Speech-Music Classification Model Based on Improved Neural Network and Beat Spectrum

Chun Huang[1], Wei HeFu[2]*
General Education and International College, Chongqing College of Electronic Engineering, Chongqing 400031, China[1]
Arts College of Sichuan University, Chengdu 401331, China[2]

***Abstract*—A speech-music classification method according to a developed neural system and beat spectrum is proposed to achieve accurate classification of speech-music through pre-emphasis, endpoint detection, framing, windowing and other steps to preprocess and collect vocal music signals. After fast Fourier transforms and triangle filter processing, the Mel frequency cepstrum coefficient (MFCC) is obtained, and a discrete cosine transform is performed to obtain the signal MFCC characteristic parameters. After calculating the similarity of feature parameters through cosine similarity, the signal similarity matrix is obtained, based on which the vocal music beat spectrum is obtained. The residual structure is optimized by adding Swish and max-out activation functions, respectively, between convolutional neural network layers to build residual convolution layers and deepen the number of convolution layers. The connected time series classification (CTC) is used as the objective loss function. It is applied to the softmax layer to build a deep optimization residual convolutional neural network for speech-music classification model. The pitch spectrum of vocal music is used as the input information of the model to realize the vocal music classification. The experiment proves that the classification accuracy of the design model is higher than 99%; when the iteration reaches 1200, the training loss approaches 0; when the signal-to-noise ratio is 180dB, the sensitivity and specificity are 99.98% and 99.96%, respectively; the accuracy of voice music classification is higher than 99%, and the running time is 0.48 seconds. It has been proven that the model has high classification accuracy, low training loss, good sensitivity and special effects, and can effectively achieve the classification of speech-music.**

Keywords*—Vocal music; classification model; beat spectrum; feature parameter extraction; cosine similarity; convolutional neural network

I. INTRODUCTION

With the rapid development of the network and computer field, multimedia information such as audio and video has gradually become the mainstream of information. Audio media is the most important media form besides visual media, accounting for over 20% of the total information [1]. Faced with the huge scale of media databases and the large amount of audio content generated by users daily, it is also difficult for people to find the information they want [2]. The traditional text-based conventional information retrieval technology has been unable to meet the users' retrieval requirements, so how to effectively retrieve this information is a critical error to be resolved. Audio classification is an effective means to achieve rapid audio information retrieval [3]. Speech and music are two of the most common signals in audio. The classification of speech and music signals not only plays an important role in audio retrieval but also plays an important role in speech recognition, beat tracking and other fields [4]. For instance, in speech signal processing, first, determine the type. If it is a speech type, the following steps can evaluate the language, gender, etc. [5]; If it is music, the following steps can evaluate the music type, beat tracking, etc. [6]. This shows that the classification of voice and music is very important, and accurate classification can set a solid base for the next retrieval and other work.

In recent years, many scholars at home and abroad have done a lot of research on audio classification issues, such as vocal music, and have made certain research achievements. For example, Sun HF et al. [7] proposed a vocal music classification algorithm based on zero crossing rate and spectrum. After endpoint detection and subsection preprocessing of voice and music signals, this algorithm combines each audio segment's zero crossing rate and spectrum amplitude characteristics for classification and recognition processing. Finally, the classification of voice and music signals is realized by calculating the probability of becoming voice or music; Zhang X L [8] and others combined the residual network and random forest methods to convert the vocal music audio data of one-dimensional time domain signal into two-dimensional data form of Mel spectrogram and pre-trained the residual network to obtain a network model with high accuracy as the characteristic extractor. The network figure is used to extract the deep features in the audio, and then the random forest algorithm is used to achieve the vocal music classification in the audio; In [9] first proposed an adaptive Mel filtering algorithm to extract the Mel spectrogram with higher discrimination and then proposed a cyclic residual structure, combined with migration and fine-tuning methods, to construct a cyclic residual network spectrum classifier, aiming at the problem that the classification accuracy of voice, music and audio signals in a small sample environment needs to be improved urgently, Combining adaptive Mel filtering algorithm and cyclic residual network spectrum classifier, an audio signal classification model mainly used in small sample environment is generated, through which vocal music in audio can be classified. Although the above three methods play a role in classifying audio, such as vocal music, they also have some defects. The first method ignores the extraction of signal features before the vocal music classification. Although the last two methods extract the signal features of vocal music, they do not consider the music's rhythmic characteristics in the

audio, which leads to inaccurate extracted features, and thus affects the classification results of the overall audio signal.

The neural network is an algorithm network composed of multiple neurons with strong self-learning features, associative storage performance and the ability to figure out optimal answers at high speed [10]. The convolutional neural network is a network obtained by improving the neural network. It is a feedforward neural network with convolution calculation and depth structure. It belongs to one of the representative algorithms of depth learning. It has the ability to represent learning and can translate and classify the input information according to the hierarchical structure [11]. Beat is an organization form representing a fixed unit's time value and the law of strength and weakness in audio. Its focus is on the relationship between the strength and weakness of each audio bar, as well as the length of time between beats. The beat in music has the characteristics of a periodic cycle. Although some research has focused on music classification, with the popularity of streaming music services, real-time music classification has become increasingly important. For this reason, according to the characteristics of neural network and beat, this paper proposes a vocal music classification model based on improved neural network and beat spectrum, that is, first obtain the beat spectrum of the audio signal of vocal music, and input it into the classification model as the intake characteristic of the audio classification figure. The classification model uses a convolutional neural network, takes the connection timing classification as the objective optimization function, and integrates the residual structure into the convolution layer; Swish and max-out activation functions will be used to improve it, and finally, an end-to-end vocal music classification model based on deep optimization residual convolutional neural network will be formed. This model has good classification performance and strong applicability.

II. Classification Model of Vocal Music

A. Feature Extraction of Vocal Music Signal Based on Beat Spectrum

Voice and music are two very important audio signals, but they have different characteristics. In the music signal, there are periodic laws with a strong beat, while most of the voice is irregular, and the beat is not obvious. Beat refers to the combination rule of downbeat and downbeat, which is composed of downbeat and downbeat in a certain sequence. In music, the beat uses a strong, weak relationship to organize the music. Therefore, the characteristics of vocal music signals are extracted based on the beat spectrum.

1) MFCC feature parameter extraction of vocal music signal: Generally, if the vocal music signal is directly analyzed, the signal resolution will be very low. However, if the voice parameters of the vocal music signal are extracted and analyzed, the recognition rate will be greatly improved [12]. Mel frequency cepstral coefficients (MFCC), to a certain extent, simulate the characteristics of speech processing by the human ear [13], have good performance in auditory perception, and also have good robustness in the case of channel noise and spectral distortion [14]. The principle of MFCC is to use a group of band-pass filters to filter the sound signal according to the frequency bands of different frequencies and the bandwidth length to obtain the output signal of each band-pass filter [15]. For example, the output signal obtained by filtering the vocal music signal through the band-pass filter is used as the basic feature of the vocal music signal, which can be used as the characteristic parameters of the vocal music signal after transformation and processing [16].

In order to facilitate the study of the human ear's perception characteristics of different frequencies of speech, Mel frequency can be used. 1Mel is 1/1000 of the tone perception degree of 1000Hz. Formula (1) is the conversion formula of frequency f and Mel frequency F:

$$F = 3322.23\lg(1+0.001)f \quad (1)$$

MFCC is proposed based on the above Mel frequency concept. The process of extracting MFCC feature parameters of vocal music signals is represented in Fig. 1.

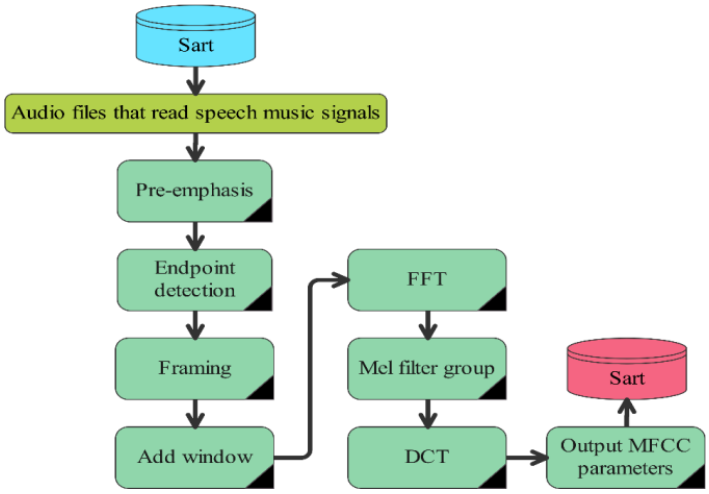


Fig. 1. Extraction flow chart of MFCC feature parameters of speech and music signals.

The specific steps for extracting MFCC feature parameters of vocal music signals are as follows:

Step 1: Read the audio file of the vocal music signal

Input the vocal music audio file into the Matlab tool, read the data and sampling frequency of the audio file using the audio read function, select the wav format for the audio file, and record the read vocal music signal as $r(n)$.

Step 2: The pre-emphasis processing of vocal music signal is aa.

Since the travelling wave propagation distance of treble is smaller than that of bass, and the energy is smaller than that of bass, it is easy to be masked by bass, so it is necessary to pre-emphasize the high-frequency signal. The pre-emphasis filter s is used to amplify the high-frequency signal to avoid the loss of high-frequency signal. The transfer function $H(s)$ of filter s is Formula (2):

$$H(s) = 1 - \ell s^{-1} \tag{2}$$

Where, $\ell \in [0.9,1]$ is generally 0.97 or 0.95.

The pre-emphasis function is Formula (3):

$$R(n) = r(n) - \ell r(n-1) \tag{3}$$

In the formula, n is the number of data points in the vocal music signal, usually 1 to 200K.

Step 3: Vocal music signal endpoint detection

Endpoint detection is an important part of vocal music recognition. Effective endpoint detection methods can not only reduce data storage and processing time but also eliminate noise interference in silent segments, making vocal music recognition more accurate. In this paper, short-time energy is used to detect the endpoint of the vocal music signal, and the start frame and end frame of the whole audio segment are determined to reduce the amount of data and calculation of subsequent audio feature extraction and improve the stability of the system.

Set the short-time energy of each frame of data in the music signal to e as the Formula (4):

$$e = \sum_{m=1}^{u_b} y^2(m) \tag{4}$$

Where, $y(m)$ is the amplitude of the m th frame of the vocal music signal, u_b is the frame length, and $m = 1,2,\cdots b$.

Step 4: Vocal music signal framing operation

If the voice of the whole vocal music signal is directly Fourier transformed, it is easy to cause a loss of timing information. However, suppose the signal is divided into several frames, and a fast Fourier transform (FFT) is performed on each frame with a fixed duration. In that case, the loss of information can be effectively avoided. Because a speech signal is a non-stationary time-varying signal, its quasi-stationary feature can only be considered as a stationary process in a short period (generally speaking, the speech signal between 10ms and 30ms is considered a stationary signal). Therefore, the voice signal can be divided into one short time segment. Each short-time segment is called a frame, and each frame contains N sampling points, usually 256 or 512 for N ; thus, the framing operation is completed.

Step 5: Windowing processing of vocal music signals

After the vocal music signal is divided into frames, the periodicity of a small signal segment will not be obvious. At this time, a Hamming window needs to be added so that the vocal music signal in a window will show periodicity. In order to avoid losing the dynamic change information of voice and music signals, there must be an overlap area between every two adjacent frames. The length of the overlap area is generally 1/2 or 1/3 of that. Multiply each frame by the Hamming window so that the continuity of each frame's left and right ends can be increased.

Set the vocal music signal of the m frame as $R(n,m), n = 0,1,\cdots,N-1$, and the vocal music signal $R'(n,m)$ after Hamming window processing is the Formula (5):

$$R'(n,m) = R(n,m)\left[0.54 - 0.46\cos(2\pi n/(N-1))\right] \tag{5}$$

Where, $0 \le n \le N-1$.

Step 6: Fast Fourier Transform (FFT) of speech-music Signal

The function of the FFT transform is to convert the time domain of the vocal music signal to the frequency domain [17], which can better reflect the characteristics of the signal and facilitate its analysis.

The frequency spectrum of the m th frame of vocal music signal after FFT transformation is Formula (6):

$$R(k,m) = \sum_{n=0}^{N'} R'(n,m)\exp(-(2\pi nk/N')) \tag{6}$$

Where, N' is the window width of FFT. After taking the modulus square of the spectrum of the vocal music signal, the power spectrum of the vocal music signal can be obtained.

Step 7: Filtering operation of vocal music signal

The obtained power spectrum of vocal music signal is filtered by a triangle filter with Mel frequency average distribution to obtain a group of coefficients $d_1, d_2, \cdots$, which are the energy output by each filter and also MFCC coefficients.

Step 8: Discrete Cosine Transform of speech-music Signal

MFCC coefficients have a high correlation. Discrete Cosine Transform (DCT) can reduce the dimension of their

correlation. Calculate the MFCC coefficient using the discrete cosine transform to obtain the L dimension MFCC parameter $A_l, l = 1, 2, \cdots, L$, as shown in Formula (7):

$$A_l = \sum_{k=1}^{p} \lg(d_k) \cos\left[(kl - (1/2)l)(\pi / p)\right] \tag{7}$$

Where, p is the number of triangular filters. After DCT, MFCC characteristic parameters of vocal music are obtained.

2) Beat spectrum of vocal music signal: Beat spectrum is a measure of acoustic self-similarity and a function of time delay. A highly structured or repetitive concert has a strong beat spectrum peak, which reveals the relative intensity of rhythm and a specific beat. Therefore, different kinds of rhythms of the same beat can be distinguished. The pitch spectrum of vocal music first needs to use cosine similarity to calculate the similarity between two pairs of MFCC feature parameters of vocal music [18] to obtain a similarity matrix and then obtain it by evaluating the autocorrelation of the similarity matrix.

The MFCC parameter of vocal music is regarded as the feature vector of the vocal music signal. The alikeness among them is evaluated by evaluating the cosine value of the angle between the two vectors. The cosine value of 0° is 1, while the cosine value of any other angle is not greater than 1, and the minimum value is - 1. Music has the characteristics of rhythm, which makes it repeatable to calculate its similarity, while voice does not. The cosine value is Formula (8):

$$\cos\theta = (A_l(i)A_l(j)) / (\|A_l(i)\| \times \|A_l(j)\|) \tag{8}$$

In the formula, $A_l(i)$ and $A_l(j)$ represent the two eigenvectors of the vocal music signal respectively, and $\cos\theta$ represents the calculated cosine value.

Using the cosine parameter of the included angle of the feature vector, the similarity matrix of the vocal music signal is evaluated through Formula (9). The autocorrelation of its similarity matrix calculates the beat spectrum of the signal to obtain the Formula (10):

$$X(i, j) = (A_l(i)A_l(j)) / (\|A_l(i)\|\|A_l(j)\|) \tag{9}$$

$$C(g, h) = \sum_{i,j} X(i, j)X(i + g, j + h) \tag{10}$$

Where, $A_l(i)$ and $A_l(j)$ are the eigenvectors of the i and j frames, respectively, $X(i, j)$ is the similarity matrix, $C(g, h)$ is its symmetric matrix, and the beat spectrum $C(h)$ is obtainable by adding them by row or column.

B. Speech-Music Classification on the Basis of Convolutional Neural Network

The neural network has the characteristics of distributed storage, parallel processing and self-learning skill, so it has a wide application view in data processing, pattern identification, classification and other areas. Convolutional Neural Networks (CNN), as an improved neural network, is one of the commonly used speech-music recognition models at present. Its unique convolution structure ensures the translation invariance of speech-music signals in time and frequency domains.

1) Convolution neural network: CNN is an algorithm for learning multilayer neural network structure, which is created of the input layer, convolution layer, pooling layer, full connection layer and output layer [19].

The sparse interaction and parameter sharing of CNN is able to decrease the number of training principles and network complexity [20]; the invariance of data scaling and translation is conducive to optimizing the network structure, making the model have a stronger generalization ability in feature extraction [21]. The network structure of CNN is shown in Fig. 2.

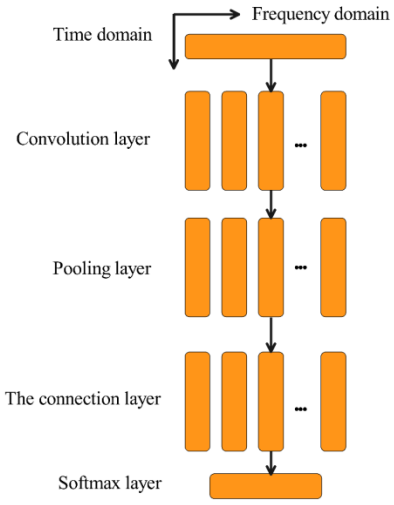


Fig. 2. Structure of convolutional neural network

CNN convolution layer has the characteristics of weight sharing and local connection [22]. Set $Q_{a(i')c(j')}$ as the i' neuron on the input a feature plane, and output the connection weight value between the j' neuron on the c feature plane. Then there is Formula (11):

$$Q_{a(i')c(j')} = Q_{a(i'+1)c(j'+1)} = Q_{a(i'+2)c(j'+2)} \quad (11)$$

In CNN, the output feature surface of each roll-up layer uniquely corresponds to the output feature surface of the pooled layer. The commonly used activation function is the Sigmoid function [23]. Through pooling operation, the features of vocal music signals are further extracted. Common pooling methods include the mean and maximum pooling methods [24]. After the convolution pooling operation, the obtained features enter the full connection layer. Each neuron in the fully connected layer is connected to each neuron in the previous layer, and the fully connected layer can receive all local information in the previous layer.

2) Residual convolutional neural network for depth optimization: Using the idea of a residual network for reference, the input of the upper layer is directly transmitted to the lower layer by adding a congruent mapping layer. This paper combines the input feature x beyond the convolution operation with the outcome after the convolution operation as the intake of the subsequent activation function. Then it outputs new information to build an optimized residual convolution layer. Aiming to raise the number of layers of the depth convolution neural network, Swish and max-out activation functions are introduced to optimize the residual structure, which can directly transmit the input information of the network to the output and improve the degradation and gradient disappearance problems of the network.

The formula of the activation function Swish is $f(x) = x * \mathrm{sig}\,\mathrm{mod}(x)$, which is non-monotonic, smooth and unsaturated.

The function of the activation function max-out is similar to that of the activation function layer in the network structure of the convolution layer and pooling layer. It is attached in front of each output neuron. The highest value of the node outcome of the same group of the hidden layer is taken, and the output of each node in the hidden layer is calculated as Formula (12):

$$W = \varphi(o), o = \omega \times x + \partial \quad (12)$$

In the formula, the linear activation vector o of the neuron is calculated by inputting the sample data vector x, the weight coefficient vector ω and the offset term coefficient ∂. Then the nonlinear transformation is performed through the max-out activation function, where $\varphi(o)$ represents $\max(o, 0)$. Maxout divides the M hidden layer units of the specified layer into I groups. If each group contains V units, the output of the i group is Formula (13):

$$W_{i'} = \max_{v=0}^{V-1} z_{iV+v}, i = 0,1,\cdots,I-1 \quad (13)$$

Since the piecewise linear function can fit any convex function with any precision, and the V hidden layer neuron nodes taken by max-out also have the characteristics of linearity, the operation of taking the maximum value also has the characteristics of piecewise linearity, where the number of segments is related to the size of V value. The max-out activation function has a very strong fitting ability and can fit any convex function [25].

The building of the deep optimization residual convolutional neural network is shown in Fig. 3.

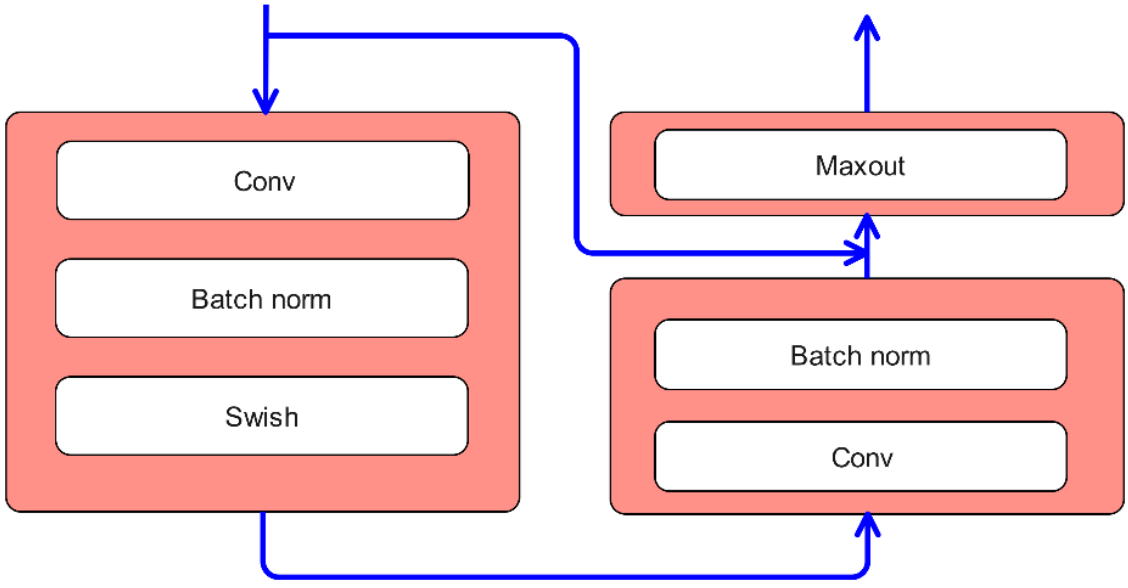


Fig. 3. Building of deep optimization residual convolutional neural network.

Since the max-out function has no weight sharing, it is easy to overfit, so only the max-out activation function is added to the external output of the residual structure. In contrast, the internal convolution layer activation function uses the Swish function to replace the ReLU function. Although both Swish and ReLU function images have no upper and lower bounds, they can increase the nonlinear mapping performance of the network model from the input feature space to the high-dimensional space. The difference is that Swish has the characteristics of non-monotone and smooth. When the sample data of the input dimension is negative, the gradient of the ReLU activation function is 0, and the network model parameters cannot be updated effectively, resulting in the problem of gradient disappearance. However, the first derivative of the Swish activation function is smooth and will not produce saturation. The network model parameters can be updated normally to avoid the disappearance of the gradient, which can improve the training effect of the depth model to a certain extent.

The above depth optimization residual convolutional neural network process can be expressed as follows: Firstly, the acquired vocal music signal beat spectrum is input into the convolutional neural system. Following the convolution process, batch normalization and Swish activation function processing, it is summed with the original input that has not undergone the convolution operation of this layer. The output results obtained are used as the input of the max-out activation function of the next layer, thus forming a group of optimized residual convolutional network structures; a complete depth optimization residual convolutional neural network model is constructed by superimposing the layers of multilayer optimization residual convolutional network structure.

3) Classification of the connection sequence: In the process of vocal music recognition, vocal music generates a real value of training during training, which is in comparison with the anticipated value in the vocal music type. The loss function in the output layer is used to anticipate the degree of inconsistency between the anticipated and the real value. The loss function represents the robust performance of the established model. The smaller the loss function, the better the robust performance of the model.

CTC is a loss function which is used to measure the difference between the input sequence data and the real output after passing through the neural network [26]. CTC introduces an empty node, which does not require full alignment of voice frames. CTC is applied to the speech-music classification in this paper. As the objective function of the softmax layer, CTC will optimize the likelihood between the input and output target sequences.

CTC adopts the maximum likelihood function as Formula (14):

$$B(q) = \sum_{(x,z)\in q} B(x,z) \tag{14}$$

The CTC loss function is defined as Formula (15):

$$B(q) = -\ln \prod_{(x,z)\in q} P(z|x) = -\sum_{(x,z)\in q} \ln P(z|x) \tag{15}$$

Where, $P(z|x)$ represents the probability of the output sequence z for a given input x, and q is the training set. When the input is given, the function of CTC is to find the output sequence with the highest probability.

4) Speech-music classification model based on DCNN-CTC: The convolution layer and pooling layer structure in the convolution neural network can help the model accurately recognize the slight deformation and displacement of the input features. As an end-to-end structure, CTC can resolve the issue of sequence misalignment in the recognition procedure of vocal music. However, the disadvantage of the shallow neural network model is that the feature information extracted by the convolution neural network is not significant enough, which affects the final classification and recognition effect. Therefore, this paper proposes a new DCNN-CTC classification model based on the optimized depth residual convolutional neural network. The model raises the number of convolution layers through the residual structure and constructs the end-to-end model structure of the deep convolution neural network, which is created of six convolution layers, a pooling layer and two filled connection layers. The DCNN-CTC vocal music classification model is shown in Fig. 4.

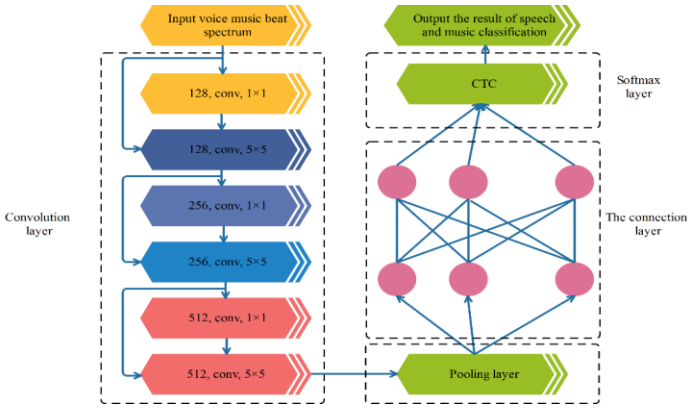
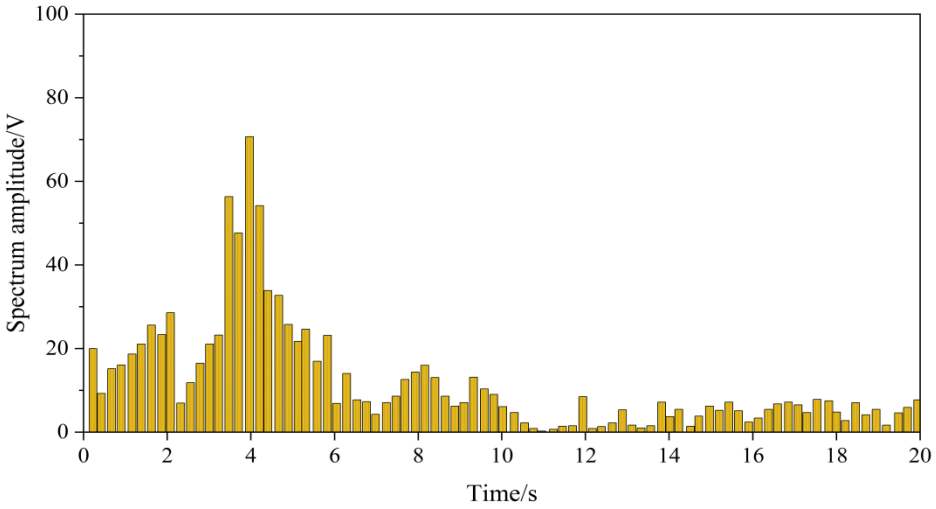


Fig. 4. Speech-music classification model of DCNN-CTC.

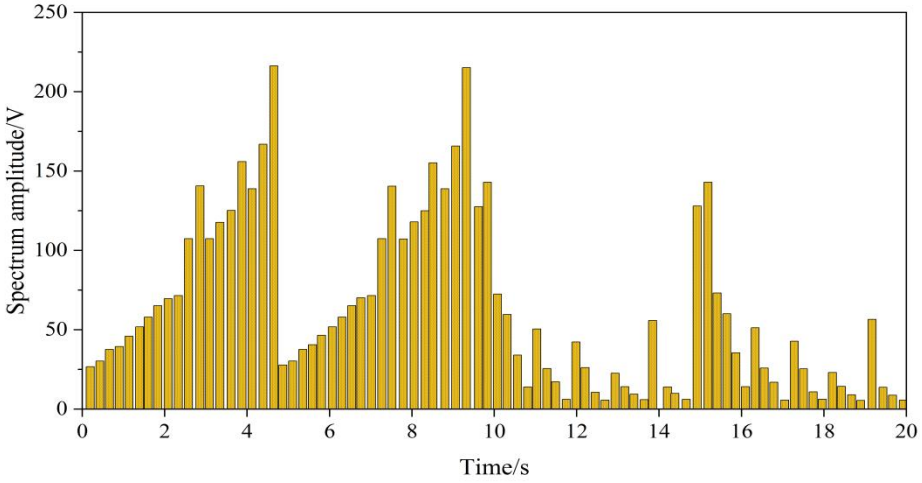
When using the DCNN-CTC vocal music classification model for training, the pitch spectrum of the vocal music signal is first input into the depth convolution neural network. The convolution layer uses the residual structure. After six layers of convolution operations, the pitch spectrum of the original vocal music is extracted for depth features so that the local information of the audio features can be fully integrated. The step size of the convolution layer is set to 1×1. The first convolution layer in the residual structure is to reduce the dimension of the vocal music signal. The size of the convolution core is 1×1. Then the output of the convolution layer is imported to the pooling layer for sampling through maximum pooling. The full connection layer of the two layers has 1024 nodes, and its activation function is maxed out. The softmax layer adopts an end-to-end CTC structure to classify the input vocal music. The end-to-end CTC structure can further improve the robustness of the model, speed up decoding, and obtain the state output of features after classification. The supervised gradient descent method is used for model training in this paper.

III. Experimental Analysis

Take two audio segments in a radio station as the experimental object, and the time length is the 20s. In order to confirm the validity of this type, the test accomplished a classification test on this segment of vocal music. First, the vocal music signal is pre-emphasized and windowed by frames. Set the frame length to 30, the frameshift to 15, and add 0 if the frame length is less than 30. The window function is Hamming window. After MFCC feature parameter extraction and cosine similarity calculation, the vocal music beat spectrum is obtained. Input the acquired vocal music beat spectrum into the model in this paper. The convolution layer of the model in this paper is set as 6 layers. After three residual block structures, the extracted characteristics are compressed and extracted through the maximum pooling method. Finally, they are transferred to the softmax layer by two filled connection layers, which use the CTC loss function. The supervised random gradient descent method is used for model training, with 128 iterations and an initial learning rate of 0.01. After the experiment, the classification results of this segment of vocal music are shown in Fig. 5 and Table I.



（a） Audio 1 signal recognition

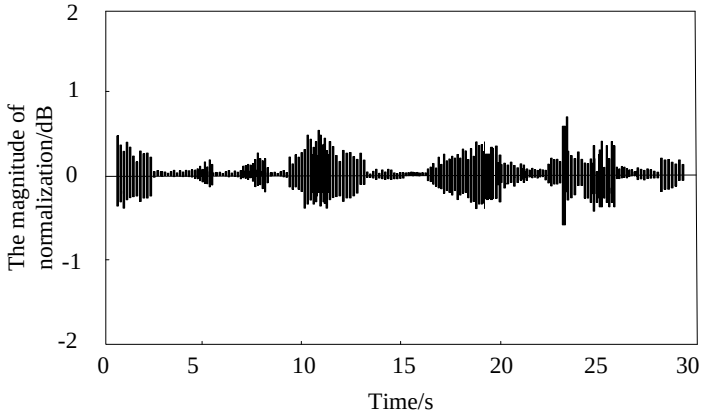


（b） Audio 2 signal recognition

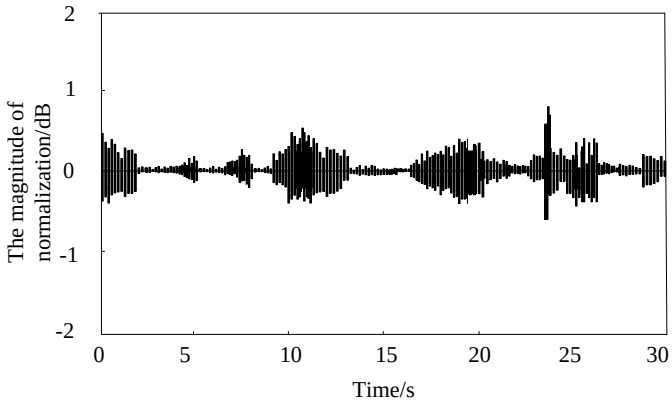
Fig. 5. Recognition results of speech and music signals.

TABLE I. AUDIO CLASSIFICATION RESULTS

The name of the audio	The classification results
Audio 1	Speech signal
Audio 2	Music signals



（a） Signal diagram before endpoint detection of speech and music signal.



（b） Signal diagram after endpoint detection of speech and music signal.

Fig. 6. Signal comparison before and after endpoint detection of speech and music signals.

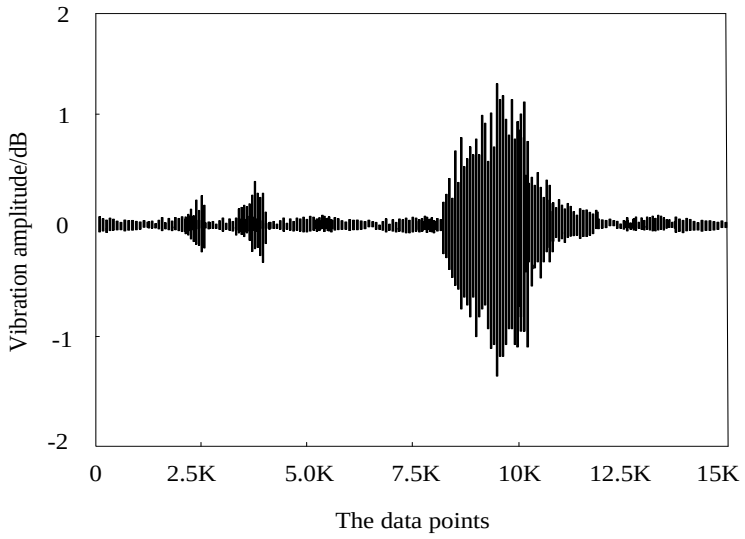
Due to the periodicity of the music signal, it has a strong beat, while most of the voice is irregular, and the beat is weak. It is perceivable from Fig. 5 that the peak value in the vocal music signal obtained by using the model in this paper corresponds to its main rhythm component, and the amplitude of different peaks reflects the strength of the corresponding rhythm of the signal. In addition, the peak value change of some vocal music signals with a strong sense of rhythm will be relatively obvious. In contrast, the peak value change of vocal music signals with a weak sense of rhythm will be slightly weak, which is able to invert the main features of the vocal music signal better. It can be seen from the identification diagram of audio 1 signal that the signal has no regularity and a weak beat; it can be seen from the identification diagram of audio 2 signal that the signal changes periodically, with a regular period from 0s to 9.6s and a regular period from 9.7s to 20s, and the beat is strong.

Through the above analysis, it shows that this model is effective for the vocal music classification.

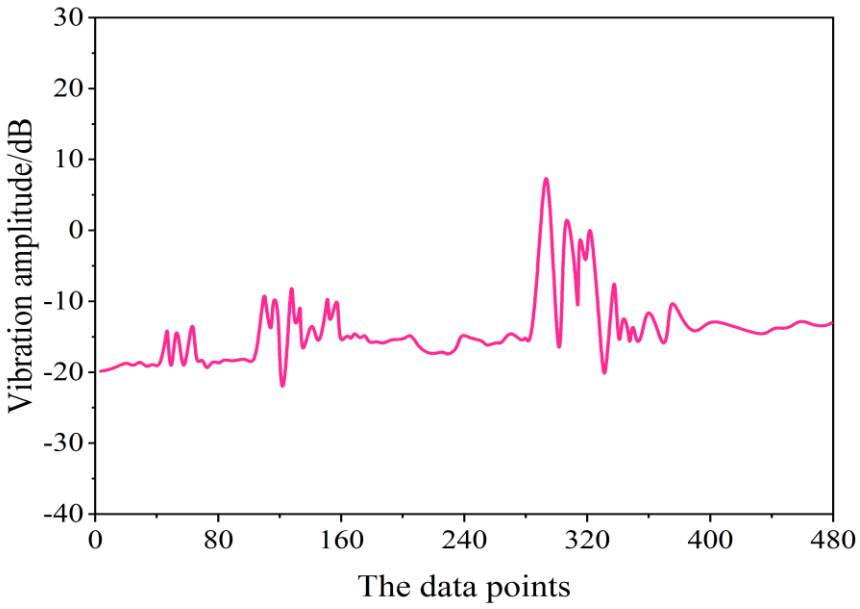
In order to measure the effect of endpoint detection on the vocal music signal mentioned in this paper, the experiment collected a section of vocal music signal on the radio station. The sampling frequency is 8kHz, the accuracy is 16bit, and the duration is 30s. The result of endpoint detection of this signal is shown in Fig. 6.

It is perceivable from Fig. 6 that after endpoint detection of the vocal music signal using the model in this paper, the audio period with low head and tail energy of the signal is removed. The start frame and end frame of the whole audio signal are determined, which can effectively reduce the amount of data and calculation of subsequent audio feature extraction and improve the stability of the model.

To judge the excellence of MFCC characteristic parameter extraction proposed in this model, an audio file intercepted from the radio station is input into the Matlab tool in the experiment. The audio file data and sampling frequency are read using the audio read function, and the audio file is in wav format. Set the audio frame length to 256 points, frameshift to 80, filters to 24 groups, and dimensions to 12. Use this model to extract the MFCC feature parameters of this audio, and the results are shown in Fig. 7.



(a) Raw audio signals



(b) Audio signal after MFCC feature parameter extraction

Fig. 7. Audio signal comparison before and after MFCC feature parameter extraction.

In Fig. 7, comparing the collected original signal with the signal after MFCC feature parameter extraction, we can see that the collected original signal has 15K data points, and the audio signal data points after MFCC parameter extraction are only 480. After amplification and comparison, the audio signal waveform after MFCC parameter extraction is more consistent with the trend and trend of the original signal waveform; after MFCC parameters are extracted, 480 data points completely reflect the characteristics and change trend of 15K data points. From this, it can be seen that the MFCC feature parameter extraction proposed in this model can reflect the characteristics of the entire audio signal, which is conducive to subsequent analysis of the signal and lays a good foundation for the next step of speech-music classification.

The convolution layer is an important part of feature learning and extraction in the convolutional neural network structure. Different convolution layers will lead to different feature representations. Therefore, the experiment tests the performance of the residual convolutional neural network with depth optimization when the convolution layer is 3, 4, 5 and 6 layers. The preprocessed vocal music signal data set is divided into a training data set and a test data set at a ratio of 7∶3 for experiments to determine the number of convolution layers with the best classification performance. The experiment sets the input length, the characteristic dimension and the number of iterations obtained by the full connection layer as 150, 3600 and 1200, respectively. It obtains the classification accuracy and training loss of the depth-optimized residual convolutional neural network under different convolution layer structures, as shown in Fig. 8 and Fig. 9.

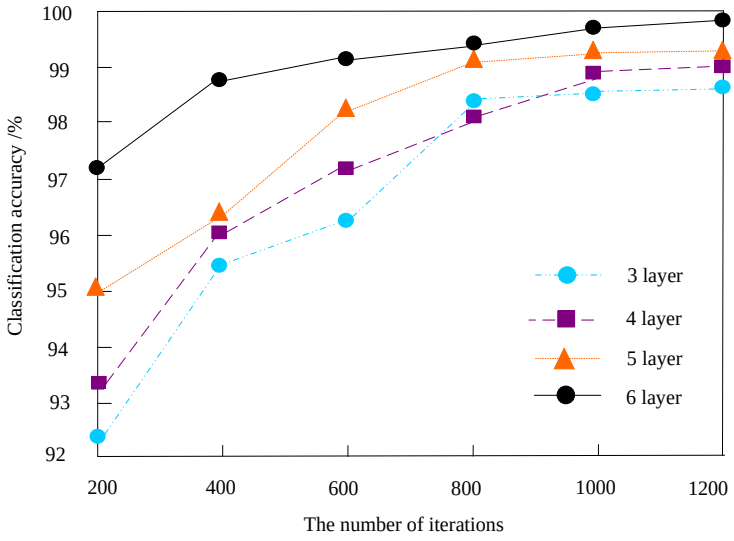


Fig. 8. Comparison of classification accuracy of deep optimized residual convolutional neural networks with different convolutional layer structures.

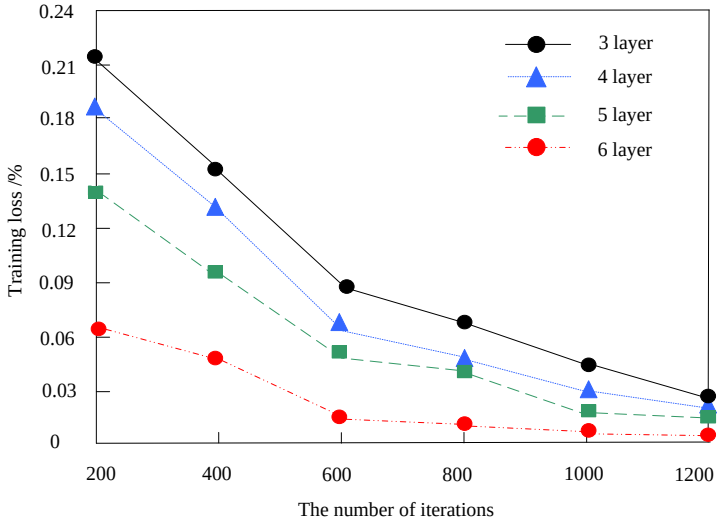


Fig. 9. Comparison of training losses of deep optimization residual convolutional neural networks with different convolutional layer structures.

It is perceivable from Fig. 8 that for the depth-optimized residual convolutional neural network figure with a 6-layer convolution layer structure, when the number of iterations is 600, the classification accuracy has reached more than 99%; when the number of iterations is 1200, the classification accuracy has reached 99.87%, approaching 100%; while for the depth optimized residual convolutional neural network model with other hierarchical structures, the performance in classification accuracy is slightly worse, which shows that, When six convolution layers are set in the depth optimization residual convolutional neural network model, it has a strong classification ability.

It is perceivable from Fig. 9 that when the number of iterations is between 400 and 600, the training loss of the deep optimization residual convolutional neural network decreases the most, especially for the convolution neural network with six convolution layers; the training loss is only about 0.01 at this time. By the time the number of iterations is 1200, the training loss is close to zero, and the training loss of other hierarchical networks is inferior to that of the convolution neural network with six convolution layers; at the same time, it also shows that the network can better avoid the overfitting phenomenon.

In conclusion, when the convolution layer of the depth optimization residual convolutional neural network model proposed in this method is 6 layers, it performs well both in classification performance and training loss. It is superior to other hierarchical convolution neural networks, with good classification performance.

For classification models, sensitivity and specificity are two important performance evaluation indicators. Sensitivity represents the proportion of all positive samples to be paired, which measures the recognition ability of classification

models to positive samples; Specificity represents the proportion of all negative samples divided into pairs and measures the recognition ability of classification models to negative samples. The sensitivity and specificity are Formula (16) and (17), respectively:

$$sensitive = \frac{TP}{P} \tag{16}$$

$$specificity = \frac{TN}{N} \tag{17}$$

Where, P represents a positive sample, N represents a negative sample, TP represents a positive sample classified as a positive sample, and TN represents a negative sample classified as a negative sample.

The experiment selects 100000 vocal music audios from the audio database of the radio station as the sample set, of which 46000 are voice audio, set as positive samples, and 54000 are music audio, set as negative samples. Under the conditions that the signal-to-noise ratio is 20dB, 40dB, 60dB, 80dB, 100dB, 120dB, 140dB, 160dB, and 180dB, the model in this paper is tested from the perspective of sensitivity and specificity. The final results are shown in Fig 10.

It is perceivable from Fig. 10 that under different signal-to-noise ratios, the method in this article has a good performance both in sensitivity and specificity. When the signal-to-noise ratio is 100dB, the sensitivity and specificity reach 99%. When the signal-to-noise ratio is 180dB, the sensitivity and specificity reach 99.98% and 99.96%, respectively. It is perceivable that the model in this article has good recognition ability and classification effect for speech-music.

To measure the function of the model in this article further, for vocal music classification, the experiment selects different types of audio from the audio library of the radio station as the sample set; the sampling rate is 32kHz, the format is PCM format, the whole number of demos is 30min, in which the voice duration is 18min, including the pronunciation of men and women of different ages, different speakers, and different sentences; The music lasts for 12 minutes, including various instrumental music, such as violin, piano, zither, flute, etc. 70% of the samples in the sample set were chosen as training demos and 30% as test demos. The experiment compares the classification model of zero crossing rate and spectrum in reference [7], the classification model of the residual network and random forest in reference [8], and the classification model of Meyer spectrum and cyclic residual in reference [9]with the model in this paper, and analyzes the classification results of different models for vocal music, as presented in Table II.

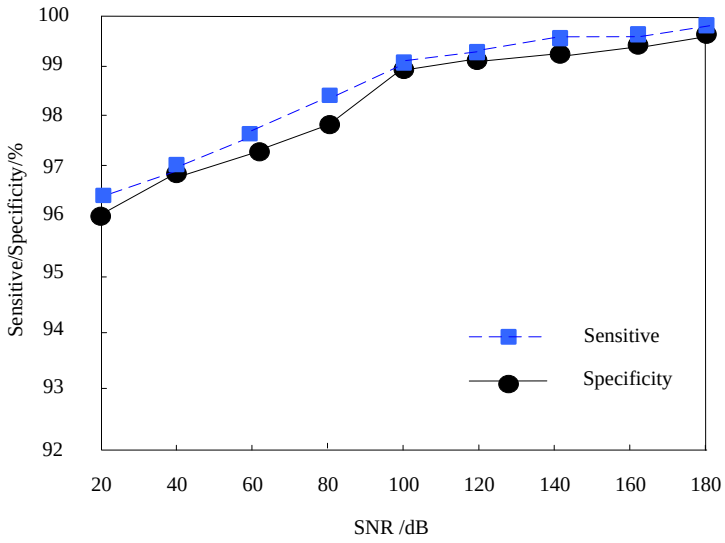


Fig. 10. Sensitivity and special effect of the proposed model under different SNR.

TABLE II. COMPARING CLASSIFICATION ACCURACY OF DIFFERENT MODELS

Methods	Audio type	The total number of samples	Number of corrected samples	Number of wrong samples	Classification accuracy /%	The average classification accuracy /%	Running time /s
Model in reference [7]	Voice	273	265	8	97.07	95.47	0.75
	Music	147	138	9	93.88		
Model in reference [8]	Voice	273	269	4	98.53	97.23	0.59
	Music	147	141	6	95.92		
Model in reference [9]	Voice	273	259	14	94.87	93.35	0.67
	Music	147	135	12	91.84		
In this paper, the model	Voice	273	272	1	99.63	99.82	0.48
	Music	147	147	0	100.00		

It can be seen from Table II that different models are used to classify different types of vocal music. The accuracy of the model in this paper for music recognition reached 100%, and there was only one error in speech recognition, with an average classification accuracy of 99.82%. It has been proven that the proposed model has good classification performance and the running time is the shortest compared to other algorithms, only 0.48s. While the average classification accuracy of the reference [7] model for vocal music is 95.47%, the running time is 0.75 seconds, the mean classification correctness of the reference [8] model is 97.23%, and the running time is 0.59 seconds. The mean classification correctness of the reference [9] model is 93.35%, and the running time is 0.67 seconds. It can be seen from the above data that compared with the other three models, the vocal music classification in this model is the most accurate, and the classification speed is the fastest. It has a good classification function and convergence effect.

IV. Discussion

The speech-music classification model based on improved neural networks and beat scores is a research aimed at improving the accuracy and efficiency of music classification. This study designed a novel model that achieved impressive results by combining an improved neural network architecture and beat spectrum feature extraction method.

Firstly, the model in this study performed well in terms of music classification accuracy, achieving an accuracy rate of over 99%. This means that the model can highly accurately classify different types and styles of music, providing users with more accurate music recommendations and personalized experiences.

Secondly, after 1200 iterations of training, the training loss of the model approaches zero. This indicates that the model can fully learn and adapt to music datasets, and has strong learning and generalization abilities.

In addition, the model exhibits very high sensitivity and specificity at high signal-to-noise ratio (180dB), reaching 99.98% and 99.96%, respectively. This means that even in noisy environments, the model can still accurately recognize and classify music, providing users with stable and reliable classification performance.

Finally, the study also focuses on the runtime of the model. After optimization, the running time of the model is only 0.48 seconds, which means that the model has high efficiency and practicality, and can quickly classify music in real-time applications.

In summary, the speech-music classification model based on improved neural networks and beat scores is an exciting research topic. By designing models with high accuracy, low training loss, high sensitivity, and special effects, and completing classification tasks in a short running time, this study has made significant contributions to the development of music classification and provided strong support for music recommendation systems and other related applications.

V. Conclusion

With the development of cloud storage and Internet technology, more and more multimedia data, such as audio, has entered people's lives. In order to save local storage space, many individuals and enterprise users store multimedia data in the cloud, which increases the technical burden of multimedia data retrieval. Classifying audio is an effective means to achieve fast retrieval. As the two most important types of audios - voice and music, their classification has important application value in content-based audio retrieval, video retrieval and summarization, and voice document retrieval. It is an important preprocessing work in sound signal processing. Therefore, this paper proposes a vocal music classification model based on an improved neural network and beat spectrum. This model combines CTC and residual network design ideas, proposes an improved depth optimization residual convolutional neural network structure, and introduces the beat spectrum as the model's input. The model in this paper effectively solves the problem of model overfitting and improves classification accuracy. The innovation work of this paper mainly includes:

1) Design the pitch of vocal music. In the characteristic extraction of the original audio signal, the beat feature is added to enhance the precision of characteristic extraction.

2) The introduction of residual structure. The residual structure is introduced into the convolutional neural network to increase the number of convolution layers and increase the number of convolution layers to 6 layers, which deepens the depth of the network model, extracts the features of deeper vocal music signals, and better enhances the classification correctness of the network figure.

3) Introduction of CTC. The softmax layer of the convolutional neural network adopts an end-to-end CTC structure, which can resolve the issue of sequence misalignment in the recognition procedure of vocal music, improve the robustness of the model, speed up decoding, and obtain the state output of features after classification.

Through experiments, it has been proven that the application of this model can effectively achieve accurate classification of speech and music, with small errors, high accuracy, and fast speed.

In order to further improve the performance of the design in practical applications, in-depth research will be conducted in the following aspects in the future:

1) Dataset expansion: Find more datasets for training and evaluating model performance. These datasets can contain different types and styles of music, as well as different language and cultural backgrounds.

2) Model architecture optimization: try different neural network architectures, such as Convolutional neural network (CNN), Recurrent neural network (RNN) or attention mechanism, to improve the modeling ability of the model for music features.

3) Multimodal learning: Combining audio data with data from other modalities (such as images or text) to improve the

performance of music classification models. For example, audio data can be combined with lyrics or album cover images for joint training.

DATA AVAILABILITY

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

CONFLICTS OF INTEREST

The authors declared that they have no conflicts of interest regarding this work.

ACKNOWLEDGMENT

The work is not supported by any funding.

REFERENCES

[1] X. Huayu, Y. Qin, R. Pin, and L. Ruisen, "Audio classification based on machine learning," Computer Engineering and Design, vol. 42, no. 1, pp. 156-160, 2021.

[2] P. Dhakal, P. Damacharla, A. Y. Javaid, and V. Devabhaktuni, "A near real-time automatic speaker recognition architecture for voice-based user interface," Machine learning and knowledge extraction, vol. 1, no. 1, pp. 504-520, 2019.

[3] K. Zhang, Y. Su, J. Wang, S. Wang, and Y. Zhang, "Environment sound classification system based on hybrid feature and convolutional neural network," Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University, vol. 38, no. 1, pp. 162-169, 2020.

[4] M. T. M. Scheffers, "Discrimination of fundamental frequency of synthesized vowel sounds in a noise background," The Journal of the Acoustical Society of America, vol. 76, no. 2, pp. 428-434, 1984.

[5] C. Lim and J. H. Chang, "Enhancing support vector machine-based speech/music classification using conditional maximum a posteriori criterion," IET signal processing, vol. 6, no. 4, pp. 335-340, 2012.

[6] D. Sammler, "Splitting speech and music," Science, vol. 367, no. 6481, pp. 974-976, 2020.

[7] H. Sun, H. Long, Y. Shao, and Q. Du, "Speech music classification algorithm based on zero-crossing rate and frequency spectrum," Journal of Yunnan University (Natural Science Edition), vol. 41, no. 5, pp. 925-931, 2019.

[8] X. Zhang, "An audio recognition method based on residual network and random forest," Computer Engineering & Science, vol. 41, no. 04, p. 727, 2019.

[9] L. Zhu, K. Qian, Z. Wang, B. Hu, Y. Yamamoto, and B. W. Schuller, "Heart Sound Classification based on Residual Shrinkage Networks," 2022: IEEE, pp. 4469-4472.

[10] J. Tyler, H. Zou, H. Zhou, H. Su, and J. Braasch, "Automated Mandarin tone classification using deep neural networks trained on a large speech dataset," The Journal of the Acoustical Society of America, vol. 145, no. 3, pp. 1814-1814, 2019.

[11] T. Oikarinen et al., "Deep convolutional network for animal sound classification and source attribution using dual audio recordings," The Journal of the Acoustical Society of America, vol. 145, no. 2, pp. 654-662, 2019.

[12] A. Abeysinghe, M. Fard, R. Jazar, F. Zambetta, and J. Davy, "Mel frequency cepstral coefficient temporal feature integration for classifying squeak and rattle noise," The Journal of the Acoustical Society of America, vol. 150, no. 1, pp. 193-201, 2021.

[13] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," Expert Systems with Applications, vol. 90, pp. 250-271, 2017.

[14] Y. W. Chen, K. Li, Y. Han, and Y. P. Wang, "Musical Note Recognition of Musical Inst ruments Based on MFCC and Constant Q Transform," ed, 2019.

[15] Z. Lin, C. Di, and X. Chen, "Bionic optimization of MFCC features based on speaker fast recognition," Applied Acoustics, vol. 173, p. 107682, 2021.

[16] T. Wang, Q. Bao, and P. Qin, "Environmental sound classification method based on Mel-frequency cepstral coefficient, deep convolution and Bagging," Journal of Computer Applications, vol. 39, no. 12, p. 3515, 2019.

[17] L. Meng and J. Johnson, "High performance implementation of the TFT," 2014, pp. 328-334.

[18] Y. Nagai and K. Katayama, "Multivariate curve resolution combined with estimation by cosine similarity mapping of analytical data," Analyst, vol. 146, no. 16, pp. 5045-5054, 2021.

[19] B. Matityaho and M. Furst, "Classification of music type by a multilayer neural network," The Journal of the Acoustical Society of America, vol. 95, no. 5, pp. 2959-2959, 1994.

[20] H. Sinha, V. Awasthi, and P. K. Ajmera, "Audio classification using braided convolutional neural networks," IET Signal Processing, vol. 14, no. 7, pp. 448-454, 2020.

[21] C. D. Feng, L. I. Shao-Bo, Y. Yao, and J. Yang, "Environmental sound recognition with improving convolutional neural networks and learning rate decay," 2019.

[22] W. Liu, Q. Zeng, Y. Bu, and Z. Zheng, "Speech recognition method based on dual micro-array and convolutional neural network," Journal of Computer Applications, vol. 39, no. 11, p. 3268, 2019.

[23] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," 2017: IEEE, pp. 1-5.

[24] Y. Lu, Z. Zhang, G. Lu, Y. Zhou, J. Li, and D. Zhang, "Addi-reg: A better generalization-optimization tradeoff regularization method for convolutional neural networks," IEEE Transactions on Cybernetics, vol. 52, no. 10, pp. 10827-10842, 2021.

[25] Z.-W. Wang, S.-K. Wang, B.-T. Wan, and W. W. Song, "A novel multi-label classification algorithm based on K-nearest neighbor and random walk," International Journal of Distributed Sensor Networks, vol. 16, no. 3, p. 1550147720911892, 2020.

[26] H. Li and W. Wang, "Reinterpreting CTC training as iterative fitting," Pattern Recognition, vol. 105, p. 107392, 2020.