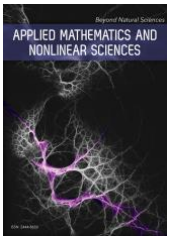
sciendo Applied Mathematics and Nonlinear Sciences, 9(1) (2024) 1-13



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Integration Strategies of American Voice Singing in Singing and Vocal Teaching Based on Multiscale Feature Fusion

Yun Liu[1,†]

1. Conservatory of Music, Sanya University, Sanya, Hainan, 572000, China.

Submission Info

Communicated by Z. Sabir
Received December 20, 2023
Accepted December 28, 2023
Available online January 31, 2024

Abstract

Incorporating American voice singing in singing and vocal music teaching has always been a problem with no standards and excessive teacher subjectivity. In this paper, based on a convolutional neural network, the American voice features are extracted using Mel spectrum, and the CBAM attention module is introduced to refine the American voice features and improve the influence of background noise on American voice extraction. The emotional features extracted under different scale frame lengths are feature fused to build a neural network model of multi-scale feature fusion to quantitatively evaluate the students' American voice singing. After obtaining the consent of the leadership of a university, it was put into use in its music department to score the students by analyzing the status of their pronunciation of American vowels, consonants, and legato, and output the problems that the students had in a specific bar. The results showed that the more frames in the input vector, the more correct the model was, 89.32 ± 0.21 for 20 frames, increasing to 90.05 ± 1.32 for 40 frames. Student 1 scored 0.654 for American vowel pitch and 0.643 for consonants. Student 10 had the highest vowel pitch with a score of 0.718. Vowel scores were generally around 0.6. Students had the highest wavelength of 0.61, which does not correspond well with the original score. This study shows the direction for the strategy of integrating American singing into the vocal program and promoting the healthy development of vocal teaching.

Keywords: Convolutional neural network; Multi-scale feature fusion; Meier spectral features; CBAM; American vocal singing; Vocal music teaching.
AMS 2010 codes: 97M80

†Corresponding author.
Email address: Lww01220507@163.com

ISSN 2444-8656
https://doi.org/10.2478/amns-2024-0237

1 Introduction

The creation of the American vocal singing method not only has a close connection with the development process of European music but also as a part of human cultural ideology, it is a product of the development of the society and the times [1-3]. The American vocal singing method has not only had a profound influence on Chinese vocal music teaching, but it has also had an important influence on the development of world music and art. The integration of the American vocal singing method into Chinese vocal music teaching helps to promote the development of Chinese vocal music art, improve the quality of Chinese vocal music teaching, and is of great significance in promoting students' mastery of comprehensive and systematic vocal music knowledge and performance skills [4-6].

The incorporation of American vocal singing into singing and vocal music teaching has provided individuals with rich musical experiences and teaching revelations. Integrating the essence of American vocal singing into vocal teaching can improve students' singing skills, emotional expression, and stage performance, and lay a solid foundation for their music learning [7-8]. In the diversified music world today, it is important to emphasize, inherit, and carry forward the unique art form of American vocal singing. The integration of American vocal singing has injected new vitality into singing and vocal teaching, which helps to cultivate more singing talents who love music and have a high degree of artistic literacy [9-11].

Characterized by its beautiful sound quality, clear articulation, and perfect control, American Voice singing provides singers with an expressive art form. In today's teaching of singing and vocal music, the inclusion of American voice singing is of great significance. Through training in breathing control, vocal technique, biting position, and range extension, students can master these skills in practice and improve their singing level [12]. In literature [13], it is discussed that vocal teaching should not only focus on professional exercises like vocalization and voice protection but also emotional input. Literature [14] explains the connection and influence between neuroscience and vocal music teaching and describes the characteristics of neuroscience-based imaginative thinking in singing art activities, such as image display and scene integration, which provides a reference for the study of neuroscience-based vocal music church. Literature [15] proposes a method based on motor learning principles to examine the teaching behavior of classical singing teachers. The results of the study can enlighten teachers to innovate and improve their teaching management in classical singing. Literature [16] describes the gradual trend of real-time magnetic resonance imaging (rtMRI) video as a teaching tool, which is a new learning resource for students of singing and spoken performance. Literature [17] aimed to detect subjective and objective voice quality and the effect of singing profession or singing studies on voice quality in women aged 60-75 years, and found mild impairment of objective and perceptual voices and no difference in speech impairments in this age group, as well as a significant lack of coarseness in the voices of women in professional singing-related fields compared to non-professional ones. Literature [18] reviewed the framework for teaching classical singing training and pointed out that a scientific framework for classical singing training has not yet been developed, and suggested that a high-quality methodology should be used to conduct a study on the design of the framework. Literature [19] designed the system architecture of personalized teaching methodology based on the system structure of generic ITS, integrating the teaching characteristics of music discipline and the model of music sight-singing intelligent guidance system. This method's ability to improve students' independent learning ability is confirmed by the teaching case analysis of the experimental class. Literature [20] designed and constructed a mobile multichannel surround sound teaching system based on practical teaching experience. The system's feasibility is validated through theoretical and practical case studies, which can assist students in correctly comprehending audio and mastering digital technology to disseminate art.

In this paper, we combine the Meier spectral feature method to extract the American voice features of students in vocal teaching and use a convolutional neural network to analyze the voice to create a quantitative evaluation model of American vocal singing in vocal teaching. To explore more specific, rich, and effective emotional features in American vocal singing, the emotional features extracted under different scale frame lengths are feature fused under the premise of ensuring that each frame signal meets the smooth characteristics. The CBAM attention mechanism is introduced to refine sound features, denoise irrelevant background sounds, and construct the CRNN network with multi-scale feature fusion and attention mechanism. After the model design is completed, the human voice separation method is used to separate the accompaniment sound and extract the clear singing voice of the American singer, to train the model's recognition effect on the American singing style, and then the model is practically used in the music department of a university, which can be used to recognize and score the students' pronunciation of the American singing style with the teaching of the teachers.

2 Convolutional neural network application strategy in vocal music teaching

Sound detection is a technique that utilizes the features of a sound signal to predict its sound event category, and it is an important research direction in the field of computer hearing. The primary method of sound event detection involves detecting a sound signal by extracting features and using neural networks and other methods to predict the corresponding sound event category. The goal of sound event detection in vocal music courses is to identify the type of sound event by analyzing the characteristics of the sound signal. The framework of the sound event detection algorithm based on neural network technology is shown in Figure 1. The algorithm generally includes audio input, sound feature extraction, neural networks, and system output.

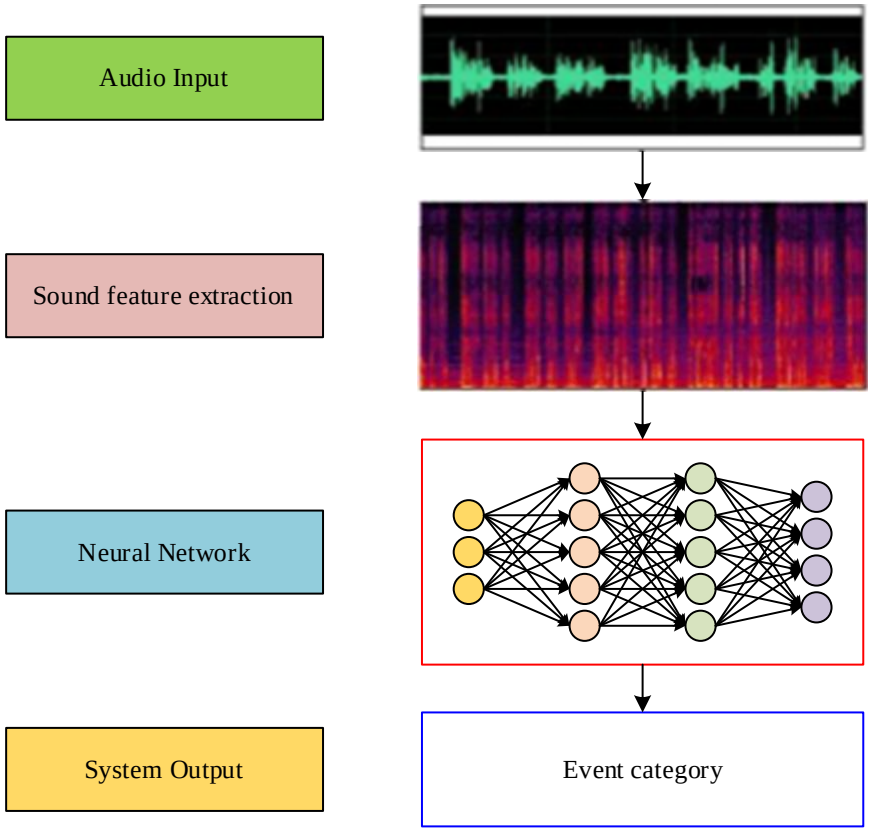


Figure 1. Sound detection principle process

2.1 Sound Feature Extraction

From the study of the auditory characteristics of the human ear, it is found that the use of Mel frequency is more in line with the actual human ear hearing, i.e., presenting a linear relationship under $1000Hz$ and a logarithmic relationship above $1000Hz$. For the sound of the frequency is relatively close, the human ear generally can not distinguish, this is the concept of critical bandwidth, when the frequency difference between the two sounds to a certain range, the human ear can be distinguished. Formula (1) for the Mel frequency and Hz frequency (i.e., the actual frequency) of the relationship between the expression:

$$f_{mel} = 2595 * \lg\left(1 + \frac{f}{700}\right) \tag{1}$$

Pre-emphasis. Currently, sound generation systems tend to suppress high frequencies when sound signals are emitted, as they need to be compensated for using pre-emphasis. In general, this is multiplied by a factor that is larger at higher frequencies and smaller at lower frequencies, allowing for the portion at higher frequencies to be compensated for. Pre-emphasis is the filtering to get the higher frequency part, i.e., the high-pass filter, and formula (2) is the formula for pre-emphasis:

$$H(z) = 1 - \alpha z^{-1} \tag{2}$$

Split framing is similar to STFT, where a long signal is turned into a short signal using split frames, and then the audio signal is processed using, for example, the Fast Fourier Transform. The length of each frame is typically 20-40ms, with some overlap between two adjacent frames.

Adding windows. To smooth the signal and prevent spectral leakage, it is necessary to add a window, which is usually a Hamming window. Equation (3) shows the process of calculating it:

$$w(n, \alpha) = (1 - \alpha) - \alpha \cos\left(2\pi \frac{n}{N - 1}\right) \tag{3}$$

In the expression, N represents the length of the Hamming window used, $0 \le n \le N - 1$, α are the parameters of the Hamming window. Generally, α is taken as 0.46.

Fast Fourier Transform is an effective method to obtain the spectral information of the sound signal, which can well transform the sound signal from the time domain to the frequency domain. Equation (4) shows its specific calculation process:

$$X_i(k) = \sum_{n=0}^{N-1} x_i(n) e^{-j\frac{2\pi nk}{N}}, 0 \le k \le N - 1 \tag{4}$$

$x_i(n)$ the n rd time point of the i nd frame of a signal in the time domain, $X_i(k)$ the k th frequency point of the i th frame of a signal in the frequency domain.

Find the energy spectrum. Based on the above results, calculate the energy for each frequency point in each frame of the signal. In general, the energy spectrum of the signal for each frame is obtained by squaring the output of the Fourier transform. Equation (5) shows its calculation:

$$S_i(k) = \left|X_i(k)\right|^2, 0 \le k \le N - 1 \tag{5}$$

where S_i represents the energy spectrum of the i nd frame signal.

2.2 CBAM Attention Module

For the existence of some irrelevant background noise and other influences in the vocal image to be recognized, CBAM is chosen here for processing, which belongs to a hybrid attention mechanism that helps to refine the features. The overall arithmetic formula is as follows:

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s\left(F'\right) \otimes F' \end{aligned} \tag{6}$$

where $\otimes$ denotes the element level multiplication, $M_c(F)$ denotes the channel attention feature map, and $M_s(F)$ denotes the spatial attention feature map.

In the channel attention module, two channel descriptions of $1\times1\times C$ are first obtained using both maximum pooling and average pooling to achieve compression and aggregation of the feature space. Then, they are fed into a two-layer network of shared multilayer perceptron machines, respectively. The two obtained feature maps are then summed element by element and activated by a sigmoid function to obtain the channel attention feature map M_c. The process is computed as in equation (7):

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)) \tag{7}$$

Where, σ is the sigmoid activation function. $AvgPool(F)$ and $MaxPool(F)$ denote the average pooling and maximum pooling operations on the features, respectively.

In the spatial attention module, the feature maps after average and maximum pooling are spliced into the channel dimension, and then a convolution operation is performed on them. The computational procedure of the spatial attention module is shown in (8):

$$M_s(F) = \sigma\left(f^{k\times k}([AvgPool(F); MaxPool(F)])\right) \tag{8}$$

where $f^{h\times k}(\cdot)$ is the convolution kernel of $k\times k$ performing the convolution operation, and $[AvgPool(F); MaxPool(F)]$ denotes the on-channel splicing of the average pooled feature map and the maximum pooled feature map B.

2.3 Application of multi-scale feature fusion approach in teaching American voice

To mine more specific rich and effective emotional features in MFA, the emotional features extracted under different scale frame lengths are feature fused under the premise of ensuring that each frame signal meets the smoothness characteristics in the following way. In this paper, we propose a multi-scale feature fusion and attention mechanism CRNN network, the architecture of which is shown in Fig. 2. MFA-CRNN includes input, output, multi-scale attention module, BGRU, full connectivity layer, and multi-scale feature fusion branch, where the MF includes feature down-sampling, up-sampling, and residual hopping operations.

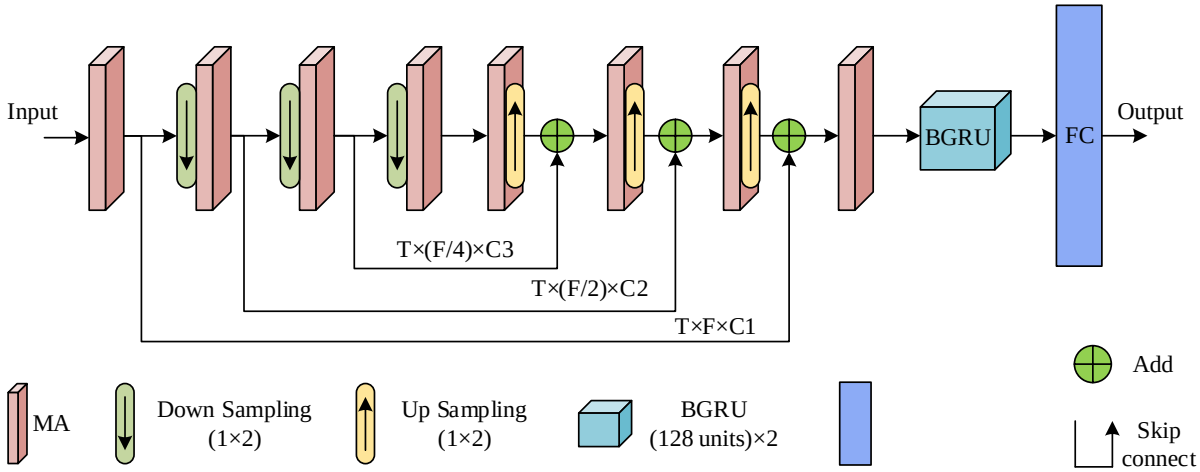


Figure 2. MFA-CRNN network structure

The direct input of MFA-CRNN is the Logmels 2D time-frequency features of the original audio, and the Logmels are first fed into the stacked MA module for deep feature extraction, highlighting the key information in the local time-frequency features and global channel features to improve the feature extraction capability of the model. Meanwhile, the MF -branch fuses the shallow and deep convolutional features to obtain richer contextual information and improve the feature expression capability of the model. Then the obtained multi-scale features are input into BGRU to capture the temporal correlation between frames and model the temporal dependence with the number of hidden layers 2 and 128 units in each hidden layer. The last layer FC is used as a classification layer, after a sigmoid activation function, to predict the sound event category corresponding to each frame, with 14 hidden units, corresponding to the number of sound event categories. The output of the model is a two-dimensional matrix of dimension $T \times K$, where T corresponds to the time domain dimension and is the number of frames of continuous audio, and $K = 14$ is the number of sound event categories in the dataset of this paper.

In the MA feature detection task, most use an attention mechanism on the time axis, focusing on the part of the frame where the event occurs and ignoring the influence of irrelevant frames. The influence of different frequency features of different events cannot be demonstrated by this attention approach, which is solely based on the time axis. The MA structure is shown in Fig. 3. It mainly consists of a time-frequency unit attention module and weighted enhancement of channel features, while local and global attention mechanisms are introduced to enhance the feature extraction capability of the model.

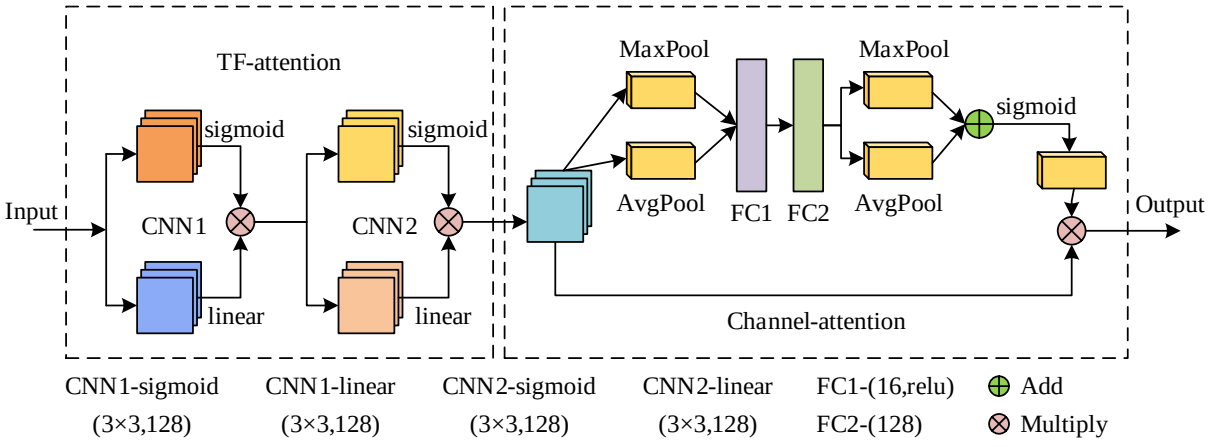


Figure 3. MA structure

In the TF-attention part, a set of parallel convolutional layers with a convolutional kernel size 3×3 and a channel number of 128 are first used for feature learning. To highlight the important information about time-frequency units and reduce the interference of useless information, this paper uses the sigmoid nonlinear activation function (denoted as σ) and the linear activation function to control the flow of time-frequency information. The function value of the sigmoid is between 0 and 1, which indicates the degree of retention of the time-frequency information, and if its value is close to 1, then attention is paid to the corresponding time-frequency unit and is ignored if it is close to 0. There are two sets of stacked convolutional CNN1 and CNN1 with the size of 128 channels in the TF-attention part. Attention has a total of stacked two sets of convolutional CNN1 and CNN2, which ultimately input the locally strengthened time-frequency features into the Channel-attention module, and further construct the interdependence between the channels for the global information, highlighting the features with stronger correlation between the channels.

In Channel-attention, in order to aggregate the input time-frequency features by channel dimension, global average pooling and global maximum pooling are performed first to obtain the global compression vector. Then the interdependence between the channels is modeled, using two layers FC to highlight the dependencies between the channels, the first layer $FC1$ for channel dimension compression, with hidden units of 16 and relu as the activation function. The second layer $FC2$ performs channel dimensionality reduction with 128 hidden units. The normalized weight matrix is obtained by sigmoid activation function, and finally the output features of TF-attention are multiplied element by element with the channel weight matrix obtained by Channel-attention. On the basis of local time-frequency feature reinforcement, the channel features are further reinforced to obtain multi-scale attention features.

Denoting the input feature map as $X \in R^{T\times F\times C}$, the convolutional layer output can be expressed as:

$$G = W^{*}X + b \tag{9}$$

where $*$ denotes the convolution operation, W and b denote the convolution kernel parameters, weight matrix and bias, respectively. Output feature $G \in R^{T\times F\times C'}$, where $T\times F$ denotes the time-frequency unit of the feature map and C' is the number of channels.

The TF-attention structure introduces local attention into the convolutional layers, and the two-layer convolutional outputs G_1 and G_2 of $CNN1$ are linearly weighted:

$$G = G_2 \otimes \sigma(G_1) \tag{10}$$

Where σ is the sigmoid activation function, $\otimes$ denotes the element-by-element multiplication, and the output features $G \in R^{T\times F\times C}$ and $T\times F$ denote the time-frequency units of the features.

There are two sets of stacked convolutions CNN1 and CNN2 in TF-attention, which finally get the locally strengthened time-frequency features $Y \in R^{T\times F\times C}$. Y is inputted into the global attention Channel-attention module, which further constructs the interdependence between the channels on the global information, and highlights the features that have stronger correlation between the channels.

In Channel-attention, global average pooling and global maximum pooling are first performed on the channel information, and the channel compression vectors d_{avg} and $d_{\max}$ of size $1\times1\times\mathrm{C}$ are obtained:

$$d_{avg} = \frac{1}{T \times F} \sum_{i=1}^{T} \sum_{j=1}^{F} Y_{ij} \tag{11}$$

$$d_{\max} = Max\left(Y_{ij}\right), (i = 1, \cdots, T; \ j = 1, \cdots, F) \tag{12}$$

Second, two fully connected layers with 16 unit nodes model the interdependence between channels to obtain the channel weight matrix $S \in R^{1\times1\times C}$:

$$S = \sigma\left(w_2\left(w_1 d_{avg}\right) + w_2\left(w_1 d_{\max}\right)\right) \tag{13}$$

where σ is the sigmoid activation function, $w_1 \in R^{Clr\times C}$ and $w_2 \in R^{C\times Clr}$ are the parameter matrices of the two layers FC and r denotes the channel compression factor.

Eventually, the output feature map Y of TF-attention is multiplied element-by-element with the channel weight matrix S to obtain the output $M \in R^{T\times F\times C}$ of the MA module, as shown in Eq. (14):

$$M = s \cdot Y \tag{14}$$

3 The integration strategy of American vocal singing in singing and vocal music teaching

3.1 Establishment of a corpus and articulatory characterization system for American voice singing

Since there is no publicly available dataset of American vocal singers' clean singing, the vocal separation method is used to separate the accompaniment sound, extract the singer's clean singing voice, and remove the part that does not contain the human voice to form a clean singing voice clip that only contains the human voice. In this experiment, a total of 20 American singers and 10 non-American singers were collected, with 400 sound clips for each singer, the length of each clip ranging from 1s to 8s, and each clip labeled with the singer's name. All the sound clips were stored in WAV format with a sampling rate of 16KHZ, 16-bit, and a single channel. To verify the effect of the number of singers on the effectiveness of the spatial representation of timbre embedding, four different subsets of singers were set up for training, respectively S4, S8, S15, and S30. Where S8 includes S4, and so on.

In the experiment, the CQT features of each sound clip are first extracted and stored in any format. Due to the K-fold cross-validation method, all the CQT feature files are randomly divided into five parts, and the ratio of training and testing is 3:1. Then the training data are fed into the convolutional network in the form of pairwise training, and the stochastic gradient descent method is utilized for the iterative training, with the learning rate set to 0.01, and the maximum number of iterations is set to 40,000. After the iteration is completed, the training, validation, and testing data are embedded into the respective model according to the trained model. Data, validation data, and test data are embedded into their respective timbre embedding spaces, and the 3-dimensional timbre features corresponding to each input vector in the timbre embedding space are obtained. Finally, based on the 3-dimensional timbre features of the training data and the test data, a classifier for American singing recognition is constructed to verify the effectiveness of the characterization of the 3-dimensional timbre features in the timbre embedding space.

When constructing the classification model for singer recognition, the training data, validation data, and test data come from the 3-dimensional timbre features of 80,000, 30,000, and 30,000 segments in the timbre embedding space, respectively, where the training data are used to construct the classifier, the validation data are used to find the optimal parameters, and the test data are used to evaluate the classification accuracy of the 3-dimensional timbre features in the timbre embedding space. The higher the classification accuracy, the more effective the 3-dimensional timbre feature representation obtained from the timbre embedding space of the human voice.

The model output of this paper is shown in Table 1. The more frames of the input vector, the higher the correct rate of the model, which is 89.32±0.21 at 20 frames and increases to 90.05±1.32 at 40 frames, which means that the feature abstraction and dimensionality reduction effectiveness of the timbre embedding space increases with the increase of the number of frames of the input audio signals, which means that the more the frames of the input audio signals are, the more the deep learning model of vocal timbre embedding can learn from it the timbre-related essential feature representations, and the better the ability of timbre representation. The more singers in the embedding space, the lower the correct rate of this paper's model, which means that the effectiveness of feature abstraction and dimensionality reduction in the timbre embedding space decreases with the increase of the categories in the embedding space. The number of frames in S4 is 89.32±0.21, and then it decreases to 35.38±0.58 in S30. There is a certain amount of timbral similarity between singers themselves, and the timbral similarity among different singers is set to 0 in the current experiments. As the number of singers increases, the effect of timbre similarity between singers is greater, and the number of cases in which singers with similar timbre are misclassified in the model increases.

Table 1. Identification of training data

Training data situation				
Frame number	S4	S8	S15	S30
20	98.23±0.03	98.54±0.05	92.23±1.32	38.53±0.62
40	98.95±0.03	98.92±0.05	93.57±0.54	65.23±1.32
60	100+0.01	99.12±0.05	99.57±0.02	83.53±1.19
Verify data				
Frame number	S4	S8	S15	S30
20	89.32±0.21	79.28±1.68	58.28±1.59	35.38±0.58
40	90.05±1.32	81.11±2.16	71.85±2.58	41.68±1.56
60	91.17±0.16	81.58±1.59	72.69±1.29	47.58±0.19

The confusion matrix is generated and analyzed with the results of the test data for the subset of singers S8. Figure 4 shows the confusion matrix for S8. The results show that the last four singers have high error rates with values ranging from 0.5 to 1. The four singers were all American female singers. The voices of male singers are seldom misclassified as those of female singers, and the voices of female singers are also less likely to be misclassified as those of male singers. To a certain extent, it can be shown that the similarity between singers and singers affects the correct classification rate of the model.

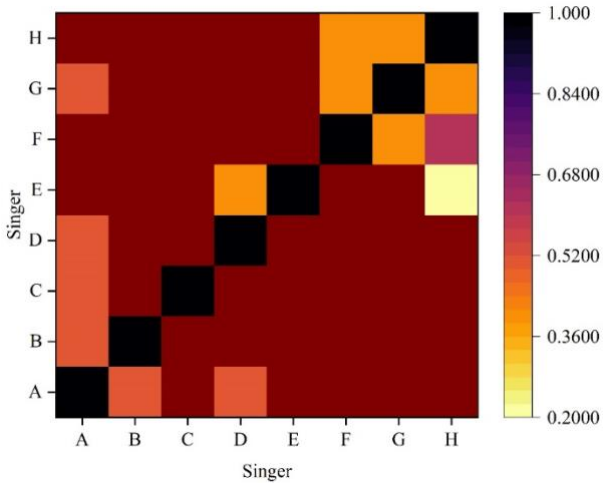


Figure 4. S8 Confusion matrix

3.2 Applied Research in Singing and Vocal Music Teaching in Colleges and Universities

Thirty undergraduate college students majoring in vocal music from a music department were selected, including 15 girls and 15 boys, aged 18 to 23, all of whom had been engaged in vocal studies for more than 2 years and were free of laryngeal diseases and upper respiratory tract infections. Recording the vocal signals of the research subject in an ethnic singing style, using samples with a sampling frequency of 48KHz and 16B quantization, with ambient noise below 45dB.

Clear pronunciation is emphasized in American vocal singing and students must master a range of vowels and consonants, including pronunciation skills for vowels, consonants, legato, and pauses. Accurate pronunciation is an important way to convey the emotion of a song, so teachers need to emphasize the importance of students' pronunciation training in their teaching. Pitch is an important indicator of a singer's skill level. The model was applied to singing and vocal music teaching at a college after verifying its ability to recognize American singing. Table 2 illustrates the performance of certain students in a music class taught at school. The output of the neural network shows that Student 1 has an American vowel pitch score of 0.654 and a consonant score of 0.643. Student 10 had the highest vowel pitch with a score of 0.718. The students at this school generally have a lower American pronunciation pitch of around 0.6. Intensive practice is needed.

Table 2. Accuracy of student pronunciation

Number	Vowel	Consonant	Quintuplet	Pause technique
1	0.654	0.643	0.678	0.643
2	0.642	0.658	0.636	0.633
3	0.678	0.677	0.643	0.643
4	0.698	0.691	0.671	0.623
5	0.617	0.622	0.633	0.645
6	0.655	0.699	0.645	0.711
7	0.712	0.714	0.736	0.743
8	0.764	0.755	0.752	0.714
9	0.655	0.658	0.651	0.643
10	0.718	0.722	0.737	0.691

The model in this paper incorporates multi-scale features to output the student's score. Taking student 10 as an example, Table 3 shows the specific performance of the student in a particular music class. The part of the model outputs 0 means no errors, and 1 is different from the original score. It can be seen that the student made a mistake on the 2nd note of the second measure, which is a consonant, and Student 10's pitch is 5 degrees above the original score. The 3rd note of the same measure is 3 degrees lower.

Table 3. Student 10 specific problems

	Subsection	Note	Vowel	Consonant	Quintuplet	Pause technique	Problem
Student 10	2	2	0	1	0	0	Tone+5
	2	3	1	0	0	0	Tone-3
	3	3	0	0	1	0	Loudness+2
	3	2	1	0	0	0	Loudness-5
	4	1	1	1	1	0	Tone+2
	4	4	1	1	0	1	Tone-2
Overall evaluation	-		0.718	0.722	0.737	0.691	-

The vocal features of the students were extracted and then compared. Figure 5 shows the situation of students singing in an American voice. The vocal teacher's performance of American voice singing is exhibited by the blue curve in the figure. It can be seen that the teacher's amplitude is the highest at 0.72, while the student's wavelength is slightly shorter with a maximum of 0.61. The overlap between the two is not high, and the student's tone intensity and stopping skills need improvement.

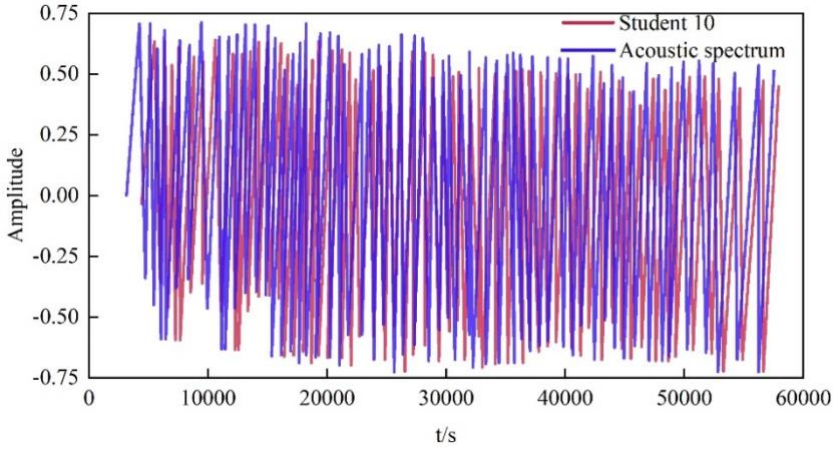


Figure 5. Student 10 waveform condition

Further analysis of the students in this school, after using this model to teach American vocal integration strategies, the students' intonation in the second semester is shown in Figure 6. The wall chart's lower layer displays the students' performance during the previous semester, while the upper layer displays the pitch after using the present model for one semester. Group 1 shows vowel intonation, where students improved by an average of 0.05 to 0.07 points, which is a relatively large improvement. Group 2 is the consonant condition, and student S9 had the highest improvement score, from 0.658 to 0.683. Group 3 shows the status of allophonic articulation, with Student S10 improving from 0.737 to 0.756. The last group shows an improvement in stopping skills. Student S10 improved from 0.691 to 0.742, the highest improvement. The model in this paper has certain effects on assisting American voice singing, which provides a reference for teachers to teach vocal music.

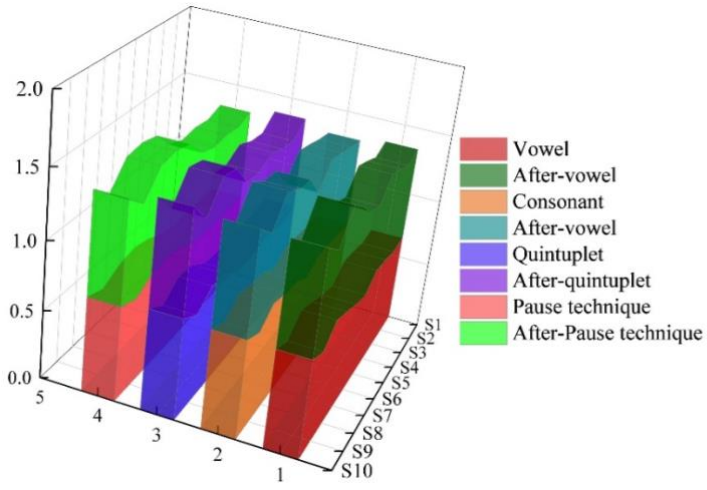


Figure 6. The student sound status of the second semester

3.3 Recommendations

The integration of American vocal singing into singing and vocal teaching brings a rich musical experience and a teaching revelation. By integrating the essence of American vocal singing into vocal teaching, students' singing skills, emotional expression, and stage performance can be improved, which can lay a solid foundation for their music learning. In today's diversified music world, the unique art form of American vocal singing should be emphasized, inherited, and carried forward. The integration of American vocal singing has injected new vitality into singing and vocal teaching, which helps to cultivate more singing talents who love music and have a high degree of artistic literacy. New quantitative and evaluation methods for integrating American vocal singing into vocal teaching have been provided by the development of intelligent algorithms and artificial intelligence. In the future, the convolutional network should be trained more systematically and comprehensively to provide more accurate and convenient teaching aids, so that students can easily master the skills of American vocal singing and improve their singing level and artistic cultivation.

4 Conclusion

In this paper, we use the CRNN network with multi-scale feature fusion and attention mechanism to recognize the students' American voice singing in the voice class, and quantitatively evaluate the students' pitch of vowels, consonants, and legato, respectively, and the following conclusions are obtained. The more frames of input vectors, the higher the correct rate of the model, which is 89.32±0.21 at 20 frames and increases to 90.05±1.32 at 40 frames. That is to say, the effectiveness of feature abstraction and dimensionality reduction of the timbre embedding space increases with the increase of the number of frames of the input audio signal. The model possesses the ability to recognize the students' intonation effectively. Student 1's American vowel pitch was 0.654 points and 0.643 points for consonants. Student 10 had the highest vowel pitch with a score of .718. The student had a problem with the 2nd note of the second measure, which is a consonant, and Student 10's pitch was 5 degrees above the original score, and the 3rd note of the same measure was 3 degrees lower. After a semester of teaching in conjunction with the model the student's pronunciation was more standardized and scores improved by an average of 0.05 to 0.07 points, which still needs to be strengthened.

References

[1] Ornoy, E. (2022). Affective responses to european art music by israeli arabs and israeli jews: a cross-ethnic study:. Musicae Scientiae, 26(1), 46-70.

[2] Western, T. (2019). “start the forgetting machine! a review of online sound archives of european traditional music.”. Yearbook for Traditional Music, 51, 325-330.

[3] Bull, A., & Scharff, C. (2021). Classical music as genre: hierarchies of value within freelance classical musicians’ discourses. European Journal of Cultural Studies, 24(3), 136754942110060-.

[4] Peretz, I., Ayotte, J., Zatorre, R. J., Mehler, J., Ahad, P., & Penhune, V. B., et al. (2020). Effects of vocal training in a musicophile with congenital amusia. Neuron, 33(2), 0-191.

[5] Xu, Y. (2021). Systematic study on expression of vocal music and science of human body noise based on wireless sensor node. Mobile information systems.

[6] Zhao, X. (2018). Application of situational cognition theory in teaching of vocal music performance. NeuroQuantology, 16(6), 308-313.

[7] Nichols, B. E. (2021). Effect of vocal versus piano doubling on children’s singing accuracy:. Psychology of Music, 49(5), 1415-1423.

[8] Clifton, K. E. (2017). Vocal music: the operas of maurice ravel. Notes, 73.

[9] Burgstaller, G. (2020). Voices of identities: vocal music and de/con/struction of communities in the former habsburg areas. Notes, 77.

[10] Freitas, R. (2018). Singing herself: adelina patti and the performance of femininity. Journal of the American Musicological Society, 71(2), 287-369.

[11] Tan, D., Diaz, F. M., & Miksza, P. (2018). Expressing emotion through vocal performance: acoustic cues and the effects of a mindfulness induction. Psychology of Music, 48(2), 030573561880987.

[12] Blaylock, R., & Narayanan, S. S. (2017). Novel imaging tools for supporting the teaching of singing and spoken performance. Journal of the Acoustical Society of America, 142(4), 2585-2585.

[13] Nian, L., & Wang, F. (2017). On the importance of emotional cultivation in vocal music teaching. International Journal of Technology, Management.

[14] Zhang, Y. (2018). Enlightenment on vocal music classroom teaching from the perspective of neuroscience. NeuroQuantology, 16(6), 132-137.

[15] Crocco, L., Mccabe, P., & Madill, C. (2020). Principles of motor learning in classical singing teaching. Journal of Voice, 34(4), -.

[16] Reed, B., & Narayanan, S. S. (2017). Novel imaging tools for supporting the teaching of singing and spoken performance. The Journal of the Acoustical Society of America, 142(4), 2585-2585.

[17] D’Haeseleer, E., Claeys, S., Bettens, K., Leemans, L., Calster, A. S. V., & Damme, N. V., et al. (2017). The impact of a teaching or singing career on the female vocal quality at the mean age of 67 years: a pilot study. Journal of Voice, 31(4), 516.e19-516.e26.

[18] Crocco, L., Madill, C. J., & Mccabe, P. (2017). Evidence-based frameworks for teaching and learning in classical singing training: a systematic review. Journal of Voice, 31(1), 130.e7–130.e17.

[19] Zhe, T. (2021). Research on the model of music sight-singing guidance system based on artificial intelligence. Complexity, 2021.

[20] Li, P., & Li, Y. (2021). Feasibility analysis of letting multi-channel surround sound system into college singing class. International Journal of Electrical Engineering Education, 002072092098608.