International Journal of Web-Based Learning and Teaching Technologies
Volume 18 • Issue 1

Research on Musical Tone Recognition Method Based on Improved RNN for Vocal Music Teaching Network Courses

Kaiyi Long, Hunan Mass Media Vocational and Technical College, China*

ABSTRACT

The test results show that the fast Fourier process with multiple time superposition and a dimension length of 40 is most beneficial to the accuracy of the model. The loss curve value of the convolutional recurrent network model (CRN) is much lower than the other three models. The music tone recognition model learns better. The accuracy rate value and recall rate value of the CRN are the highest, and the accuracy rates of the four music tone indicators are 94.6%, 92.4%, 93.5%, 92.5%, and the recall rates were 93.2%, 94.9%, 95.2%, and 88.6% respectively; the improved algorithm was the most accurate in terms of F1 values and is suitable for use in vocal music teaching courses. The results show that the algorithm can be broadly performed in the zone of music tone recognition and has a certain contribution to the development of the field of music tone recognition.

KEYWORDS

Gating Mechanism, Meier Inversion Coefficients, Musical Note Recognition, Multilayer Perceptron, Recurrent Neural Networks

1. INTRODUCTION

Reforms in computer science have driven the development of online web-based teaching, which has led to the diversification of vocal music courses. A large number of institutions have introduced online vocal teaching courses, which can help students to correct their vocal style and improve their model singing ability, make it easier to scientifically fulfil the training objectives of music majors, help students to better understand music, and occur an essential position in the quality education of students. However, there are many problems that cannot be ignored in the process of training musicians and students learning vocal music in schools. Different students have different basic musical qualities and it is difficult to achieve the teaching objectives by relying on teachers to require uniform educational work. It is also difficult for teachers to analyses and understand each student comprehensively, which does not ensure that students receive targeted training in learning vocal music (Li 2021, Fu 2021). The basic idea of musical sound recognition is to extract the vocal characteristics of a musical

DOI: 10.4018/IJWLTT.327948 *Corresponding Author

instrument and input them to a classifier for musical sound recognition, the ability of which depends on the performance of the musical features and the discriminatory ability of the recogniser. With the advent of deep learning techniques, the development process of audio recognition technology has been significantly accelerated. Computer technology can save a great deal of energy, time and financial resources to replace the teacher in the task of identifying different musical notes in a vocal music teaching course, and computer algorithms can also perform the identification of singing and even automate the creation of musical scores. For the same instrument, the notes, loudness and timbre can bring different musical feelings. With the continuous development of music information retrieval technology, there is a rich variety of musical characteristics, and the existing music recognition technology is mostly a simple combination of all musical characteristics, but this does not meet the high technical requirements of music recognition. The Mel frequency cepstral coefficient and linear predictive cepstral coefficient are important technical tools for the study of the cepstral domain, and the selection, screening and combination of features are the next research focus of the current music recognition model. On the other hand, there is a large amount of noise interference in the music teaching curriculum, which leads to the performance of existing audio classifiers not meeting the practical needs. Most of the current research methods are still based on traditional signal processing methods, and there is less research on music recognition using deep neural networks, and there is still much room for improvement in terms of recognition accuracy and efficiency. In particular, the Mel frequency cepstrum coefficients are mostly selected manually, making the development of the basic feature recognition module of the music recognition model more restricted; and the traditional sound recognition algorithm model is still not good enough in terms of recognition accuracy and efficiency as well as the detection of abnormal noise. Against this background, in order to enhance the ability to extract musical tone features and the perception of timbre recognition, improve the accuracy of musical tone recognition, design a high-performance musical tone recognition model; and combine it with the vocal music teaching course to improve the teaching shortcomings of the traditional vocal music teaching course and realise the informatization of the vocal music teaching course, the research uses multi-layer perceptron and Meier's inverse spectral coefficient for musical tone feature recognition, changing the The fusion of multiple network models differs from a single network model in that it is more resistant to noise interference and has higher recognition accuracy. The article is divided into four parts: the first part summarises and outlines the relevant research and studies in this field and analyses the shortcomings of existing research; the second part provides a detailed description of the proposed methodology; the third part discusses and analyses the experimental results of the performance verification; and finally, the conclusions of the study are outlined.

2. RELATED WORKS

Recent years, audio recognition is one of the hot issues of sound information extraction research, which has attracted comprehensive attention from scholars. Sound recognition and classification have also been studied extensively. The first is about feature extraction and recognition of musical sounds of musical instruments. To advance the scientific advancement of music teaching, Sumarno and his team constructed an electronic organ tone recognition database based on artificial intelligence, which can flexibly generate a library of electronic organ sounds with different timbres. The harmonic peak method with improved confidence allows for pitch recognition, and combined with the timbre parameters the frequency domain information of each synthetic timbre frame can be calculated. The overall recognition rate of 3762 notes and 286 beats is experimentally validated to be 88.6% (Sumarno & Chai 2021). Liu Nan fuses the local feature extraction ability of the two layers to summarize sequence features to design a convolutional recurrent neural network (CRNN) for music hum recognition using TensorFlow and Keras as the framework. The results verified that the model can greatly enhance the accuracy of hum recognition, shorten the recognition time and obtain audio features with higher complexity (Liu 2022). Cui and his team raised an improved audio recognition

system based on the target detection method You Only Look Once-v4. The system uses a stacking approach to fuse independent sub models of different channels and combines them with a spatial pyramid pooling module to enhance the generalization of data in diverse audio formats. Experiment verify that the proposed model upgrades the performance of audio recognition and exhibits better generalization capabilities compared to other deep learning techniques (Cui & Wang 2022). To achieve retrieving and recognizing cross-media audio, Wang Tianshu proposed a neural network-based dynamic threshold-based segmentation and weighted integrated matching algorithm, which dynamically sets the magnitude difference step and segments notes (Wang 2022). For recognizing the performance skill of traditional instruments, Li and his team built an 8-layer neural network based on residual networks incorporating support vector institutions for identifying the similar instruments performing. The results demonstrated that the recognition accuracy was 95.7%, 82.2%, 88.3% and 97.5% for wind, plucked, string and percussion instruments, respectively (Li & Zhang 2022). Tanaka K and his team proposed a representation learning method for decomposing instrument sounds through variational autoencoders, introducing a metric learning technique that brings similar timbres close to each other and different timbres far from each other. Under weak supervision of machine learning to determine whether musical timbres are the same, experimental results show that the method improves the ability to generalise musical sounds for different instruments, with structured decomposed representations for all different instruments (Tanaka et al. 2022). Optical music recognition can process images of parts of a pentatonic score and retrieve the musical notation present in them. garrido Munoz Carlos and his team propose a neural network structure for reading musical notation from document images in an end-to-end manner, and experimental results demonstrate the effectiveness in retrieving graph structures from extracts of handwritten musical notation (Garrido et al. 2022).

In addition to the extraction of musical instrument pitch and timbre features, there have been many studies of recognition and classification systems for other sounds. To solve the high-cost problem, time consuming and low accuracy of bird sound recognition, Xiao and his team combined attention mechanism and residual network to construct a new automatic bird call recognition model, which can automatically extract and select high-dimensional features. Experimentally, 12,651 bird sound samples from real environments were selected for training 10-fold cross-validation and testing, and the model was found to be suitable for accurate bird sound classification with 92.6% accuracy, 3.1% higher than the other optimal models, and also performed well in terms of recall values and accuracy (Xiao et al. 2022). Wu addresses the car engine sound recognition feature extraction. A time-frequency image recognition method based on deformable feature map residual network is proposed to address the problem of difficult feature extraction for car engine sound recognition. The extracted features are combined with Meier frequency cestrum coefficients and the fused results are classified by squeezing and excitation block recalibration. The experiments show that the method shows a large improvement in the accuracy of recognizing car engine sounds (Wu et al. 2022). The main constraint to the development of airborne target recognition is the lack of airborne target audio data. To address this obstacle, Wu proposes an extended learning method to use wave nets as a generator of airborne target audio. Experimental results show that this audio recognition generator can achieve up to 99.50% accuracy by mixing the original and generated audio in a four-to-one ratio (Wu et al. 2022) To improve speech recognition accuracy in noisy systems, Maghraby E has developed a speech recognition system that combines acoustic and visual speech information based on bi-directional long-term memory to perfect audio recognition performance in furious conditions (Maghraby et al. 2021).

Chen Yue and his team used support vector machines and self-organising graphs to simulate pitch perception in Mandarin. The perceptual model identifies speech classes directly from syllable-sized continuous speech signals, and experimental results show that direct pitch recognition has better performance and is less computationally intensive than other methods of extracting features (Chen et al. 2022). In noisy conference sites, writing meetings with noise interfering with the interference of voice recognition, Khan Isra and his team proposed a framework for speech recognition and audio data classification, where the model first removes noise from the audio, extracts sound features from

the processed audio, performs sound classification training, and allows for audio-to-text conversion, and the proposed model has high accuracy and effectiveness (Khan et al. 2022). To address the weak generalisation ability of existing bird song recognition models, Qiu Zhibin and his team proposed a migration learning-based bird song recognition method. A sample set of bird song signals was built based on historical bird information. The model preprocessed the bird song signals by framing, windowing, noise reduction and clipping operations, extracted the spectrogram and mapped it to a 64th-order filter set to obtain the spectrogram. experimental results showed that the method improved the generalization ability of the model and the accuracy of bird song recognition (Qiu et al. 2022). phan Thi Thu Hong and his team established a honeycomb sound recognition model based on machine learning to study the Mel spectral coefficient features. Experimental results show that the model can significantly improve the accuracy of these Merle spectral coefficient-based models in identifying and classifying bee sounds with other environmental noises, outperforming existing deep learning algorithms (Phan 2022).

It appears that the use of deep learning for acoustic event detection has been widely researched and developed, but deep learning algorithms that consider both audio signal processing and classification are still relatively rare. In order to strengthen the audio classifier's ability to recognise and classify features, it is necessary to consider both feature extraction and model recognition and classification capabilities to design music sound recognition models.

3. CRNN-BASED MODEL DESIGN FOR MUSICAL TONE RECOGNITION IN VOCAL MUSIC TEACHING COURSES

3.1 MLP-Based Model Design for Music Sound Data Pre-Processing and Feature Extraction

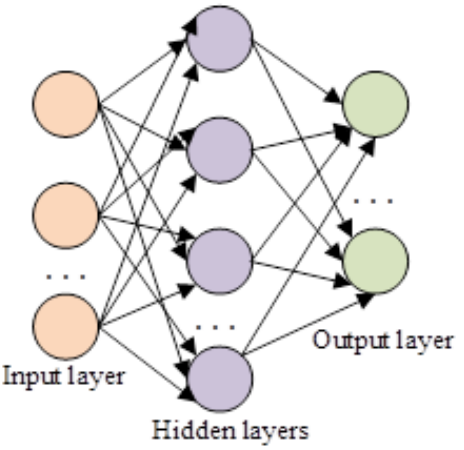
In order to design an algorithmic model for the accurate recognition of a wide range of musical tones, this study firstly uses the Multilayer Perceptron (MLP) to extract musical features using MeI-Freguency CeptraI Coefficients (MFCC); then uses an improved CNN and recurrent neural to construct a musical tone recognition model.

The audio data is first pre-processed, with operations including pre-emphasis, framing and windowing. The audio data is then transformed: from Time Domain-Frequency Domain, and multiple samples over a period of time are assembled into a single frame, with the values of the samples being considered in the context of the data samples. The Hamming window is given in equation (1) expressed as $w\left[i\right]$. N_l refers to the window length. Finally, the endpoint detection is performed using a double threshold technique based on the over-zero rate.

$$w\left[i\right] = 0.54\text{-}0.46\cos\left(\frac{2\pi i}{N_l}\right), 0 \le i \le N_l - 1 \tag{1}$$

After pre-processing, audio data contains a lot of semantic information, and audio feature extraction is a key step in the training of music sound recognition models. The study builds an audio feature extraction system using MLP, also known as Feedforward Neural Network (FNN), where the model is trained by minimizing the loss function to optimize the model parameters and the MLP neurons perform classification or regression tasks according to their activation functions (Xu et al. 2022; Yildirim et al. 2021). The input, hidden, and output layers form the most typical MLP structure (Figure 1), where the upper neurons are connected to all the lower neurons. MLPs can be used in areas such as image and audio recognition or machine translation.

Figure 1. MLP structure diagram



MLP consists of two learning processes, forward propagation and backward propagation. After inputting data samples, the forward propagation process is shown in equation (2). In equation (2), n_l represents the neurons number in the MLP's layer, l represents the output of the neurons after the activation function. a_i^l, l and z_j^{l+1} mean the output of the j-neuron in the $l+1$-layer without the activation function. w_{ij} is the connection weight between the neurons i and j. b_j is the bias parameter of the j-neuron in the $l+1$-layer.

$$z_j^{l+1} = \sum_{i=1}^{n_l} w_{ij}^{\ l} a_i^l + b_j^{\ l+1} \tag{2}$$

The output value of each neuron a_j^{l+1} is shown in equation (3). Such forward propagation is performed between all the multiple hidden layers of the MLP, and the loss function is calculated built on the output of the output layer.

$$a_j^{l+1} = f^{mlp}\left(z_j^{l+1}\right) \tag{3}$$

In equation (2), $f^{mlp}\left(\cdot\right)$ denotes the activation function and the Sigmoid is used in the MLP. The neuron residuals were calculated using the back propagation process, which is shown in Eq. (4).

$$\delta_i^l = \frac{\partial J\left(w,b\right)}{\partial z_i^l} \tag{4}$$

In Eq. (4), δ_i^l represents the residual of the i neuron of the l layer and $J\left(w,b\right)$ represents the loss function. The loss function is then used to derive the weight parameter and bias parameter, and the derivation process is shown in Eq. (5).

$$\begin{cases} \dfrac{\partial J\left(w,b\right)}{\partial w_{ij}^{l}} = \dfrac{\partial J\left(w,b\right)}{\partial z_{j}^{l+1}} \times \dfrac{\partial z_{j}^{l+1}}{\partial w_{ij}^{l}} = \delta_{j}^{l+1} a_{i}^{l} \\ \dfrac{\partial J\left(w,b\right)}{\partial b_{j}^{l+1}} = \dfrac{\partial J\left(w,b\right)}{\partial z_{j}^{l+1}} \times \dfrac{\partial z_{j}^{l+1}}{\partial b_{j}^{l+1}} = \delta_{j}^{l+1} \end{cases} \tag{5}$$

From (5), the residuals of the i neuron of the l layer can also be calculated by chaining the derivatives, as shown in equation (6).

$$\begin{aligned} \delta_{i}^{l} &= \frac{\partial J\left(w,b\right)}{\partial z_{i}^{l}} = \sum_{j=1}^{n_{l+1}} \frac{\partial J\left(w,b\right)}{\partial z_{j}^{l+1}} \times \frac{\partial z_{j}^{l+1}}{\partial z_{i}^{l}} \\ &= \sum_{j=1}^{n_{l+1}} \delta_{j}^{l+1} \times \frac{\partial \left(\sum_{k=1}^{n_{l}} w_{kj}^{\ l} \times f^{mlp}\left(z_{k}^{l}\right) + + b_{j}^{\ l+1} \right)}{\partial z_{j}^{l}} \\ &= \left(\sum_{j=1}^{n_{l+1}} \delta_{j}^{l+1} w_{ij}^{\ l} \right) \times f^{'mlp}\left(z_{i}^{l}\right) \end{aligned} \tag{6}$$

The process of deriving the derivatives for the weight and bias parameters gives the direction of the gradient of the network parameters, which is updated using the gradient descent algorithm for all parameters in the network, and the update process is shown in equation (7), where λ is the learning rate.

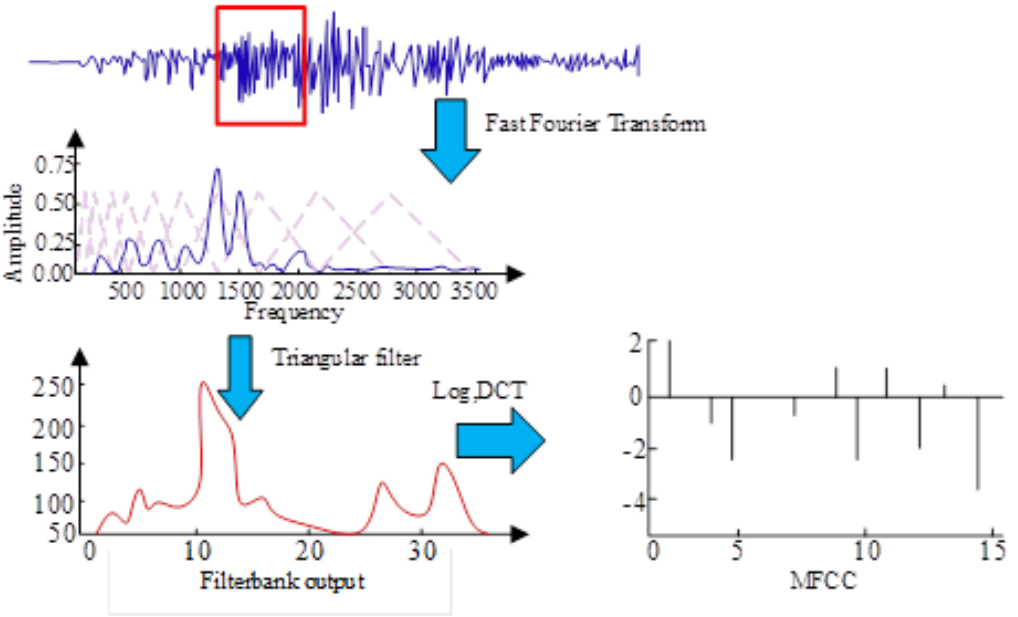
$$\begin{cases} w_{ij}^{l} = w_{ij}^{l} - \lambda \dfrac{\partial J\left(w,b\right)}{\partial w_{ij}^{l}} \\ b_{i}^{\ l} = b_{i}^{\ l} - \lambda \dfrac{\partial J\left(w,b\right)}{\partial b_{i}^{\ l}} \end{cases} \tag{7}$$

The speech feature parameter used in MLP is the Mel inverse spectral coefficient, which has more excellent robustness than the linear prediction parameter grounded on the vocal tract model. It is more consistent with the auditory features of the human ears, and still has greater recognition performance while reducing the signal-to-noise ratio. MFCC works by placing a set of band-pass filters in the frequency band in line with the size of the critical bandwidth to filter the input signal, and the output of each band-pass filter. The signal energy output from each bandpass filter is performed as the basic feature of the signal, and after processing it can be used as the input feature of speech. The MFCC working process is schematically shown in Figure 2 (Prawin 2021; Birch 2021).

To acquire each frame spectrum, the speech signal after framing and windowing is subjected to Fast Fourier Transform, the Fast Fourier Transform process is shown in equation (8). In equation (8) $y\left(i\right)$ represents the sound signal after pre-emphasis and N represents the FFT length.

$$\begin{cases} g\left(i\right) = y\left(i\right) w\left(i\right) \\ G\left(K\right) = \sum_{i=0}^{N\text{-}1} g\left(i\right) e^{-j\frac{2\pi}{N}mk}, 0 \le k \le N \end{cases} \tag{8}$$

Figure 2. Schematic diagram of MFCC working process



The energy spectrum is obtained by taking the square of the modulus of the spectrum after the Fourier process change, and the energy spectrum is smoothed by a set of triangular filter sets to highlight the resonant peaks of the original speech to eliminate harmonic effects, thus avoiding the impact of different speech pitches on audio discrimination and reducing the number of operations. After triangular filtering, the output logarithmic energy $s\left(u\right)$, the calculation process is shown in equation (9). u is in one of the triangular filters, $H_u\left(k\right)$ refers to the frequency response of it.

$$s\left(u\right)=\ln\left(\sum_{k=0}^{N-1}\left|G\left(k\right)\right|^2 H_u\left(k\right)\right) \tag{9}$$

Discrete cosine transform is used to separate the redundant data and the Discrete cosine transform is calculated to obtain the MFCC coefficients, see equation (10). *M* in Eq. (10) means the number of triangular filters. n is the order of the MFCC coefficients, which are typically taken as 12-16.

$$C\left(n\right)=\sum_{u=0}^{U-1}s\left(u\right)\cos\left(\frac{\pi n\left(u-0.5\right)}{M}\right) \tag{10}$$

Finally, the first and second order differentials of MFCC are further solved, which together form the Mayer inverse spectral coefficients. The differential spectrum can describe the dynamic characteristics of speech, and the differential formula is given in equation (11).

$$d_t = \begin{cases} C_{t+1} - C_t, t < K \\ \dfrac{\sum_{k=1}^{K} k\left(C_{t+k} - C_{t-k}\right)}{\sqrt{2\sum_{k=1}^{K} k^2}}, else \\ C_t - C_{t-1}, t \geq Q - K \end{cases} \tag{11}$$

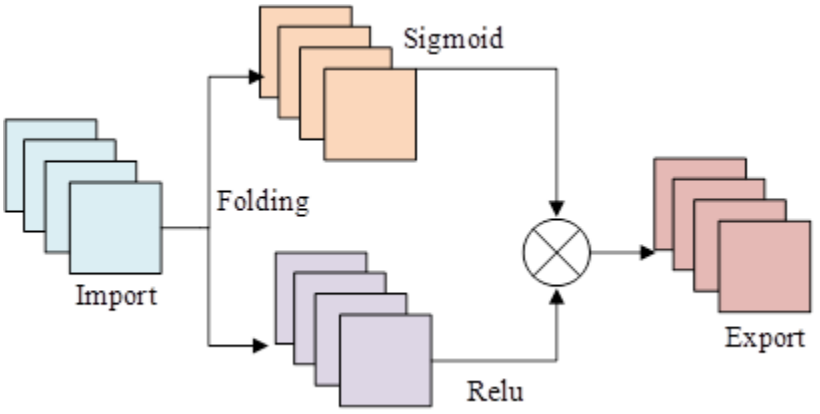
In Eq. (11), d_t is the first order difference of t. C_t is the corresponding inverse spectral coefficient. Q is the inverse spectral coefficient order. K is the time difference of the first order inverse coefficient, which is derived as 1 or 2; the first order difference is then substituted into Eq. (11) to acquire the 2nd order difference.

3.2 Design of Music Sound Recognition Model Fusing RNN and CNN

The audio signal of a music file is divided into many frames and contains strong local features between adjacent frames, so a CRNN is used for audio recognition analysis. The CNN contains an input, a convolutional, a pooling, an activation function and a fully connected layer. The convolutional operation of the convolutional layer calculates a linear response; the pooling layer pools the feature data after convolution, and the number of training parameters decreases significantly after down sampling; the activation function layer adds non-linear variation. The activation function layer increases the non-linear variation to update the generalization ability (Zeng et al. 2022). As the audio data of musical tones will be interspersed with different levels of noise, the noise will have an impact on the model recognition and detection. For ensuring the results' accuracy, the Attention Mechanism is combined with Multiple Scale Convolutional on the ground of the conventional CNN (Lw et al. 2020; You et al. 2022).

The Sigmoid activation function is first added to the original Relu activation function, and is seen as a gating unit, with the attention gating schematic listed in Figure 3. After the data has been output by the convolution layer, the two are multiplied element by element by the two activation functions respectively, and the calculation process is shown in equation (12). In equation (12), K and V are the weight parameters of the convolution kernel, b and c are the bias parameters, and I represents the input time-frequency map or feature map.

Figure 3. Schematic diagram of attention gating cycle



$$Y = \mathrm{Re}\,lu\left(I * K + b\right) \odot Sigmoid\left(I * V + c\right) \tag{12}$$

CNN need to make detection for each frame of audio, and the large size of pooling layers in traditional CNN is not conducive to the accuracy of audio recognition. Multi-scale convolutional fusion is performed based on the gating mechanism, using convolutional kernels of multiple sizes for convolutional operations, and then the feature maps of different convolutional parts are stitched together. 1×1, 3×3 and 5×5 convolutional kernels are used in this study.

RNNs are suitable for training samples whose input is a continuous sequence of varying lengths, such as a time-based sequence: a continuous segment of speech, a continuous segment of handwritten text. (Xing 2022). RNNs have a "memory function", in which the learned information is remembered and applied to the current learning computation during the training. The nodes between the hidden layers are interconnected to co-determine the output of the next moment. From Fig.4, compared with the fully connected network in Figure 1, the recurrent network has one more recurrent layer. w is the weight matrix between each time point, and the calculation process is shown in equation (13), where t and t-1 represent the time series, $s\left(t\right)$ is the memory of the sample at time t, $f\left(\ \right)$ refers to the activation function of $\tanh$, and $g\left(\ \right)$ means the softmax activation function.

$$\begin{cases} h\left(t\right) = Ux\left(t\text{-}1\right) + Ws\left(t\text{-}1\right) \\ s\left(t\right) = f\left(h\left(t\right)\right) \\ o\left(t\right) = g\left(Vs\left(t\right)\right) \end{cases} \tag{13}$$

Recurrent networks extend a long time, the network structure becomes increasingly complex and information is gradually lost during transmission, leading to network gradient explosion and gradient dispersion problems. To avoid this situation, the study uses Bidirectional LSTM (BILSTM) networks (Luo & Zhang 2022; Kota & Munisamy 2022). The LSTM network contains four internal network layers, and the LSTM module contains three gates to control cell state changes, i.e., the forgetting, the input and the output gate. The BILSTM network is an optimization of the LSTM with the addition of a back-propagation LSTM module, which allows the data to propagate in both directions in the BILSTM network, solving the biased nature of the LSTM, **as** shown in Figure 5(Zhang C 2022;Geng B 2022).

Figure 4. Schematic diagram of circular network

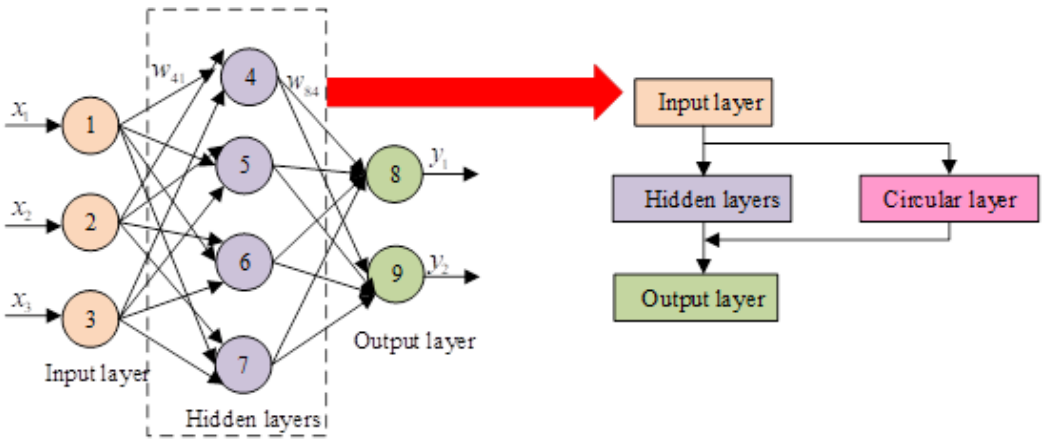
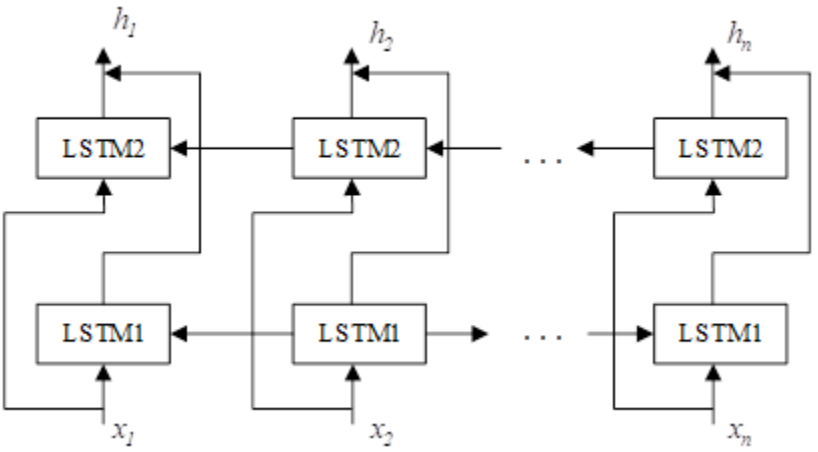


Figure 5. Schematic diagram of BILSTM network results



Different voices in the music have different characteristics, and using a fixed frame length will ignore these characteristics and affect the recognition results. Therefore, in the recurrent network part, the CNN's output is segmented using the multi-sliding window framing method, and the corresponding frame dimensions of the feature map of the CNN will be spliced according to the sliding window length to obtain advanced features with multiple frame length dimensions. The equation of Sliding Window is given in equation (14), where $X_{windowi}$ is the new feature obtained after the first sliding window of i, T is the sequence length, and $h_{i,j}$ is the output of BILSTM at . j

$$\begin{cases} H_{windowi} = BILSTM_i\left(X_{windowi}\right), i = 1,2,...,n \\ H_{windowi} = \left[h_{i,1}, h_{i,2}, ..., h_{i,T}\right] \end{cases} \tag{14}$$

Finally, the output of the RNN is passed through the Feedforward Neural Network to obtain the detection results, and then the detection results of each frame are weighted for importance recognition, and the final recognition results are obtained using the activation function Sigmoid and the activation function Softmax, respectively, as shown in equation (15). The output of the Sigmoid activation function is the probability that the frame belongs to each category, while the output of the Softmax activation function is the importance factor of the frame belonging to each category.

$$\begin{cases} p_t = FNN - Sigmoid(h_t) \\ z_t = FNN - Soft\max(h_t) \\ O_t^{'} = p_t \odot z_t \\ O^{''} = \dfrac{\sum_{t=1}^{T} O_t^{'}}{\sum_{t=1}^{T} z_t} \end{cases} \tag{15}$$

4. PERFORMANCE TEST OF MUSICAL SOUND RECOGNITION MODEL WITH IMPROVED RNN

4.1 Design of Test Experimental Protocol and Analysis of Model Parameters

A test experiment was designed to validate the performance of the musical tone recognition model constructed for the study. $240 \times 64 \times 1$ The data used in the experiment came from student performance audio files provided by an online vocal teaching course in China, with a total of 3568 audio files. 70% of the sample data set was used as training set samples and 30% as test set samples. The parameters of the CNN part of the model are shown in Table 1. The output part of the CNN uses three sliding windows of size 1, 2 and 4; the RNN part uses three BILSTM networks with input sequences of length 360, 240 and 120, with 256 neurons in the input and output layers, 128 hidden layers and 1 hidden layer; the Feedforward Neural Network part contains 240 fully connected layers, each with 17 neurons. The algorithm was trained for 50 rounds, with 100 iterations per round and a learning rate of 0.001. The loss function used was binary cross-entropy.

This research firstly examined the effect of the fast Fourier variation length on the performance of the algorithm model, the length settings were selected as multi-time-length overlay, single length N=1024, 2048, 3072 and 4096, and the algorithm accuracy was used as the evaluation metric for learning training, the results are Figure 6. From Figure 6, the horizontal axis means the number of times the algorithm worked on the entire training dataset, and the vertical axis is the recognition accuracy rate. As the iteration amount rises, the accuracy graphs for different Fast Fourier Transform lengths show an increasing trend, and the model converges roughly at around 40 iterations. The accuracy of the Fourier process curve is the highest when the Fourier length is 1024, and the accuracy of the model converges to 0.74. When the Fourier length is growing to rising, the curve level gradually declining and the accuracy rate becomes lower and lower, and when N=4039, the accuracy rate decreases to 0.51. The fast Fourier process with multi-length superposition is the most favorable for the model performance. was the most favorable for model performance, so subsequent experiments used this as a parameter for model training and comparison.

Continuing to examine the effect of different MFCC dimension lengths on the accuracy of the network model, the dimension lengths were chosen as 20, 40, 60 and 80, and the training outcomes are exhibited in Figure 7. As for Figure 7, the accuracy of the curves at different dimensional lengths roughly tends to rise as the iteration number increases, with the accuracy curve at dimensional length 60 falling at 30 iterations before rising again. When the dimension length is in the range of 20-40, the accuracy of the model grows as the dimension grows, with the highest accuracy rate reaching 0.89 at dimension length 40; however, when the dimension length continues to grow to 60, the accuracy rate decreases as the dimension length grows. The length of the dimension is the length of the feature, which is related to the sum of training parameters. The slope of the rising accuracy curve is larger when the dimension is smaller, and the model is faster to train.

Table 1. Parameter settings of CNN part

Layer	Unit1	Unit2	Unit3	Unit4
Convolutional kernel 1	$3 \times 3 \times 64$	$3 \times 3 \times 128$	$3 \times 3 \times 128$	$3 \times 3 \times 256$
Convolutional kernel 2	$1 \times 1 \times 64$	$1 \times 1 \times 128$	$1 \times 1 \times 128$	$1 \times 1 \times 256$
Convolutional kernel 3	$5 \times 5 \times 64$	$5 \times 5 \times 128$	$5 \times 5 \times 128$	$5 \times 5 \times 256$
Pooling layer	2×1	2×1	2×1	4×1

Figure 6. Effect of different FFT lengths on model accuracy

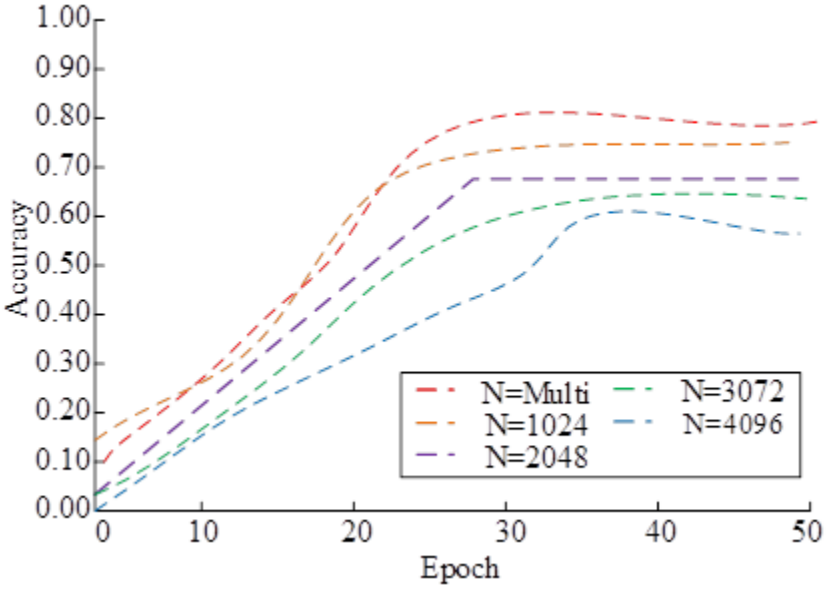
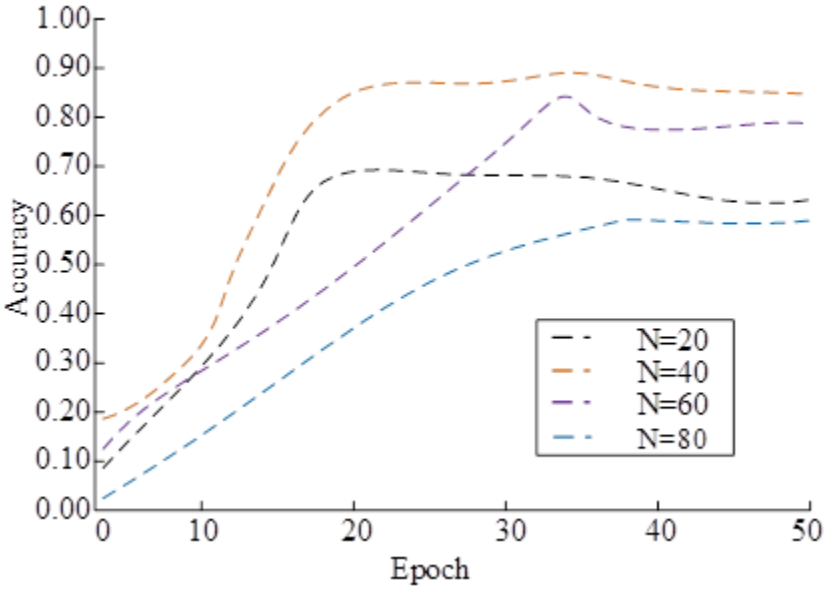


Figure 7. Effect of different MFCC dimensions on model accuracy



4.2 Analysis of the Quality of the Recommended Results

The model setup was carried out according to the optimal parameters in subsection 3.1, and the improved RNN model CRNN was compared with the traditional RNN algorithm model, the Gate Recurrent Unit (GRU), and the LSTM model using a one-way LSTM. The algorithm models are

analyzed separately for the recognition of pitch, time value, volume and timbre of musical tones, and the algorithm's loss value, accuracy, recall and F1 metrics are put into assessment indicators.

Firstly, the loss values on the test dataset after 50 iterations of different neural networks were compared. A total of 50 loss values were obtained during the learning process of the algorithm, and the loss curves are shown in Fig. 8. For Fig.8, the loss curve values of the CRNN model are much lower than those of the other three models, with steeper descent segments, indicating that the CRNN model can learn the category features of musical tones better on the data set samples, and the algorithm model learning ends when the iterations are completed. In contrast, the RNN model has the largest fluctuation in the loss curve, and at about 20 iterations, the loss curve shows a large upward and downward trend with a large oscillation, and the algorithm has difficulty in training; the smoothness of the loss curve of the LSTM and the GRU is not as good as that of the CRNN model.

The four different algorithms were used to analyses the pitch, time value, volume and timbre of musical tones, and the outcomes of the accuracy and recall experiments are listed in Figure 9. The GRU is the 2nd most accurate model, with accuracy values above 85%, and is more accurate in the recognition of time value and volume, which is not much different from the CRNN model. The LSTM model has the lowest recognition accuracy and outperforms the RNN model only in the recognition of the temporal value of music.

Figure 8. Iterative loss values for different algorithm models

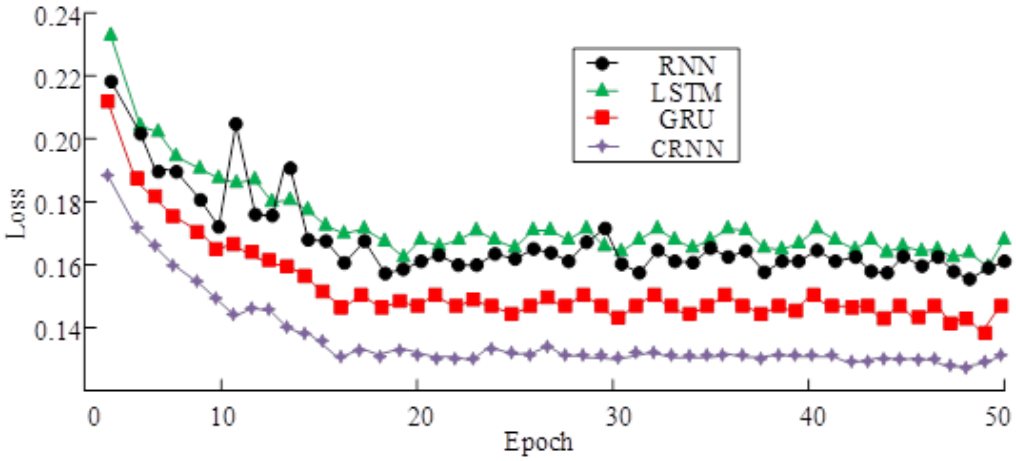
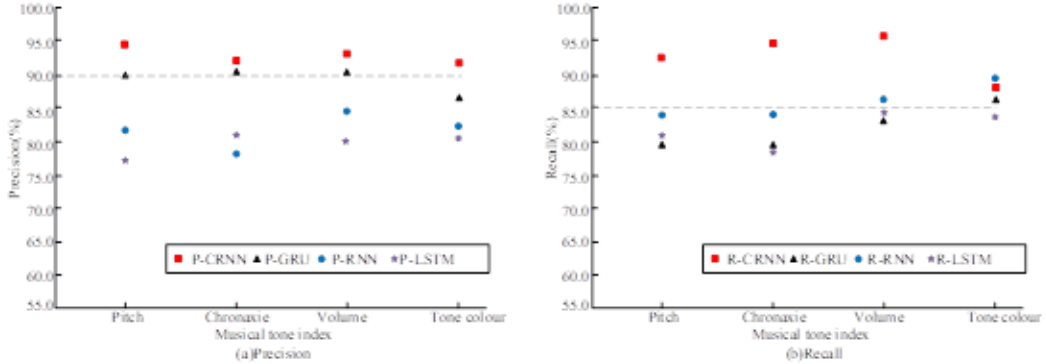


Figure 9. Accuracy and recall rates of music recognition for four networks



The precision and recall values are shown in Table 2, with Pre representing the precision and re the recall. From Table 2, the CRNN model has the highest pitch recognition accuracy value of 95.1%, while the GRU model has an accuracy value of 86.4%, which is 6.1 percentage points lower than the CRNN model. Although it is ideal to aim for high accuracy and recall, the actual situation is that high accuracy results in low recall and high recall results in low accuracy. In line with the results shown in Table 2, the GRU model has a higher precision rate than the RNN model and the LSTM model, while its recall value is significantly lower. However, the recall values of the CRNN model are still at the higher value level, indicating that the CRNN model does not improve the accuracy of the algorithm by sacrificing the recall, and maximisers the pair of contradictory indicators.

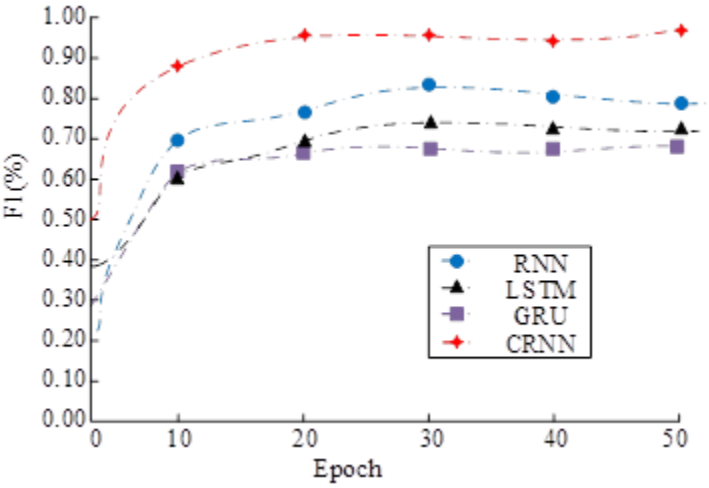
The F1 Score is the summed average of Precision and Recall, and can be performed to test the accuracy of a binary classification model, as neither precision nor recall can be used as an indicator of how good a model is. F1 considers both the accuracy and recall of a classification model, and its maximum and minimum value is 1 and 0. The training results of different algorithms for music sound recognition F1 values are shown in Figure 10. The F1 value of CRNN is the highest, with a maximum value of 94.6%; considering the accuracy and recall indexes, the performance of the CRNN is greater than the other three models, indicating that the model constructed in this study is more accurate for the recognition of different music tone indexes, and the model is more capable of learning and suitable for use in vocal music teaching courses.

The ROC is the ratio of the recall rate to the false positive rate, and the AUC is the area enclosed by the ROC curve and the coordinate axis. The area of the right-hand part of the ROC curve should

Table 2. Recognition accuracy and recall rate of different musical tone indicators

Model	CRNN		GRU		RNN		LSTM	
Evaluating indicator	Pre	Re	Pre	Re	Pre	Re	Pre	Re
Pitch	94.6%	93.2%	90.2%	78.6%	82.2%	84.5%	77.2%	80.3%
Chronaxie	92.4%	94.9%	91.6%	76.9%	77.6%	84.4%	79.1%	77.2%
Volume	93.5%	95.2%	91.9%	82.3%	84.6%	86.2%	77.2%	83.4%
Tone colour	92.5%	88.6%	86.4%	86.5%	83.4%	88.1%	77.1%	82.6%

Figure 10. F1 values for different models

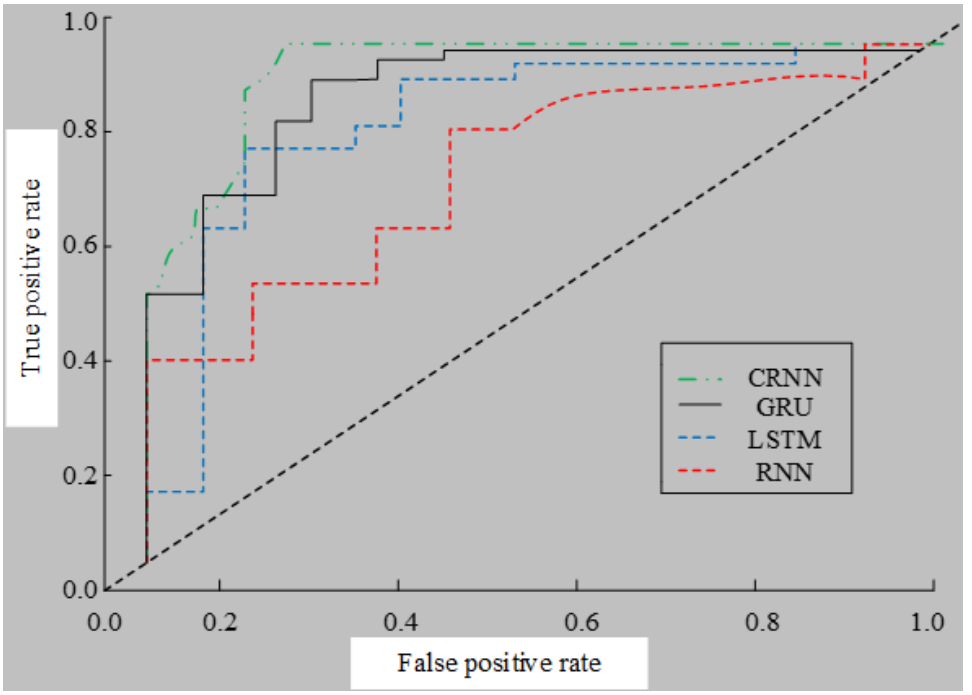


be roughly in the range of 0.5 to 1. The minimum AUC value is 0.5, and if it is less than the threshold value, the algorithm's detection is invalid. The AUC values of different combined models are shown in Figure 11. The ROC curve of the CRNN model is located at the top of the coordinate system, with an AUC value of 0.91; the ROC curves of GRU and LSTM are closer, with AUC values of 0.73 and 0.69 for the two models, respectively. In a comprehensive manner, the recognition model constructed by the study incorporating RNN and CNN has good performance in practical use.

5. CONCLUSION

In order to address the problem of low accuracy and efficiency of music tone analysis and recognition models in vocal music teaching online courses, this study designed a convolutional recurrent neural network model based on attention gating mechanism and multi-scale convolutional fusion improvement. The results of the performance test show that in order to improve the accuracy of the model, the optimal parameter setting scheme is to choose a fast Fourier process with multiple duration superposition and set the dimension length to 40, the accuracy rate is up to 0.89, the relationship between dimension length and accuracy rate is not a simple linear relationship. the loss curve value of CRNN is much lower than the other three models music sound recognition effect is better; CRNN accuracy rate value is the highest The values of the four music sound metrics were 94.6%, 92.4%, 93.5% and 92.5% respectively; the recall rate values of the CRNN model were also at the highest level, 93.2%, 94.9%, 95.2% and 88.6% respectively, while the F1 value of CRNN was as high as 94.6%, CRNN achieved the contradictory metrics of accuracy rate and recall rate The balance of multiple indicators of the model is optimised. The CNN part of the model is more efficient in extracting time-frequency features, and the RNN part takes into account the temporal correlation of features, making the model suitable for use in voice teaching courses. Traditional vocal teaching courses rely more on the accumulated teaching experience and teaching methods, and teach the basic principles

Figure 11. AUC values for different algorithms



of vocal music through the explanation of video and image materials, the teaching effect is far from the abstract and intuitive nature of vocal teaching courses. However, by combining CRNN with vocal teaching, students can have a more intuitive experience of the singing system and the teaching of vocal technique is more straightforward and clear. The design of the CRNN model promotes the cross-fertilisation of information technology and music art disciplines, opens up a wide range of applications for the development of computer music and contributes to the reform of the traditional vocal teaching classroom model. In the future vocal music teaching classroom, the music tone recognition model is the basis for realising human-computer interaction, which can help the computer to carry out the automatic generation of music scores, and can achieve real-time accompaniment and performance. However, this study did not design a study on the influence of the parameters of the network structure itself, such as the learning rate size, training steps, etc. The study is not comprehensive enough. In addition, the study did not investigate the identification of musical notes for different instruments, which limits the use of the model in different types of vocal performance situations.

FUNDINGS

The research is supported by: Hunan philosophy and Social Science Foundation Project “Research on Inheritance and innovation of Suining traditional music culture under the background of integrated development of culture and tourism”, Project No.: 18WTC28.

REFERENCES

Alper, Mehmet, & Arif. (2022). One-hour-ahead solar radiation forecasting by MLP, LSTM, and ANFIS approaches. *Meteorology and Atmospheric Physics, 135*(1), 946-958.

Birch, B., Griffiths, C. A., & Morgan, A. (2021). Environmental effects on reliability and accuracy of MFCC based voice recognition for industrial human-robot-. *Proceedings of the Institution of Mechanical Engineers. Part B, Journal of Engineering Manufacture*, *235*(12), 1939–1948. doi:10.1177/09544054211014492

Carlos, G. M., Antonio, R. V., & Jorge, C. Z. (2022). A holistic approach for image-to-graph: Application to optical music recognition. *International Journal on Document Analysis and Recognition*, *25*(4), 293–303. doi:10.1007/s10032-022-00417-4

Fu, L. (2021). Discussion on the Differences between Theory and Practice in Vocal Music Teaching. *Region - Educational Research Review*, *3*(1), 6–9.

Geng, B. (2022). Text segmentation for patent claim simplification via Bidirectional Long-Short Term Memory and Conditional Random Field. *Computational Intelligence*, *38*(1), 205–215. doi:10.1111/coin.12455

Isra, K., Muhammad, E. S., Ashhad, U., & Ullah, A. (2022). Rafi. An Intelligent Framework for Person Identification Using Voice Recognition and Audio Data Classification. *Applied Computer Systems*, *27*(2), 183–189. doi:10.2478/acss-2022-0019

Kota, V. R., & Munisamy, S. D. (2022). High accuracy offering attention mechanisms based deep learning approach using CNN/bi-LSTM for sentiment analysis. *International Journal of Intelligent Computing and Cybernetics*, *15*(1), 61–74. doi:10.1108/IJICC-06-2021-0109

Li, R., & Qin, Z. (2022). Audio recognition of Chinese traditional instruments based on machine learning. *Cognitive Computation and Systems*, *4*(2), 108–115. doi:10.1049/ccs2.12047

Li, S. (2021). A Probe into the Integration of Traditional Music Culture in Vocal Music Teaching in Colleges and Universities. region - Educational Research Region -. *Educational Research Review*, *3*(2), 65–69.

Luo, J., & Zhang, X. (2022). Convolutional neural network based on attention mechanism and Bi-LSTM for bearing remaining life prediction. applied Intelligence. *The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *52*(1), 1076–1091.

Lw, A., Xs, A., Min, X. A., Jia, L.A., & B, Y.X. (2020). Portfolio trading system of digital currencies: a deep reinforcement learning with multidimensional attention gating mechanism - ScienceDirect. *Neurocomputing, 402*, 171-182.

Maghraby, E., Gody, A. M., & Farouk, M. H. (2021). Audio-Visual Speech Recognition Using LSTM and CNN. *Recent Advances in Computer Science and Communications, 14*(6), 2023-2039.

Nan. (2022). Study on the Application of Improved Audio Recognition Technology Based on Deep Learning in Vocal Music Teaching. *Mathematical Problems in Engineering, 43*, 897-909.

Phan, T. T. H., Dong, N. D., Du, N. H., Van Hanh, N., & Thai, P. H. (2022). Investigation on new Mel frequency cepstral coefficients features and hyper-parameters tuning technique for bee sound recognition. *Soft Computing*, *27*(9), 5873–5892. doi:10.1007/s00500-022-07596-6

Prawin, J. (2021). Breathing crack damage diagnostic strategy using improved MFCC features. *Journal of Intelligent Material Systems and Structures*, *32*(20), 2437–2462. doi:10.1177/1045389X211001446

Qiu, Z., Wang, H., Caibo, L., Lu, Z., & Kuang, Y. (2022). Sound Recognition of Harmful Bird Species Related to Power Grid Faults Based on VGGish Transfer Learning. *Journal of Electrical Engineering & Technology*, *18*(3), 2447–2456. doi:10.1007/s42835-022-01284-z

Sumarno, L., & Chai, R. (2021). DCT based feature extraction and support vector machine classification for musical instruments tone recognition. *Institute of Advanced Engineering and Science, 2021*(10), 2796-2803.

Tanaka, K., Nishikimi, R., Bando, Y., Yoshii, K., & Morishima, S. (2021). Pitch-timbre disentanglement of musical instrument sounds based on VAE-based metric learning. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 111-115.

Wang, T. (2022). Neural Network-Based Dynamic Segmentation and Weighted Integrated Matching of Cross-Media Piano Performance Audio Recognition and Retrieval Algorithm. *Computational Intelligence and Neuroscience*, *13*(May), 187–201. doi:10.1155/2022/9323646 PMID:35602641

Wu, Z., Jiang, J., Xiao, Z., & Huang, M. (2022). Audio-based expansion learning for aerial target recognition. *Applied Acoustics*, *188*(Jan), 2650–2658. doi:10.1016/j.apacoust.2021.108551

Wu, Wan, Ge, & Pan. (2022). Car engine sounds recognition based on deformable feature map residual network. *Scientific Reports, 12*(1), 2744-2756.

Xiao, H., Liu, D., Kai, C., & Mi, Z. (2022). AMResNet: An automatic recognition model of bird sounds in real environment. *Applied Acoustics*, *201*(Dec), 3218–3232. doi:10.1016/j.apacoust.2022.109121

Xing, B., Xu, E., Wei, J., & Meng, Y. (2022). Recurrent neural network non-singular terminal sliding mode control for path following of autonomous ground vehicles with parametric uncertainties. *IET Intelligent Transport Systems*, *16*(5), 616–629. doi:10.1049/itr2.12161

Xu, W., Hu, Y., & Li, J. (2022). A data-driven Dir-MUSIC method based on the MLP model. *IET Science, Measurement & Technology*, *16*(6), 367–376. doi:10.1049/smt2.12110

You, G. R. (2022). Enhancing ensemble diversity based on multiscale dilated convolution in image classification. *Information Sciences*. *International Journal (Toronto, Ont.)*, *606*(2), 292–312.

Yue, C., Gao, Y., & Yi, X. (2022). Computational Modelling of Tone Perception Based on Direct Processing of f0 Contours. *Brain Sciences*, *12*(3), 337–349. doi:10.3390/brainsci12030337 PMID:35326294

Yun, C., & Fu, W. (2022). Research on Audio Recognition Based on the Deep Neural Network in Music Teaching. *Computational Intelligence and Neuroscience*, *27*(May), 1782–1796.

Zeng, Z., Sun, S., Sun, J., Yin, J., & Shen, Y. (2022). Constructing a mobile visual search framework for Dunhuang mu rals based on fine-tuned CNN and ontology semantic distance. distance. The Electronic Library. *The International Journal for Minicomputer, Microcomputer, and Software Applications in Libraries*, *40*(3), 121–139.

Zhang, C., Wang, W. Z., Zhang, C., Fan, B., Wang, J. G., Gu, F., & Yu, X. (2022). Extraction of local and global features by a convolutional neural network–long short-term memory network for diagnosing bearing faults. *Proceedings of the Institution of Mechanical Engineers. Part C, Journal of Mechanical Engineering Science*, *236*(3), 1877–1887. doi:10.1177/09544062211016505