ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

The mechanism of the choral conductor's gestural language in the emotional communication of the work

Shuai Wang[1,*]

[1] School of Art, North University of China, Taiyuan, Shanxi, 030051, China

SUMMARY: *Choral conductor gesture is the core medium for conveying the emotion of a musical work, but its intrinsic mechanism of action has long remained in subjective descriptions. This paper reveals the role of choral conductor gesture language in the emotional communication of musical works through a data-driven approach. The article first optimizes the 3DCNN network structure for the characteristics of choral conductor gesture data signals, designs a spatio-temporal separation convolutional neural network, introduces a spatio-temporal channel adjustment factor, and quickly changes the model size, obtaining a classification effect with low computational cost and high recognition accuracy. The empirical results show that the accuracy of the model in choral conductor gesture and emotion recognition on different datasets reaches more than 85%, and the correlation analysis of the gesture and emotion recognition results in the multimodal data of "Sunset xiao drums" performed by three conductors reveals the mechanism of the role of the gesture and emotion, and the conductor realizes the emotion of the performer through the amplitude of the gesture, the speed of the gesture and other spatial dynamics features interaction. The results of this paper provide a theoretical and technical basis for the teaching, evaluation and intelligent assistance of music conducting art.*

KEYWORDS: *spatio-temporal channel regulator; spatio-temporal convolutional neural network; choral conductor's gesture; emotional communication*

1 Introduction

The primary responsibility of the choral conductor is to lead the choir in the choral process to the audience to convey the emotion of the work, the whole process is composed of two important links, one is the rehearsal link, in the rehearsal stage, the choral conductor needs to be communicated through the conductor gestures, verbal descriptions, verbal prompts, and other aspects of the comprehensive communication to realize the communication and interaction with the members of the group, and then to cultivate a kind of tacit understanding [1-3]. Another link for the stage performance link, this link no longer has a verbal expression, by the gesture command for all the command means, the conductor through gestures to prompt the members, for the audience to bring with rich emotion, beautiful and beautiful music [4]. Jansson, D et al. [5] conducted a survey of 294 choral conductors in Norway, and constructed a three-tiered model consisting of 17 elements of competence, for the understanding of their perceptions and priorities, and their own perceived level of proficiency was rated. In actual band conducting work, the conductor exists as the core of the band, and needs to organize, plan, train, and arrange repertoire for the choral band, etc. In the process of choral team practice and

*18335163369@163.com
https://doi.org/10.65102/is2026231 Published: 30 April, 2026

performance, the conductor needs to have strong organizational skills, interpersonal skills, and coordination skills, as well as excellent music fundamentals, music analysis skills, and music innovation skills [6].

Many amateur conductors in the conductor and conductor gestures in the essence of understanding there is a certain bias, resulting in weekdays lack of corresponding conductor basic training and aesthetic interest training, resulting in actual choral performances in the conductor without feelings, gestures waving and music to convey the emotions do not accord with such phenomena is not conducive to the development of choral orchestra and the audience's emotional conveyance [7, 8]. Napoles, J [9] a study on the impact of the perception of expressive choral performance, pointed out that there are significant differences between different presentation and performance styles, and in all modes of performance, the audience tends to think that performances with rich physical and gestural movements are better than those with stereotyped movements. Many choral conductors stay in this superficial stage and are satisfied with the status quo, and their conducting skills cannot be improved for a long time, which affects the quality of the chorus and the audience's listening experience [10]. In the performance process often see many conductors on stage waving command gestures, but with the overall choral performance atmosphere is out of place, this phenomenon is obviously due to the existence of the conductor of the concept of command understanding bias, the conductor is not skilled enough to lead to [11].

Poor quality conductor often brings confusion for chorus members in the performance, do not know how to correctly grasp the beat and the work display, but also is not conducive to the audience's strong cultivation of music aesthetic art appreciation, but also will lead to the core idea of the musical work of the emotion conveyed is not deep enough [12]. For example, Jansson, D et al [13] identified four contextual dimensions, namely, the complexity of the music, the uniqueness of the movement, the acceptance of the singers, and the proficiency of the conductor, through in-depth interviews with 40 choir members, and conductors in Norway and Sweden. In addition, Kumar, A and Morrison, S [14] selected two music clips and paired them with videos of two conductors using specific gestures to explore whether the conductor's gestural movements affect the audience's level of attention to the orchestra's performance, and the study found that the conductor's gestural language directs the audience's attention, as well as a deeper interpretation of the work. Conducting skills are therefore highly specialized and diverse, requiring choral conductors to develop appropriate skill innovations based on musical connotations in practice [15].

Other scholars believe that the emotional experience and emotional communication determines the conductor's conductor behavior, conductor behavior only in line with the musical situation and the core connotation of the music to produce visual aesthetics, and blind imitation can only be used to achieve the artistic aesthetics of the art of the achievement is impossible to talk about [16, 17]. Kumar, A and Morrison, S [18] selected two music clips and paired them with a video of two conductors using specific gestures to explore whether the conductor's gestural movements affect the audience's level of attention to the orchestra's performance, and the study found that the conductor's gestural language guides the audience's attention as well as a deeper interpretation of the work. Poggi, I and Ansani, A [19] identified 23 types of gestures and 11 parameter values for intensity indication for choral conducting gesture language, such as "forte", "asymptote" and "diminutive", etc. Additionally, 21 types of gestures performed by hands, arms and shoulders related to intensity were determined through five symbolic mechanisms. Platte, S et al [20] A controlled study of 18 healthy choral singers (9 males and 9 females) which found significant abdominal volume differences between two different gestures suggests that the conductor's choice of gesture affects the respiratory behavior and tonal quality of the singers, challenging the existing view of gestures as equivalent.

Conducting gestures in the rehearsal and performance stage of the information conveyance, both professional and popular characteristics, is the common possession of the artistic language treasure, choral conductor through the musical works of perception of the conductor performance, constitutes a personalized behavioral art [21]. Cottrell, S [22] explored the role of conductor gestures in orchestral performances through a qualitative study, where they argued that the role of the conductor is ambiguous, the musicians have autonomy, and the performance relies on co-participation and waves of emotions conveyed, emphasizing the symbolic significance of the conductor's gestures and the important role they play in controlling time. Different conductor gestures contain a sense of musical art, an artistic communication of the work, and different demands on the sound [23]. Confident, concise, easy to understand, strong and not lacking in artistic expression of the gesture language can give chorus members in the singing of the possession of confidence and mastery of the performance of the rhythm, in the music performance, the voice of the guidance has been impossible to appear, only through the conductor gestures to be able to the emotional information correctly and efficiently conveyed [24].

The purpose of this study is to construct and validate a multilevel gesture-music-emotion model through computational empirical mathematical methods, and systematically investigate the role of choral conductor's gesture language in conveying the emotion of a work. A multi-task classification model based on spatio-temporal convolutional neural network is proposed. The spatio-temporal channel regulator introduced in this model can efficiently and accurately process the spatio-temporal dimensional features and emotional features of the original data at the same time, which improves the efficiency of multimodal feature extraction. In order to explore the specific mechanism of action, an experimental scheme of multimodal data recognition classification and correlation analysis is designed, and the multimodal data of the three conductors' interpretation of “Sunset Drumming” is collected, which provides a solid data foundation for the subsequent mechanism validation.

2 The role and design principles of choral conductor's gesture language

2.1 The Meaning and Role of Sign Language for Choral Conductors

The conductor is the creator of the collective singing art. Usually, the chorus team consists of a large number of members together, different members of the art of understanding and expression of different members, to some extent added to the diversity of the choral process, but still the pursuit of the overall sound coordination and harmony as the core, a higher standard of strict timbre of each voice, as far as possible, the volume control in a relatively balanced state, and fully embodies the sense of hierarchy between the various parts of the voice.

2.2 Basic techniques of command gestures

(1) Gesture expressing flexible rhythm - “point”

Short and flexible movements are two of its more notable features, and it is mostly used in the conductor of the introductory part of the entire musical work. The elastic “point” has certain techniques in its application, such as using the wrist movement as the axis, with other parts of the body actively cooperating, and the whole process is always in a relatively relaxed state. When the emotion in the musical work is stronger, the expression, strength and wrist movement change, but the final presentation of the musical performance is relatively consistent in effect.

(2) The gesture of expressing coherent lyricism - “line”

This gesture is mainly used in works with large emotional ups and downs, and is characterized by a strong and bold rhythm. The small climax and big climax part of the work is mostly used in this gesture, using the big arm to drive the small arm to the hand, the overall swing, so that the line movement is full of rich emotions, so the distinction between its requirements are more detailed.

2.3 Gesture Design Principles

Principles of gesture movement: First, accuracy. From the perspective of the chorus conductor, its main responsibility is to guide all chorus members, auxiliary personnel scientific and reasonable judgment of the direction of the conductor gestures, in the full expression of the linear emotion in the musical works at the same time, the basic knowledge of the conductor more clearly passed to the members, through the accurate expression of the movement and emotion can help members of the overall improvement of the musical expression. Second, standardization. Conductor in chorus rehearsal or performance, need to strictly follow the relevant normative gesture action standard or action illustration for the conductor, to maximize the avoidance of swinging blindness, once the scene command confusion, will affect the normal performance of the entire choir. Third, naturalness. The conductor is like the brain, all the musicians must follow its state of play, to always maintain a good mental outlook state, to maintain a proper posture language posture, to members of the natural energy, in order to ease the overcoming of pressure and tension, all the energy to focus on the chorus performance. The main characteristics of gesture movement: the so-called “gesture movement” is essentially a kind of basic graphic line trajectory, usually, the pattern consists of three parts, that is, the beat line, beat point, and the line of reflection. In the process of shaping and presenting the image of a musical work, the choir's main source of expression is based on the points and lines of the conductor's gestures, and through the effective fusion of the two, the choir is able to construct a very colorful musical melody.

3 Gesture and emotion recognition based on spatio-temporally separated convolutional neural networks

3.1 Convolutional Neural Networks

The main structure of a convolutional neural network contains an input layer, a convolutional layer, a pooling layer, a fully connected layer, and a classification layer.The LeNet-5 network model is a classical model of convolutional neural networks, which successfully implements the task of handwritten digit recognition. The relative simplicity of its network structure makes it an introductory algorithm for deep learning techniques. The structural diagram model of the convolutional neural network is shown in Figure 1.

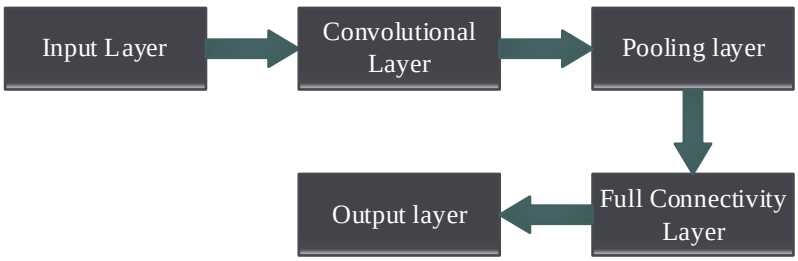

Figure 1: The basic structure of the convolution neural network

3.1.1 Convolutional layers

The function of the convolutional layer is to obtain a new feature image by the operation of convolution of the convolution kernel with the feature image, and to repeat the action on the image in order by several convolution kernels on the regions at different positions, i.e., the sensory field. And the result generated after the convolution operation is combined into the feature image. The convolutional layer in front of the convolutional neural network structure mainly extracts low-level image features, while the convolutional layer in the back extracts more complex image features. In fact the convolution operation is a linear operation and its main mathematical expression is shown in equation (1).

$$y_{conv} = f\left(\sum_{j=0}^{J=1}\sum_{i=0}^{I=1} x_{m+i,n+j} * w_{ij} + b\right), 0 \le m \le M, 0 \le n \le N \tag{1}$$

In the formula x is a 2D vector with acceptance region size $M * N$; w is a convolution kernel with length j and width i; b is the bias applied to the output feature mapping, which is 1 by default; y_{conv} is the output of the convolution operation; and y_{conv} is the length and width, respectively, of the length and width of the 2D vector of the image, respectively. f is the activation function in the neural network.

According to the needs of different projects and scenarios, the relevant parameters of the convolutional layer in the convolutional neural network will be changed accordingly. The parameters that affect the properties of the convolutional layer are mainly the size of the convolutional kernel, the move step, and the filling method after the convolutional operation. And these parameters have a very important effect on the convolution operation. The size of the convolution kernel determines the size of each feature map, which can be any size. Commonly used convolution kernel sizes are 1*1, 3*3, 5*5, or larger. Let the step size be S, and the length and width of the convolution kernel be filter_height and filter_width, respectively, then the expression for the size of the new output feature map after each convolution is shown in Equation (2).

$$\begin{aligned} &\text{new\_height=(input\_height-filter\_height)/S+1} \\ &\text{new\_width=(input\_width-filter\_width)/S+1} \end{aligned} \tag{2}$$

3.1.2 Pooling layer

The main role of the pooling layer is mainly to reduce the size of the feature map for selecting relevant features and information filtering operations for the output after convolution. The size of the newly generated feature map after the pooling operation of the image will be significantly reduced. And after many pooling operations, the number of parameters of the whole network will also be significantly reduced, through the pooling method can reduce the amount of computation of convolutional neural network and improve the efficiency of network training. According to the division of pooling operation, the commonly used pooling methods are average pooling and maximum pooling.

The principle of average pooling is to take the average of the relevant values in the filter neighborhood to get a new value as the result of the pooling operation; the principle of maximum pooling is to take the maximum of the relevant values in the filter neighborhood as the result of a pooling operation. Pooling operation also brings some problems. First, the feature map size is significantly reduced after pooling operation, which will inevitably lose part of the original feature map information and affect the accuracy of network training. Second, the size

of the filter neighborhood is limited, which can lead to large variance of the estimation results.

One of the pooling methods applied by most of the convolutional neural networks is maximum pooling. Maximum pooling preserves the most important feature map information in the filter neighborhood. Take the maximum pooling method with a scale of 2 and a step size of 2 as an example. Its mathematical expression is shown in equation (3).

$$f_{pool} = Max\left(x_{m,n}, x_{m+1,n}, x_{m,n+1}, x_{m+1,n+1}\right) \tag{3}$$

where f_{pool} denotes the output after maximal pooling.

3.1.3 Full connectivity layer

The image is formed into a set of feature vectors after several convolution and pooling operations of the neural network, and the feature images are next classified by the fully connected layer and the classification layer. The fully connected layer generally appears at the end of the neural network after successive convolutional pooling operations and is in fully connected form to receive neuron signals from the upper layer. Each of its nodes is connected to all the nodes of the previous layer. Generally the fully connected layer will be used as a classifier of the network for deep feature classification and finally passed to the next layer through the activation function to continue the computation. But since the fully connected layer has to be connected in the form of fully connected, the number of parameters involved in this layer is more. In convolutional neural network model, most of the parameters are concentrated in the fully connected layer. Therefore the network training takes longer time. How to reduce the parameters in the fully connected layer without affecting the network training effect as much as possible is a big idea for network structure improvement.

3.1.4 Output layer

In a convolutional neural network structure, the output layer typically uses a loss function to estimate the degree of deviation between the predicted and true values of the network model. The loss function is a quantitative function that measures the magnitude of the network's prediction error from statistical methods. In general, the value of the loss function is closely related to the robustness of the network; the smaller the value of the loss, the smaller the robustness. We hope that a network model after training the lower the loss value the better, the higher the prediction accuracy of the neural network loss value can tend to zero. There are several common loss function models.

(1) Logarithmic loss function

The logarithmic loss function, also known as the log-likelihood loss, is a function based on probability estimation that evaluates the error between the output value of the classifier and the true value. The mathematical expression of the log loss function is shown in equation (4).

$$L(Y, P(Y \mid X)) = -\log P(Y \mid X) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log\left(p_{ij}\right) \tag{4}$$

where Y is the output variable, X is the input variable, and L is the loss function. The number of samples in the input is N and the number of categories in the image is M. The y_{ij} on the right side of the formula represents an indicator of whether the category j is a true category of the input instance. The probability that the classifier predicts that the input instance x_i belongs to the category j is denoted by p_{ij}.

However, the logarithmic loss function can only classify instances of two categories, and is not applicable to multi-classification scenarios. And most of the image recognition and classification involves more than three categories of classification tasks. Therefore, in practical applications SoftMax function is widely used in image classification tasks.

(2) Softmax Classifier

Softmax classifier plays the role of multi-classification in neural networks, maps the output values in the (0,1) interval, and is a function based on the Logistic function for promotion. Since Logistic function can only perform binary classification task, Softmax classifier is often used in multiclassification task and its mathematical expression is shown in equation (5).

$$SoftMax\left(z_i\right)=\frac{e^{z_i}}{\sum_{c=1}^{C}e^{z_c}} \tag{5}$$

where is the output value of the ith node, and C is the number of output nodes, i.e., the number of categories for classification. In contrast to Hardmax, Softmax determines which classification the instance is most likely to belong to by generating the probability of each classification result as the next prediction target.

3.2 Classification model based on spatio-temporally separated convolutional neural network

3.2.1 Space-time separated convolutional neural networks

Choral gesture language data is the time series information collected by sensor arrays that have relative positional connections in space, and its data dimension and spatio-temporal correlation are consistent with video data, so for feature extraction of choral gesture language data, it is most appropriate to use 3DCNN. However, in 3DCNN, the convolution process operates on the local temporal and spatial dimensions, which increases the difficulty of optimization during network design. In contrast, the convolution module in SSCNN splits the convolution operation in 3DCNN and decomposes the convolution process into two separate steps, which approximates the process of 3D convolution by performing 2D convolution followed by 1D convolution, and is thus called the (2+1)D module. Compared to 3DCNN, the (2+1)D module adds an extra nonlinear layer in the middle of spatial convolution and temporal convolution, thus making it possible to represent models with more complex functions for the same network parameters; in addition, the (2+1)D module is able to converge faster during the training process, yielding lower training losses. Therefore, it is more suitable to use the (2+1)D module to extract the features of choral gesture language data.

The input data format is $\left[N_{l-1},H_{l-1},W_{l-1},D_{l-1}\right]$, $N_{(l-1)m}$ represents the number of intermediate layer channels in the $l-1$th layer, and N_{l-1} and N_l represent the numbers of channels in the input and output layers, respectively. The number of 2D convolutional kernels and 1D convolutional kernels is determined by the number of channels in the input layer, intermediate layer and output layer. Then the 3D convolutional kernel w can be decomposed as:

$$w=t\otimes v \tag{6}$$

where $t\in\mathbb{R}^{1\times1\times u}$, denotes the 1D convolution kernel, $v\in\mathbb{R}^{s\times q\times1}$ denotes the 2D convolution kernel, and $\otimes$ denotes the Kronecker product, introduces the following formula for the

(2+1)D convolution:

$$L_{l,j}^{x,y,z} = f\left\{\sum_{m=0}^{N_{(l-1)m}}\sum_{u=0}^{U_{l-1}-1} t_{(l-1),m,j}^{u} f\left(\sum_{i=0}^{N_{l-1}}\sum_{s=0}^{S_{l-1}}\sum_{q=0}^{Q_{l-1}-1} v_{(l-1),i,m}^{s,q} l_{(l-1),i}^{(x+s),(y+q),z} + b_{(l-1),i,m}\right) + c_{(l-1),m,j}\right\} \tag{7}$$

Let $h_{(l-1),i,m}^{x,y,z} = f\left(\sum_{i=0}^{N_{l-1}}\sum_{s=0}^{S_{l-1}-1}\sum_{q=0}^{Q_{l-1}-1} v_{(l-1),i,m}^{s,q} l_{(l-1),i}^{(x+s),(y+q),z} + b_{(l-1),i,m}\right)$, then the (2+1)D convolution process is expressed as:

$$L_{l,j}^{x,y,z} = f\left\{\sum_{m=0}^{N_{(l-1)m}}\sum_{u=0}^{U_{(l-1)}-1} t_{(l-1),m,j}^{u} h_{(l-1),i,m}^{x,y,z} + c_{(l-1),m,j}\right\} \tag{8}$$

where $h_{(l-1),im}^{x,y,z}$ represents the feature map tensor of the intermediate layer after 2D convolution of the input layer, $v_{(l-1),i,m}^{s,q}$ represents the 2D convolution kernel between the $l-1$th layer and the lth layer; $[s,q]$ denotes the coordinates of the convolution kernel in both directions, S_{l-1}, Q_{l-1} represent the size of the convolution kernel; $t_{(l-1),m,j}^{u}$ is the 1D convolution kernel, u are the coordinates in the direction of the length of the 1D convolution kernel, and U_{l-1} represents the length of the convolution kernel; $b_{(l-1),i,m}$ and $c_{(l-1),m,j}$ are the bias values of the 2D and 1D convolution parts, respectively.

(The (2+1)D module extracts features by convolution in the temporal and spatial dimensions, respectively. The activation function added between the temporal and spatial convolution operations enhances the nonlinear expressiveness of the whole network, allowing the network to converge faster during training.

3.2.2 Spatio-temporal channel modifiers

Since the 3D convolution is decomposed into spatial and temporal convolution, the number of channels output from the spatial convolution layer needs to be given a specific value. For this reason, the spatio-temporal channel adjustment factor is proposed, and by changing the spatio-temporal channel adjustment factor, the number of intermediate layer feature channels of each (2+1)D module can be controlled, which in turn reduces the computational and parametric quantities of the model while ensuring the quality of feature extraction.

Assuming a 3D convolutional kernel $\mathcal{W} \in \mathbb{R}^{d\times d\times t}$, the parametric size of the 3D convolutional kernel is $N_{(l-1)} \times d \times d \times t \times N_l$. where N_l denotes the number of channels in the lth layer, d denotes the spatial dimension size of the 2D convolutional kernel, and t denotes the 1D convolutional kernel time dimension size. By spatio-temporal decomposition of the convolution kernel, the spatial convolution kernel dimension is obtained as $\mathcal{W}_s \in \mathbb{R}^{d\times d}$, and the temporal convolution kernel dimension is obtained as $\mathcal{W}_T \in \mathbb{R}^{t}$. The final parametric size of the (2+1)D convolution kernel is calculated as $N_{(l-1)} \times d \times d \times N_{(l-1)m} + N_{(l-1)m} \times N_l \times t$. Where $N_{(l-1)m}$ represents the number of intermediate layer feature channels in the $(l-1)$th layer, d is the size of the 2D convolution kernel, and t denotes the size of the 1D convolution kernel. Assuming that the default number of parameters of the SSCNN model when $\lambda = 1$ is the same as the 3D convolution, one can derive the relationship between the number of intermediate layer

channels and the input and output layers in the (2+1)D module:

$$N_{(l-1)m} = \left\lceil \frac{td^2 N_{l-1} N_l}{d^2 N_{l-1} + tN_l} \cdot \lambda \right\rceil \tag{9}$$

where λ is the spatio-temporal channel adjustment factor, and $\lceil \cdot \rceil$ denotes the upward rounding function. When the network model determines the number of feature channels in the input and output layers, by changing the size of λ, the number of channels in the intermediate layer can be adjusted, which in turn adjusts the size of the network model. When $\lambda = 1$, the (2+1)D module has the same parametric size as the 3D module with the same number of input and output feature channels; When $\lambda > 1$, the number of intermediate layer channels is increased with the same number of input and output channels, the spatial features extracted by the network become more, and the number of parameters of the corresponding (2+1) D module is larger than the number of parameters of the 3D module; When $\lambda < 1$, the number of intermediate layer channels is reduced, the model of the network is reduced, and the number of parameters in the (2+1)D module is smaller than the number of parameters in the 3D module.

3.2.3 A Modeling Framework for Gesture and Emotion Classification

The overall network structure of SSCNN-based choral gesture language gesture emotion feature extraction network consists of three (2+1)D convolutional modules, three maximum pooling layers, one global pooling layer, fully connected layer and output layer. The collected choral gesture language data through the feature extraction network finally get the 256-dimensional choral gesture language feature vector map, the gesture sequence $G \in \mathbb{R}^{32\times1\times48\times10\times10}$ is inputted into the choral gesture language feature extraction model, and after the three layers of (2+1) D convolution and pooling operation can obtain a feature vector with feature dimension $\mathbb{R}^{32\times256\times7\times2\times2}$. Considering that the feature distribution of each choral gesture language data sample changes after the convolution operation, batch normalization is added after the 2D and 1D convolution of each (2+1)D convolution module. Due to the small number of gesture data samples and their low-resolution characteristics, directly unfolding the tensor of $256\times7\times2\times2$ into one-dimensional vectors will increase the parameters of the fully-connected layer and is prone to overfitting in the training process. Therefore, the Dropout algorithm is used in the middle of each convolutional layer, and the spatio-temporal global average pooling is used to compress the features of the output tensor of the last convolutional layer, and finally a tensor of $256\times1\times1\times1$ is obtained. Spatio-temporal global average pooling averages all the parameters of each channel and ends up containing only one feature per channel, compressing the number of features and also drastically reducing the number of parameters in the fully connected layer. After compressing the features into a one-dimensional tensor and inputting them into the fully connected layer, the predicted label probability values are calculated by Softmax and the final classification results are obtained.

4 Recognition results of choral conductor's gestures and the emotion of the piece

4.1 Experimental setup

In this paper, the experimental environment is 64-bit Ubuntu18.04 desktop platform, the processor is AMD R7-1700 CPU, and the graphics card is NVIDIA Ge-Force 3090 24 G GPU.

The Adam optimization algorithm is used to train the model. The learning rate was set to 0.0003, and the batch training size was set to 64 according to the hardware conditions of the experimental environment, uniformly sampling 16 frames of images for each video as input, and using scaling, flipping, and adding noise to enhance the data to train 300 epochs.In the experiments in this paper, we selected DHG-14 /28, VIVA Hand Gesture two publicly available choral conductor hand gesture datasets. The proposed network is evaluated using three public emotion image datasets FI, Twitter I and EmotionRoI.

In order to improve the generalization ability and robustness of the model and reduce the overfitting of the model, k-fold cross-validation was performed on each dataset, and finally, the mean square error (MSE) of the k-fold cross-validation was computed as the evaluation criterion. Its calculation formula is:

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i \tag{10}$$

where $MSE_n = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, n$ is the sample size and Y_i is the predicted value of the real data $\hat{Y}_i$.

4.2 Results of Work Conducting Gestures and Emotion Recognition

In this paper, several classical methods are compared and experimented on the dataset DHG-14/28 and dataset TCD-Choral, and the experimental results are shown in Table 1 and Table 2, respectively. The experimental results show that the method proposed in this paper has better performance in gesture action recognition than the current state-of-the-art methods in terms of accuracy. Among them, on the DHG-14 dataset and DHG-28 dataset, the method proposed in this paper improves the recognition accuracy by 2.4 and 1.9 percentage points over ST-TS-HGR-NET, respectively. It improves the recognition accuracy by 3.6 percentage points over the previous best method on the TCD-Choral large dataset. In particular, there is a significant performance improvement relative to the traditional CNN structure and the model architecture based on the LSTM model for temporal feature analysis. As can be seen from the experimental data in Table 2, the performance of the I3D method is significantly improved over the HOG+HOG2 and CNN:LRN methods, most notably due to the fact that the I3D method has been trained with on the ImgaeNet dataset and the Kinematic dataset, which allows for a more adaptable model. In this paper, we borrowed this method and also pre-trained on the ImageNet dataset before migrating to the dataset used in this paper. The experimental results show that the gesture classification model based on spatio-temporally separated convolutional neural networks proposed in this paper can efficiently analyze global temporal features. It proves that the method proposed in this paper has good performance superiority and provides a new idea for the choral conductor gesture task.

Table 1: This method is compared to his classic method in DHG-14/28 data set

Method	Accuracy/%	
	DHG-14	DHG-28
SoCJ+HoHD+HoWR	83.4	80.1
Chen et al	84.5	80.5
CNN+LSTM	85.9	81.4
Res-TCN	86.7	83.2
STA-Res-TCN	89.4	85.4
ST-GCN	91.5	87.6
DG-STA	92.1	88.4
Parallel-Conv	91.4	84.6
ST-TS-HGR-NET	94.3	89.7
This method	96.7	91.6

Table 2: This method is compared to his classic method in the TCD-Choral

Method	Accuracy/%
HOG + HOG2	64.9
Two Stream CNNs	68.3
CNN: LRN	74.8
CNN: LRN: HRN	77.4
C3D	77.6
I3D	83.1
MTUT	86.2
This method	89.8

Using the gesture recognition method based on spatio-temporal separation convolutional neural network, 20 types of choral conductor gestures, such as preparatory beat, two-beat, three-beat, four-beat, dotted-line beat, cyclic beat, strong tone, sudden strong, crescendo, crescendo, legato, staccato, holding, releasing, vocal entry cue, intensity layer control, soft closing, breathing, tension building and overtone, etc., are recognized in TCD-Choral dataset, numbered respectively 1-20, and the recognition results are shown in Figure 2. These 20 gestures constitute the basic vocabulary of communication between the conductor and the choir. In practice, these gestures can be combined in a variety of ways to form a continuous and expressive visual language system, which directly drives the generation and change of musical emotion. For most of the gestures, the recognition accuracy can reach over 90%, for the least accurate overtone, due to its gesture movement and crescendo has a very high similarity, even so, still has 88.4% accuracy, in practice, the design of the gesture human decision, for the gesture of the large differences between the class, the accuracy will be higher.

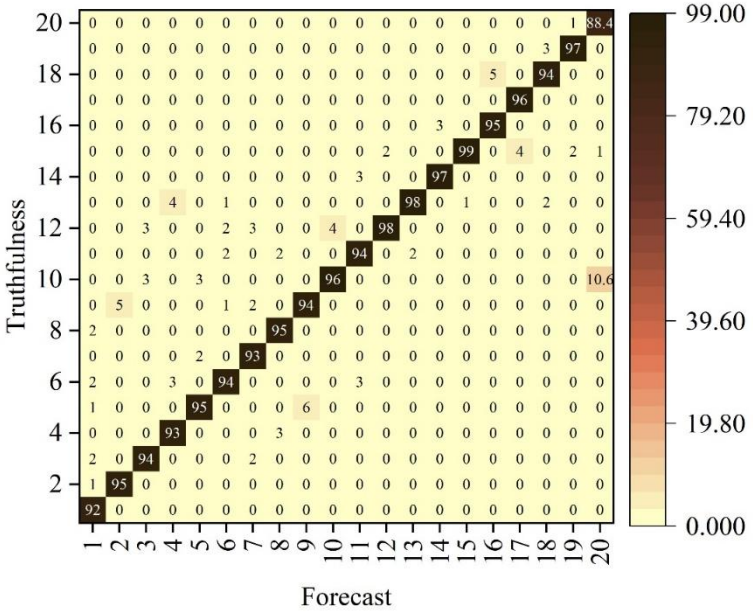

Figure 2: Identify the confusion matrix of 20 gestures

In order to verify the performance of SSCNN, this study will be compared with other existing methods, including traditional methods and deep learning methods, and the experimental results are shown in Table 3. The experimental results are evaluated in terms of classification accuracy of image sentiment, which is defined as the ratio of the number of correctly classified sentiment images to the total number of sentiment images. This study will be compared with the following traditional methods. In addition, this study uses several underlying visual features for the experiments, including GIST, SIFT and HOG underlying visual features. This study also experimented with VGG-16 using emotion image dataset fine-tuning parameters. For deep learning methods, this study firstly compares with DeepSentiBank as well as PCNN. Secondly 2 methods focusing on emotional regions are compared, namely AR and R-CNNGSR. Finally, it is compared with the multilevel emotion recognition model. It can be seen that there are some missing data in the table due to the lack of the same experimental results between SSCNN and the comparison method or the source code of the comparison method is not publicly available. SSCNN achieved 83.51% and 88.15% classification accuracies on the EmotionRoI and FI datasets, respectively, and 89.95%, 86.24%, and 82.38% classification accuracies on the 3 subsets of Twitter I, respectively. By comparing with the above methods, the classification effects of SSCNN are all better than the existing methods. By analyzing them, this is due to the fact that SSCNN takes into account the relationship between the choral conductor's gestures and the emotion of the work, rather than dissecting the emotional expression of the work as an independent entity.

Table 3: Classification accuracy of SSCNN compare with other methods

Method	Twitter I			EmotionRoI	FI
	Twitter I 5	Twitter I 4	Twitter I 3		
Gist	65.92	61.43	60.67	60.42	--
SIFT+BoW	63.13	63.62	60.35	65.34	--
SIFT+VLAD	70.28	68.86	67.09	72.19	--
SIFT+FisherVector	71.16	67.3	65.54	70.98	--
HOG+BoW	68.54	61.87	60.98	61.03	--
HOG+VLAD	72.05	67.8	66.48	63.38	--
HOG+FisherVector	76.06	70.34	68.34	65.32	--
SentiBank	71.35	68.38	66.62	66.15	--
PAE	72.93	69.67	67.84	75.2	--
DeepSentiBank	76.27	70.21	71.28	70.09	61.53
PCNN(VGGNet)	82.62	76.49	76.34	73.61	75.38
VGG-16	83.45	78.75	75.42	72.21	70.64
Fine-tuned-VGG-16	84.39	82.22	76.73	77.02	83.08
AR	88.61	85.19	81	81.27	86.44
R-CNNGSR	--	--	--	81.39	--
Zhang	89.77	85.73	81.55	83.13	87.84
SSCNN	89.95	86.24	82.38	83.51	88.15

Comparative analysis of real and predicted values of emotion types in the EmotionRoI dataset, the real and predicted values of emotion types for each choral conductor content screen are shown in Fig. 3, where the horizontal indicates the predicted class, the vertical indicates the real class, and the diagonal is the prediction.

From the table, we can find that the recognition accuracy rate of the emotion screen of the interesting category is 92%, of which 3% is misjudged as lively and 5% is misjudged as busy. The recognition accuracy of emotional images in the cheerful category is 91%, of which 1% are misjudged as interesting and boring, and 3% and 4% are misjudged as cozy and dull, respectively. The recognition accuracy of emotional images in the lively category was 76%, and more misjudgments were found in the dull and busy categories, with 7% and 13% respectively. The emotional picture recognition accuracy for the warm category was 85%, and 15% were misclassified as dull. The accuracy of recognizing emotional images in the boring category was 95%, with more remaining misclassified as dull. The recognition accuracy of emotional images in the dull category is 95%, with 1% and 4% misclassified as cheerful and cozy respectively. The emotional picture recognition accuracy for the busy category was 94%, with 5% misclassified as dull. The emotion type of the choral conductor gesture work screen with the highest accuracy is the boring and dull category, which achieves 95% accuracy, and the average accuracy of the rest of the emotion categories can be more than 85%, which makes the model's emotion classification performance acceptable.

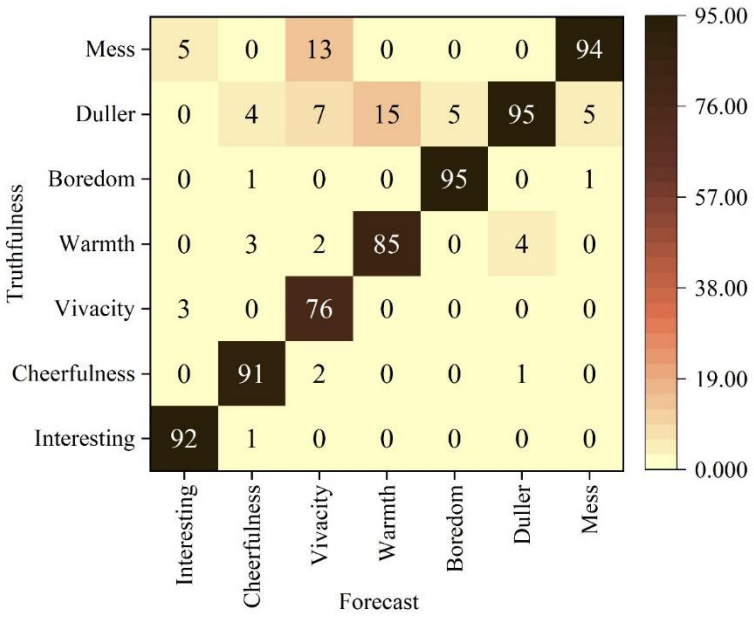

Figure 3: Image identification

5 Analysis of the role of choral conductor's gestures in the emotional communication of the works

The conductor's handling of the strength of the choral conducting gestures of “Sunset Drums” provides a microscopic model of how choral conducting gestures can guide emotional expression. Figure 4 shows the audio waveform of the whole piece of “Sunset Drums” conducted by Peng Xiuwen. Overall, the waveforms of the whole piece of music performed by Peng Xiuwen show an olive shape with two thin sides and a wide middle, and Peng Xiuwen's overall conducting and playing strength follows the law of gradual enhancement of strength and gradual weakening of strength. Especially in Variation 2, the second half of Variation 4, and Variation 5, there is a clear outline of the conductor's intensity as the melodic structure of the music progresses.

Figure 5 is the audio waveform of the overall phrase intensity of Yan Huichang's performance of “Sunset Drums”. As seen in the picture, Yan Huichang's conducting is the smallest change in intensity among the three versions, except for some phrases that have obvious enhancement of the rest of the paragraph changes are less obvious, most of the intensity to maintain a horizontal line, the overall intensity of the arrangement is relatively smooth.

Figure 6 shows the audio waveform of the strength of the whole piece of music “Sunset Drums” conducted by Liu Sha, from which it is concluded that Liu Sha's conductor is the largest of the three versions of the conductor's strength of the greatest ups and downs in the strength of the music across the largest, and Liu Sha's strength of the strength of the increase and decrease are shown in the gradual advancement of the layers of incremental increase in the gradual increase in the strength of the three versions of the conductor's strength of the greatest ups and downs of the strength.

Looking at the audio waveform analysis of the above three performers, although the layout of the playing intensity is basically similar, it presents different graphic characteristics. From the point of view of the overall intensity of the music, Liu Sha's version has the greatest intensity change among the three versions, with a range of [-1,1], followed by Peng Xiuwen's version, and the smallest intensity change is that of Yan Huichang's version.

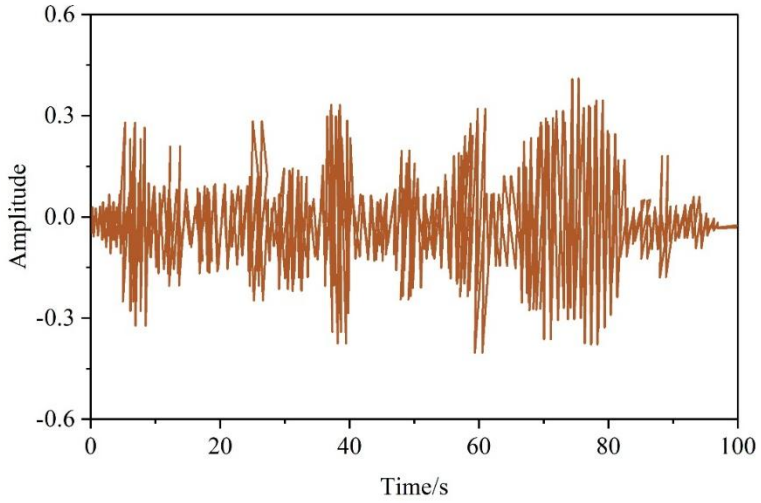

Figure 4: Peng's Full audio waveform

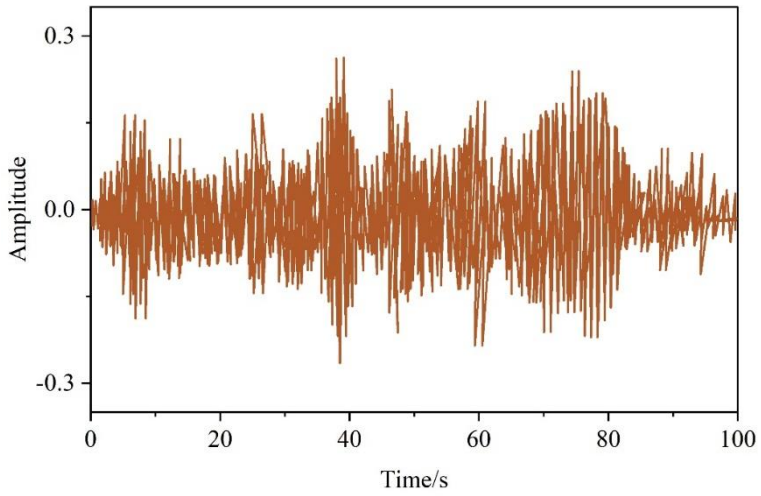

Figure 5: Yan's Full audio waveform

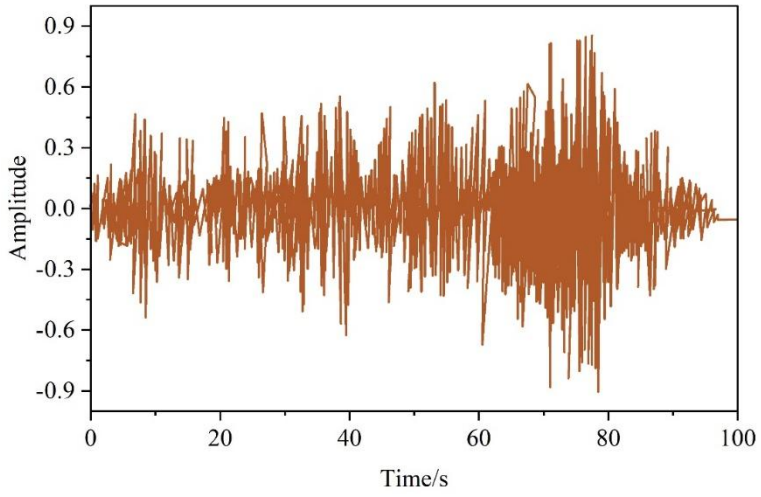

Figure 6: Liu's Full audio waveform

Peng Xiuwen's version of the strength of the performance for the sudden change of strengthening, the whole piece of conductor strength wave diagram shown in Figure 7, the performance of the strength of the changes in most of the instantaneous changes, in the performance of the introduction and variation one of the section Peng Xiuwen's strength of the alternating changes are very frequent, and shows a sudden strong and weak changes in the command effect, variation five to the end of the explosive force will be pushed to the climax of the music, and the end of the rapid quiet, and the climax section formed a stark contrast with the The coda is quickly quiet again, forming a sharp contrast with the climax section. The contrasts between phrases and sections in Peng's powerful performance were obvious, with more frequent changes in strength and weakness. The overall tone was brighter and more emotional, with a focus on the contrasting effects between the strengths, making the piece more dramatic and narrative.

Liu Sha's performance is more delicate than Peng Xiuwen's, focusing on the continuous changes in musical intensity. Although the introductory part of the piece also alternates in intensity, it is different from Peng Xiuwen's performance in that it presents a gradual advancement of intensity, forming a gesture-guided-emotion-injected-voice-strengthened-gesture-adjusted positive feedback, which pushes the whole group to the peak of emotion in a gradual progression. Liu Sha's conductor's strength wave diagram for the whole piece is shown in Figure 8. In Variation 1, Variation 3 and the climax section, Liu Sha's conductor's gesture strength tension is the largest among the three versions, and the strength maximum value is the largest value of 0.88 among the three players, while Peng Xiuwen's maximum value is 0.39. His climax strength peak is also later than Peng Xiuwen's, and his climax persistence is larger than Peng's. He gradually pushes the group from very weak to very weak, and then gradually pushes the group into a positive feedback, which will push the whole group into the emotional peak layer by layer. He gradually pushes the climax from the very weak to the peak of the climax, and the change of intensity is gradual, in a stepwise manner, constantly getting stronger. At the end of the coda, his conductor is not like Peng Xiuwen's, but through the gradual reduction of the layers of strength and weakness, the music returns to calm. On the whole, Liu Sha's conducting focuses on the internal subtlety of the music, the sense of hierarchy, the connection between the strength of the phrases and the changes in the phrases, and the overall conductor's performance of the timbre is also rich and intriguing.

Yan Huichang's conductor's intensity wave diagram is shown in Figure 9, which shows that Yan Huichang's performance is the slowest of the three versions, with the smallest overall intensity. The intensity arrangement is smoother, with less variability between overall sections, but with noticeable changes in intensity for individual tones. This reveals that his gestures may be small in amplitude, but his fingertip and wrist control is extremely fine, with extensive use of dotted line taps to sculpt individual notes. In the variation five to the climax section, although the intensity pattern is strengthened as a whole, and the intensity pattern is close to Peng Xiuwen's peach kernel shape, the maximum intensity is still embodied in the individual notes, and there is no obvious arrangement of the phrases as a whole, which realizes the implicit and introverted expression of emotion.

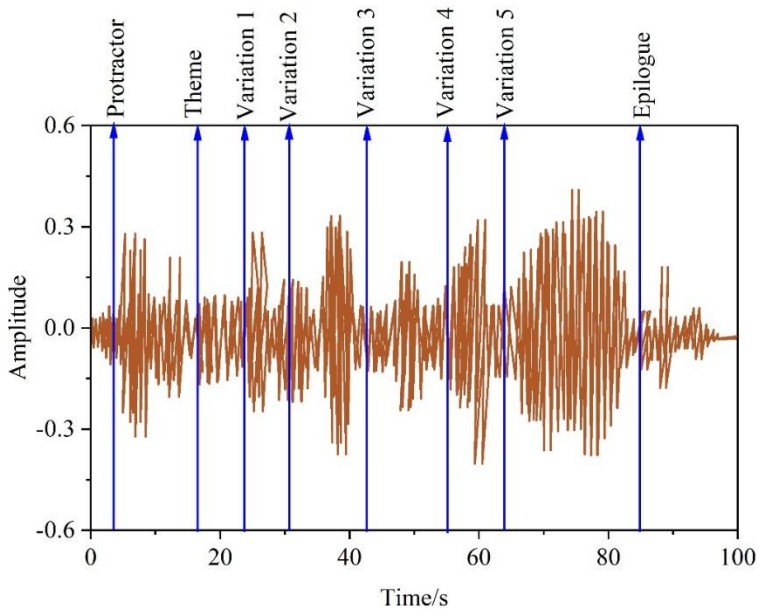

Figure 7: Peng's Full audio waveform

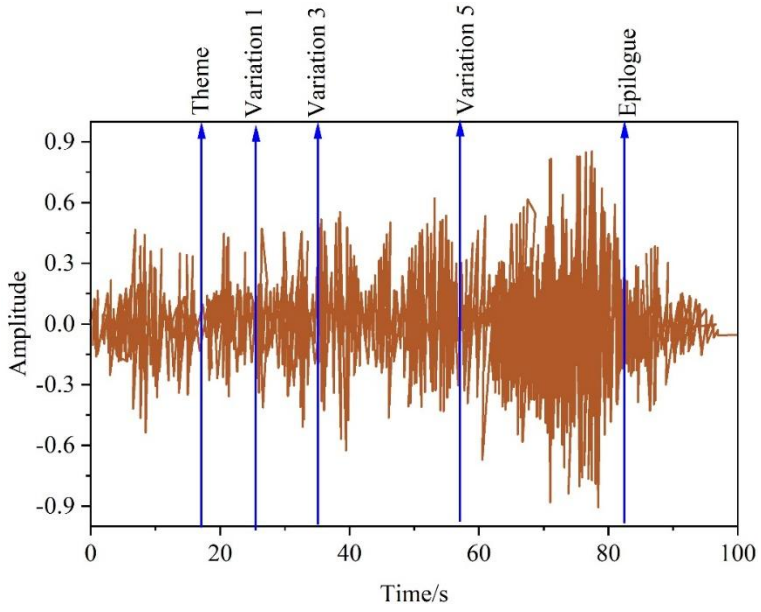

Figure 8: Liu's Full audio waveform

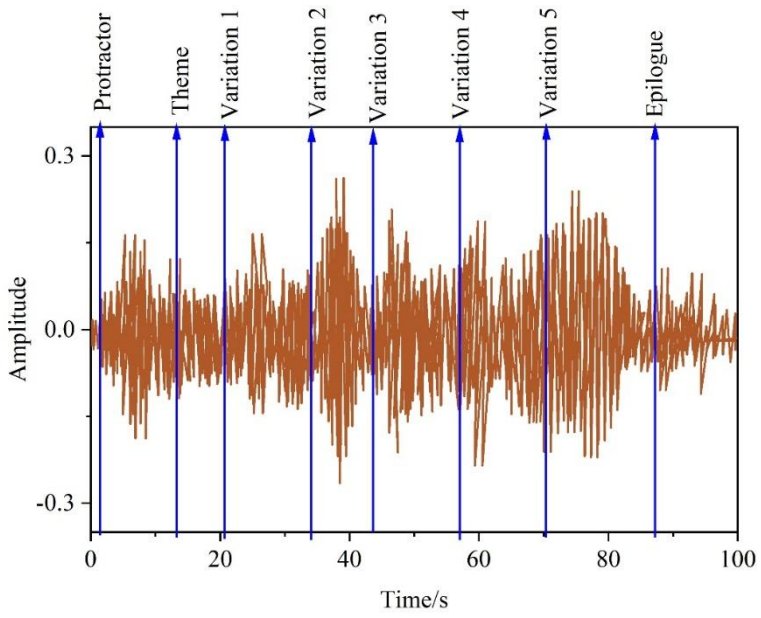

Figure 9: Yan's Full audio waveform

6 Conclusion

In this study, by integrating musicological theory and artificial intelligence technology, synchronized recognition and quantitative analysis of choral conductor gesture and music emotion based on spatio-temporal separation convolutional neural network is conducted to explore the role mechanism of choral conductor gesture language in the emotion communication of works, and the following conclusions are drawn:

(1) Choral conductor gesture language is transformed into corresponding volume, timbre and tension by chorus members through the amplitude, speed and other spatial dynamics features of the gesture. The conductor stimulates the singers' empathy through gestures full of emotional tension, thus commanding the performance of dramatic narrative, emotionally delicate and rich or subtle and introverted emotions.

(2) The differentiation of gesture strategies expresses a variety of complex emotions, and there is no single best gesture language. Peng Xiuwen realizes a grand dramatic narrative through strong contrast and high impact gesture strategies. Liu Sha constructs a delicate and deeply emotional musical narrative through progressive and cumulative gesture strategies. Yan Huichang, on the other hand, polarized the visual-auditory timbre control through refined and embellished gesture strategies, creating an implicitly poetic musical emotional context.

(3) Computational models provide a methodology for traditional art research. The spatio-temporally separated convolutional neural network model used in this paper successfully transforms abstract art descriptions into computable data models. The model's recognition accuracy of different choral conductor gestures and different emotions all reached more than 85%, which not only can recognize choral conductor gestures and emotions with high precision, but also provides a powerful analytical tool for us to parse the complex mapping relationship between gestures and emotions.

About the Author

Shuai Wang, Guyang County, Inner Mongolia, is a teacher of the Art College of North University of China. The research direction is chorus and command, music anthropology, and digital dissemination and application of Chinese excellent traditional culture. Master studied in North University of China, learning choral conductor professional; he studied music anthropology at the Northern University for Nationalities.

References

[1] Grady, M. L. (2013). Considerations of lateral and vertical conducting gestures in evoking efficient choral sound. The Phenomenon of Singing, 9, 122-129.

[2] Kilpatrick III, C. E. (2020). Movement, gesture, and singing: A review of literature. Update: Applications of Research in Music Education, 38(3), 29-37.

[3] Poggi, I. (2011). Chapter 25. Music and leadership: The choir conductor's multimodal communication. In Integrating gestures: The interdisciplinary nature of gesture (pp. 341-354). John Benjamins Publishing Company.

[4] Globerson, E., Flash, T., & Eitan, Z. (2021). Space, time and expression in orchestral conducting. In Space-Time Geometries for Motion and Perception in the Brain and the Arts (pp. 199-212). Cham: Springer International Publishing.

[5] Jansson, D., Elstad, B., & Døving, E. (2021). Choral conducting competences: Perceptions and priorities. Research Studies in Music Education, 43(1), 3-21.

[6] Niemtsova, L. O. (2021, September). A Combination of Professional and Personal Qualities in the Choir Conductor. In ATEE 2020-Winter Conference. Teacher Education for Promoting Well-Being in School. Suceava, 2020 (pp. 319-331). Editura Lumen, Asociatia Lumen.

[7] Garnett, L. (2017). Choral conducting and the construction of meaning: Gesture, voice, identity. Routledge.

[8] Xu, K. (2021). On the Body Language Art in Choral Command. World Scientific Research Journal, 7(7), 85-89.

[9] Napoles, J. (2013). The influences of presentation modes and conducting gestures on the perceptions of expressive choral performance of high school musicians attending a summer choral camp. International Journal of Music Education, 31(3), 321-330.

[10] Rolsten, K. (2016). The production of quality choral performance: A review of literature. Update: Applications of Research in Music Education, 35(1), 66-73.

[11] Nie, L., & Li, J. (2023). Strategy and Practice of Choral Conducting Ability in Normal Universities. Open Access Library Journal, 10(7), 1-10.

[12] Nápoles, J., Silvey, B. A., & Montemayor, M. (2021). The influences of facial expression and conducting gesture on college musicians' perceptions of choral conductor and

ensemble expressivity. International Journal of Music Education, 39(2), 260-271.

[13] Jansson, D., Haugland Balsnes, A., & Durrant, C. (2022). The gesture enigma: Reconciling the prominence and insignificance of choral conductor gestures. Research Studies in Music Education, 44(3), 509-526.

[14] Kumar, A. B., & Morrison, S. J. (2016). The conductor as visual guide: Gesture and perception of musical content. Frontiers in psychology, 7, 1049.

[15] Durrant, C. (2017). Choral conducting: Philosophy and practice. Routledge.

[16] Seighman, G. B. (2015). Exploring the science of ensemble gestures, emotion, and collaboration in choral music making. The Choral Journal, 55(9), 8.

[17] Odusanya, O. S., & Idọlọr, E. G. (2023). Interpretive approach to conducting african choral works: a conductors task. CENTRAL ASIAN JOURNAL OF ARTS AND DESIGN, 4(10), 36-48.

[18] Kumar, A. B., & Morrison, S. J. (2016). The conductor as visual guide: Gesture and perception of musical content. Frontiers in psychology, 7, 1049.

[19] Poggi, I., & Ansani, A. (2017). Forte, piano, crescendo, diminuendo: Gestures of intensity in orchestra and choir conduction. In Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016), Copenhagen, 29-30 September 2016 (No. 141, pp. 111-119).

[20] Platte, S. L., Gollhofer, A., Gehring, D., Willimann, J., Schuldt-Jensen, M., & Lauber, B. (2024). The effect of different preparatory conducting gestures on breathing behavior and voice quality of choral singers. Journal of Voice, 38(6), 1524-e1.

[21] Litman, P. (2006). The relationship between gesture and sound: A pilot study of choral conducting behaviour in two related settings. Visions of Research in Music Education, 8(1), 4.

[22] Cottrell, S. (2007). Music, time, and dance in orchestral performance: The conductor as shaman. twentieth-century music, 3(1), 73-96.

[23] Poggi, I., D'Errico, F., & Ansani, A. (2021). The conductor's intensity gestures. Psychology of music, 49(6), 1478-1497.

[24] Bodnar, E. N. (2017). The effect of intentional, preplanned movement on novice conductors' gesture. Journal of Music Teacher Education, 26(3), 38-50.