (without any exposure to opera works) were selected in this paper. Let the two different types of listeners score different opera works, the scoring weight of the professional listeners is three times that of the ordinary listeners, and the average score according to this program is the artificial scoring score of an opera work. From the vocal singing voice dataset constructed in this paper, 12 vocal performance voices are randomly extracted, and professional listeners, ordinary listeners and CLDNN-SAGAN model are allowed to score them respectively, and the comparison results of the differences between the model scores and manual scores are obtained as shown in Table 3.

As can be seen from the data distribution in the table, there is also variability in the scoring between different listeners. Professional listeners give different scores to different music with greater variability, and are more inclined to positive emotional opera works, while ordinary listeners are more enthusiastic about bright melodies and chords, and are not inclined to give scores with a large gap. At the same time, it can be seen that although there is a certain difference between the AI scoring and the audience scoring, the overall difference is not big, and the AI scoring is closer to the scoring of the professional audience, which also indicates that the scoring of the ordinary audience is more subjective, and is easily affected by the factors of personal feelings, or different preferences and other factors. This shows that AI scoring and manual scoring can complement each other to provide more comprehensive and accurate results for vocal singing skill scoring.

Table 3: Model scores and artificial scores

Segment	Style	Professional (*0.75)	Ordinary (*0.25)	Final score	Model
MIDI1	Sadness	90.47	90.31	90.43	91.14
MIDI2	Sadness	90.28	91.65	90.62	90.78
MIDI3	Happiness	94.06	90.21	93.10	92.95
MIDI4	Happiness	94.19	92.38	93.74	93.07
MIDI5	Happiness	94.35	92.74	93.95	93.48
MIDI6	Happiness	93.92	93.12	93.72	93.69
MIDI7	Neutrality	91.35	91.05	91.28	91.43
MIDI8	Neutrality	91.48	90.42	91.22	91.29
MIDI9	Anger	88.27	86.37	87.80	90.35
MIDI10	Fear	88.94	86.51	88.33	87.42
MIDI11	Happiness	94.27	92.18	93.75	94.26
MIDI12	Happiness	94.35	94.06	94.28	94.05

V. Conclusion

In order to better improve the vocal singing skills of performers on the opera stage, this paper proposes a vocal singing emotion feature recognition model based on the CLDNN-SAGAN model, which can assist vocal performers to optimize their vocal singing skills in opera by matching the emotional features of vocal singers and scoring them. Experiments show that the model in this paper achieves an emotion recognition accuracy of about 88.75% under different sample lengths, and the model scores up to 94.26 points, which is smaller than the professional manual scoring. It can be an effective tool for vocal performers to optimize their vocal singing skills, and it also provides a reliable guiding direction for further expanding the artistic expression of the works on the opera stage, further enhancing the attractiveness of the opera works to the audience.

References

[1] HEBEISEN-MOŞUC, E. (2024). The Vocal Technique as an Instrument of Expression on Stage. Învăţământ, Cercetare, Creaţie, 10(1), 149-170.
[2] Tăbăcaru, R. (2020). Drama-Music communication in Opera performance. Bulletin of the Transilvania University of Braşov, Series VIII: Performing Arts, 13(2-Suppl), 213-320.
[3] Kostyuk, A. A., & Alekseeva, G. V. (2023). Emotions as a Phenomenon of Vocal and Opera Music. Russian Musicology, (1), 168-177.
[4] Guanjing, Y., Shijie, H., Zhihuan, Z., Yaqi, Z., & Yuchen, L. (2024). The semantic nature of opera vocal performance interpretation: historical genesis and musical stylistic means. Cadernos de Educação Tecnologia e Sociedade, 17(se4), 268-277.
[5] Bruder, C., & Larrouy-Maestri, P. (2023). Classical singers are also proficient in non-classical singing. Frontiers in Psychology, 14, 1215370.
[6] Brigo, F., Porro, A., & Lorusso, L. (2023). The Singing Brain: Words and Music in the Opera. In Effects of Opera Music from Brain to Body: A Matter of Wellbeing (pp. 67-77). Cham: Springer International Publishing.
[7] Müller, M., Schulz, T., Ermakova, T., & Caffier, P. P. (2021). Lyric or dramatic-vibrato analysis for voice type classification in professional opera singers. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 943-955.
[8] Condon, S. (2018). Preparing an emotionally expressive vocal performance. European Journal of Philosophy in Arts Education, 3(1).
[9] Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., & Eyben, F. (2017). The expression of emotion in the singing voice: Acoustic patterns in vocal performance. The Journal of the Acoustical Society of America, 142(4), 1805-1815.
[10] Yang, Z. (2022). Scientific and technological creative stage design using artificial intelligence. Computers and Electrical Engineering, 103, 108395.
[11] Lin, S. C., Chou, C. H., Ke, M. F., Liao, S. H., Lin, Y. H., & Kuo, C. P. (2022, November). Appling Artificial Intelligence Techniques on Singing Teaching of Taiwanese Opera. In Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology (pp. 1-2).

[12] Bareggi, A., Bardazzi, F., & Amour, L. (2023, September). New Perspectives in Virtual Environments for Opera Music. In 2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA) (pp. 1-7). IEEE.
[13] Lyu, B. (2020). Research on the Possibility of the Application of Artificial Intelligence to Chinese Opera Stage Performance. Solid State Technology, 1698-1703.
[14] Fraisse, V., Wanderley, M. M., & Guastavino, C. (2021). Comprehensive framework for describing interactive sound installations: Highlighting trends through a systematic review. Multimodal Technologies and Interaction, 5(4), 19.
[15] Kim, H. J., & Lee, S. S. (2021). A Study on the implementation of immersive sound using multiple speaker systems according to the location of sound sources in live performance. International Journal of Asia Digital Art and Design, 25(1), 14-21.
[16] Vilkaitis, A., & Wiggins, B. (2019). Ambisonic Sound Design for Theatre with Virtual Reality Demonstration-A Case Study. EPiC Series in Technology, 1, 60-67.
[17] Antoshchuk, S., Kovalenko, M., & Sieck, J. (2018). Creating an interactive musical experience for a concert hall. International Journal of Computing, 17(3), 143-152.
[18] Zhongqiao Sun. (2024). Emotional Expression and Singing Techniques in Vocal Singing Research. Journal of Global Humanities and Social Sciences,5(12).
[19] Fan Zhang. (2024). Analysis of the Art and Emotional Skills of College Vocal Singing in the Age of Big Data. Applied Mathematics and Nonlinear Sciences,9(1).
[20] Pavani Cherukuru & Mumtaz Begum Mustafa. (2024). CNN-based noise reduction for multi-channel speech enhancement system with discrete wavelet transform (DWT) preprocessing. PeerJ. Computer science,10,e1901-e1901.
[21] Rohitesh Kumar & Rajib Ghosh. (2025). Person verification and recognition by combining voice signal and online handwritten signature using hyperbolic function based transformer neural network. Neurocomputing,632,129751-129751.
[22] Ali Aldhubri,Jianfeng Lu & Guanyiman Fu. (2024). SAGAN: Skip attention generative adversarial networks for few-shot image generation. Digital Signal Processing,149,104466-.