Hindawi
Mobile Information Systems
Volume 2023, Article ID 9897383, 1 page
https://doi.org/10.1155/2023/9897383

Retraction

Retracted: Analysis of Computer Visualized Sound Parameters of Vocal Music Singing Based on Deep Learning

Mobile Information Systems

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] N. Mo, J. S. Ai, and S. Y. Ran, "Analysis of Computer Visualized Sound Parameters of Vocal Music Singing Based on Deep Learning," *Mobile Information Systems*, vol. 2022, Article ID 1121092, 9 pages, 2022.

Hindawi
Mobile Information Systems
Volume 2022, Article ID 1121092, 9 pages
https://doi.org/10.1155/2022/1121092

Research Article

Analysis of Computer Visualized Sound Parameters of Vocal Music Singing Based on Deep Learning

Na Mo,[1,2] Jin Shun Ai,[1] and Shi Yi Ran[3]

[1]*Northeast Normal University, Jilin 130000, China*
[2]*Jilin University of the Arts, Jilin 130000, China*
[3]*Jilin University, Jilin 130000, China*

Correspondence should be addressed to Jin Shun Ai; jinsa@nenu.edu.cn

Received 5 July 2022; Revised 10 August 2022; Accepted 13 August 2022; Published 12 October 2022

Academic Editor: Chi Lin

The first impression of the early historical recordings on the contemporary audience is often not ideal; the overall speed is relatively fast, the rhythm is relatively loose and free, and the strings are slippery. In recent years, a large number of research achievements have emerged at home and abroad that apply computer visualization analysis methods to music performance practice. It is feasible and necessary to gradually apply the visual audio parameter analysis method to music research and performance practice teaching. In this study, a prosody hierarchy prediction model based on the CNN (convective neural network) is proposed, and word vectors are added as semantic features. The DL (deep learning) method is applied to vocal music recognition, and a recognition method based on the DL framework is proposed by combining traditional audio signal processing methods. After introducing word vectors as features in the CNN model, the F-score value increased from 77% to 80%. The feasibility of the proposed vocal music recognition algorithm based on DL is verified by experiments.

1. Introduction

The history of vocal music art can be traced back to a long time. In Egypt, Asia Minor, and some eastern countries, especially ancient Greece, artistic singing had already existed as early as a century ago. As singing sound is a kind of compound sound, we can use acoustic instruments to analyze the sound spectrum, calculate the frequency and amplitude (sound intensity) of each component in singing sound, and display it in the form of sound spectrum. For many years, most teachers have been accustomed to correcting 'students' intonation, rhythm, and other basic problems by traditional means such as piano and metronome and involve further expressive problems through the oral description, personal demonstration, or listening to other masters' recordings [1]. Prosodic structure analysis is an important part of the speech synthesis system. Accurately predicting the prosodic boundary position and its grade of text is an important link in speech synthesis, and it is an important prerequisite and guarantee for synthesizing natural and smooth output speech. Most of the teachers and students' time and energy are often spent on such universal and basic issues as intonation, rhythm, speed, intensity, and timbre.

In music performance, the speed and rhythm parameters directly related to time have the most basic structural skeleton, which determines the basic characteristics of a particular performance and has universal validity and comparability across musical instrument types and performance forms. The computer visual analysis of music parameters is not new, such as the visual presentation of waveform diagram for the layout of playing intensity and the objective description of playing timbre and overtone distribution by spectrum diagram. Marinkovic et al. studied monosyllabic samples of French speakers under different emotions, removed the influence of prosodic features on emotions, and mainly studied the differences of sound quality parameters in different emotions [2]. Aubin and Bremond put forward a method of similarity measurement based on Aubin pitch and rhythm information and

Euclidean distance and developed a new vocal music recognition system by using this technology [3]. Senevirathna and Jayaratne divided the process of vocal music recognition into three stages, namely, pitch extraction, melody matching, and note matching and put forward three algorithms for each stage, which combined the results of the three stages to recognize vocal music [4]. Zhang et al. pointed out that, in music performance, the speed of the traditional symphony can be changed between the soil topic, subtopics, paragraphs, and even within the phrase. The importance of the relative speed relationship exceeds the importance of the composer's prescribed speed. The layout is the key to grasp the performance speed of the symphony [5]. Different prosody prediction studies are different in language, language materials, and definition of prosody level, and the evaluation methods adopted for the results are different, so we cannot directly compare the advantages and disadvantages of each model method.

Combining music teaching research with modern information technology, especially computer multimedia technology, is a general trend in the international academic frontier. In the analysis of acoustic features of emotional speech, previous studies mainly focused on the features such as fundamental frequency, sound intensity, and sound length [6] because these feature values are easy to extract by computer software, and they also play a certain role in distinguishing some emotions, but they are not enough to be the basis for accurately predicting the speaker's emotions. The denser the spectral lines on the sound spectrum, the lower the sound. The thinner the spectral line, the higher the sound. Not every harmonic has to appear in the sound spectrum. If the harmonic amplitude is zero or close to zero, a spectral line will be vacant in the sound spectrum. Based on DL (deep learning) technology, this study analyzes the sound parameters of vocal music singing with the aid of computer visualization, and the conclusions obtained through data analysis of visualization software are combined with vocal music teaching, so as to continuously deepen the artistic practice.

2. Related Work

2.1. Research on the Computer Visual Audio Parameter Analysis Method. It is the most varied parameter in vocal music singing, and it is also the most difficult factor to explain clearly through subjective description. Compared with instrumental music, vocal music teaching usually takes more time and energy to correct and adjust timbre. For beginners, the piano can certainly be a less effective and imperfect auxiliary tool. For high-level singers, if the keyboard pitch of the twelve-average law is used as an absolute reference standard for a long time, it is likely to destroy the sense of intonation which should be an expressive factor.

Vall et al. introduced a method to discuss the two most important dimensions of music performance, namely, speed and intensity in the same visual model, which has been widely used in the research of music expression and performance style pattern recognition [7]. McFee et al. proposed a bidirectional echo hiding algorithm [8] by using the front masking phenomenon of human hearing. The audio generated by this method has good auditory transparency, but the accuracy of extracting secret information will decrease with the decrease of echo amplitude. The minimum distortion embedding framework proposed by Bisharad and Laskar lays the foundation for the research of mainstream adaptive steganography algorithms in recent years [9]. Hu et al. gave the carrier elements with large differences between the original audio and the compressed and decompressed reconstructed audio a higher embedded distortion cost [10].

Music alignment technology is particularly important in automatic accompaniment, which can synchronize accompaniment and singing. Automatic accompaniment is to use computer algorithms and knowledge of music arrangement to generate music through some logic. Zhang proposed the modeling problem of inferring note events by identifying the duration of music [11]. Zhang proposed to add asynchronous compensation in audio note alignment to solve the problem of the local alignment error caused by the unsynchronized playing of music melody [12]. Murthy and Koolagudi proposed an alignment framework between audio and MIDI (musical structure digital interface) notes based on spectrum decomposition and dynamic time warping [13]. Lerch et al. put forward linear scaling matching at the level of vocal singing sentences and obtained the optimal path through the dynamic programming algorithm, which overcame the disadvantage that dynamic programming might lose the global optimal path in long path search [14].

2.2. DL-Related Technology. DL (deep learning) is an important branch of machine learning. It is called the prototype of the artificial neural network model, which opens a precedent for scholars to study the artificial neural network model. However, due to the limitation of computer hardware, the later DL research developed slowly, and as a marginal discipline, it did not attract the attention of scholars. However, some theories put forward during this period still guide the development of today's discipline.

Tsagkaraki et al. put forward the perceptron model, which is the first model that can learn feature weights according to sample data [15]. Pouyanfar et al. proposed a backpropagation algorithm, which greatly reduced the time required for training neural networks. While the neural network training algorithm is improved, the rapid development of computer technology also makes the computing power increase by leaps and bounds [16]. Pouyanfar et al. improved the Gaussian mixture model by using the five-layer DBN (deep belief network) structure and achieved remarkable improvement in speech recognition accuracy. The feedforward sequential memory network is used to model sentence speech signals through multilayer convolution layers, and the long-term related information of speech is summarized and expressed, which improves the recognition rate by more than 15% on the basis of the best two-way recurrent neural network speech recognition system in academia and industry [16].

Kim et al. proposed an algorithm to recognize transcribed songs by using the CNN (convective neural network) and trained the CNN model by using cross-similarity matrix generated from a pair of songs as input [17]. Meng et al. put forward a note following system based on chromaticity features and the DTW (dynamic time warping) algorithm, which is used to assist the computer to turn pages automatically [18]. Carter and Briens used the method of layer-by-layer training based on the greedy algorithm to effectively reduce the training difficulty of the multilayer neural network [19]. With the improvement of computing power and training methods, people can further exert the learning ability of multilayer neural networks to solve complex tasks. Güder and Içekli proposed to use the one-dimensional CNN for end-to-end audio steganalysis [20]. Compared with the audio steganalysis algorithm based on feature design, the detection effect is obviously improved.

3. The Research Method

3.1. DL and Audio Analysis. It is the most varied parameter in vocal music singing and the most difficult variable to explain clearly through subjective description. For different singing methods, different singers, and different styles of music works, there is no correct timbre; timbre is always in constant dynamic change. For example, on the commanding point of melody (vowel I), the four versions all have different degrees of timbre brightening and nonmusical components (mainly consonants, lips and teeth rubbing, breath sounds, etc.) tend to increase, etc. Due to the modulation of vocal tract resonance, its overtone amplitude gradually decreases with the increase of its frequency, which will be changed by the selectivity of vocal tract. In the process of training and recognition, it is necessary to detect the starting point of the note, intercept the dataset from the starting point of the note, eliminate the interference of silent noise, and improve the validity of the data.

Compared with traditional machine learning, the biggest advantage of DL is that it can extract nonlinear information from complex data, and it does not need tedious feature engineering. Therefore, for the music recommendation studied in this study, we will explore whether DL has more advantages in learning music-related features than traditional machine learning. Audio files store the waveform information of music signals, while MIDI files store music in the format of binary sequences. Instead of outputting any music signals, they send control information instructions to sound sources for playing music, so playing MIDI with different sound sources will generate music with different timbres.

Constructing a good steganalysis feature greatly depends on the experience and knowledge of steganalysis researchers and the degree of mastering the side information such as carrier model and the steganalysis algorithm. Generally speaking, the lower two peaks, namely, the first formant and the second formant, basically define the vowel tone of sound, while the higher third formant, fourth formant, and fifth formant affect the personal characteristics and musical tone of sound.

The CNN is a kind of deep neural network, which was first used in the image field. Among them, the convolution layer can be regarded as a feature extractor based on sliding window, and each neuron in the convolution layer receives the local feature information from the previous level. The CNN convolution layer, the pool layer, and its activation function transform the input data into abstract feature vectors, and the full connection layer is the marker vector that transforms the feature vectors into the expected values of the data. Figure 1 shows the structure of the CNN, including the input layer, the convolution layer, the pool layer, the full connection layer, and the output layer.

In the latest progress, DL-based recommendation has overcome the obstacles of the traditional linear model, thus significantly improving the quality of recommendation. DL can effectively capture the nonlinear relationship between users and goods and obtain the vector representation of users or goods through vectorization or coding.

The combination of multiple perceptrons is a perceptron, while the neural network is actually a perceptron model with one or more hidden layers. The supervision information in the training stage of classifier can guide the learning of steganalysis features. The overall framework of the vocal music recognition framework based on DL is shown in Figure 2.

The training dataset is input into the vocal music recognition neural network, and the training is carried out in batches. After each iteration, the loss function value is calculated by using the verification set, and the training is stopped after reaching a certain accuracy requirement. The performance of the neural network is tested and evaluated by appropriate evaluation indexes, which can be used as the basis for repeatedly adjusting the training process and parameter selection of the neural network. One can also use a set of model parameters trained by the neural network training module to test the network performance and the end-to-end test of the prototype system on different data sets.

Pre-emphasis digital filter: this kind of filter has the ability to improve the high-frequency characteristics, and it often uses a first-order digital filter. After pre-emphasis, the audio signal can be expressed as follows:

$$H(Z) = 1 - \mu Z^{-1}, \tag{1}$$

where Z is the input vocal singing signal and μ is called the pre-emphasis coefficient, and its value is usually a fraction slightly less than 1. The value used in this experiment is 0.94.

If the signal-to-noise ratio of vocal music singing signal is high, the starting point of vocal music singing can be determined by using short-term energy characteristics. Considering that the signal-to-noise ratio is difficult to reach a high level in practical application, the short-term average zero-crossing rate is further used to assist judgment.

In this study, the double threshold method is used to detect the starting point. The double-threshold method first examines the short-term energy of vocal music singing signal, and the short-term average energy E_n of the voice signal at n time is as follows:

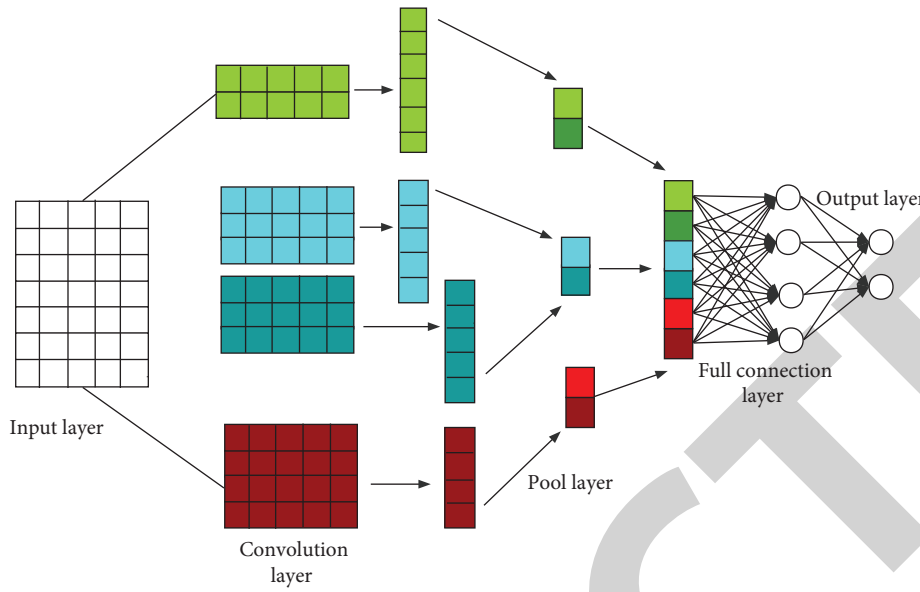

Figure 1: CNN structure diagram.

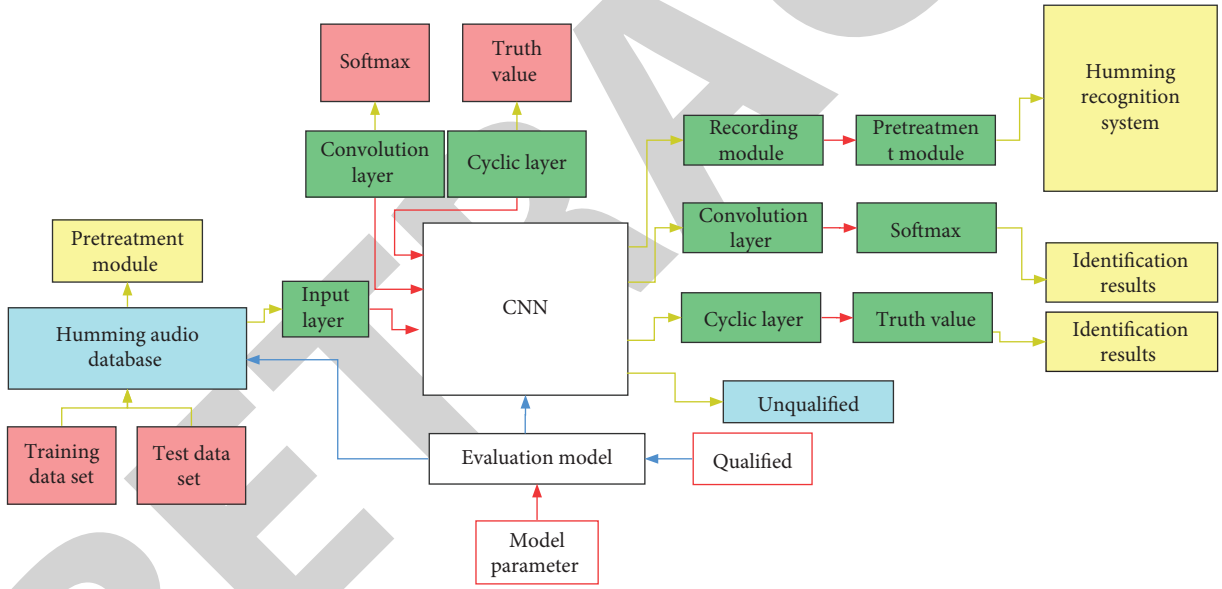

Figure 2: The DL framework of vocal music recognition.

$$E_n = \sum_{m=-\infty}^{\infty} x(m)^2 * h(n-m), \tag{2}$$

where x is the input signal and h is the weight.

HNR (harmonics-to-noise ratios) describes the ratio of the periodic part to the noise part of a speech signal, which mainly reflects the hoarseness of the voice. The autocorrelation function $r(x)$ of the speech delay signal q is defined as follows:

$$r(x) = \int s(t)s(t+x)dt, \tag{3}$$

where $s(t)$ is a stable time signal, and the function takes the global maximum when $x = 0$ is taken.

The glottic waveform diagram shows the difference between the maximum and minimum amplitude in a period. The larger its value, the greater the amplitude of glottic vibration. First, we use linear predictive coding inverse filtering to get the glottic wave and then extract this parameter from the glottic wave. The most common expression of linear predictive model is as follows:

$$\hat{x}(n) = -\sum_{i=1}^{p} a_i x(n-i)\hat{x}(n) = -\sum_{i=1}^{p} a_i x(n-i), \tag{4}$$

where $\hat{x}(n)$ is the predicted signal value, $x(n-i)$ is the previously observed value, and a_i is the prediction coefficient.

Because different users listen to songs for different times, and in order to eliminate the influence of data among users, we normalize the rating data with the maximum value and the minimum value. That is, for the j th element R_{ij} of the i th row in the scoring matrix R, as shown in the following formula:

$$\widehat{R}_{ij} = \frac{R_{ij}}{\mathrm{Max}\left(R_{ij}\right)}, \tag{5}$$

where $\mathrm{Max}(R_{ij})$ represents the maximum value in the i th row of elements and $\widehat{R}_{ij}$ represents the normalized score matrix.

3.2. Note Endpoint Extraction of Audio Signal. Time-related dimensions, such as rhythm, beat, and speed, are the most universal fundamental elements in all forms of music performance, and vocal singing is no exception. They are often used as an important means to outline the structure of music, express mood and expression, and convey the rhythm of a particular style. We can visually present and compare the abovementioned special laws quantitatively and intuitively. It will probably become one of the most important methods of this kind of learning to analyze the sound parameters qualitatively and quantitatively by the visual method. Obviously, if traditional methods such as constant metronome are still used as auxiliary teaching methods in daily teaching, it is unlikely that these deep-seated factors closely related to the second creation of music performance will be directly touched.

Generally speaking, the more compatible the stage environment is with the work, the more realistic the performance of the whole opera is, and it gives people the feeling of being there. In Turandot, in order to achieve this effect, Puccini also made great efforts to create a Chinese stage and bring the audience back to the time when the story happened. At that time, Puccini had only a synopsis of the story, and in order to meet the needs of opera performance, he recreated monologues and lyrics in other ways, making the Chinese style of his works more prominent. In the whole opera, Puccini made six variations on jasmine flower according to the needs of the plot of Turandot. Some of these passages are long, some are very short, and even there is only one line. Some only use solo and chords to express deep silence, while others use orchestra ensemble and chorus to highlight the grandeur. Macroscopically and microscopically, it is a rough and delicate Chinese life.

Excellent music performance itself should not only take "accuracy" as the first pursuit just to respect the spirit of the music score and the original work but also to have some reasonable and meaningful "deviations" from the music score. The elastic expansion and contraction of speed and rhythm is inevitable in any case, and there is only a difference in degree, but there is no difference between there and there. On such a small microscale, the fluctuation of speed has actually changed into the rhythm relationship between sound and sound. Instead of examining the tightness of speed pulsation, it is better to study which sound is longer and which sound is shorter, which can more intuitively and effectively reflect the expansion and contraction of microrhythm.

Without considering the features, the linear model is first used to train the coefficients of linear model fusion. We select the following exponential loss function during training:

$$L(y, f) = e^{-yf(x)}. \tag{6}$$

where $y \in \{-1, 1\}$ and f is the output value of the classifier.

The steganalysis noise caused by steganography is usually much weaker than the content information of the carrier itself. Therefore, using ReLU activation in the shallow layer of the neural network constructed for steganalysis task may lose the negative area information and introduce a large amount of carrier content information in the positive area. It is close to saturation, and the gradient value of the activation unit is close to 0, which is prone to gradient disappearance. Therefore, we introduce the linear truncation activation unit TLU in the first convolution layer of the network as follows:

$$TLU(x) = \begin{cases} -T, & x < -T, \\ x, & -T \le x \le T, \\ T, & x > T. \end{cases} \tag{7}$$

The characteristics of audio signals are nonstationary, and successive notes are superimposed and not hidden from each other, so the end points of notes cannot be directly detected from the time-frequency domain characteristics of audio signals.

The common practice is to convert audio frames into feature sequences that can highlight the starting point of notes after preprocessing. There are two main algorithms, namely, algorithms based on the probability model and algorithms based on signal features. Figure 3 shows a two-way LSTM (Long short-term memory) model with the attention mechanism.

The short-term features are learned by bidirectional LSTM, and then, the attention of the feature vectors output by the cyclic neural network is calculated by the attention mechanism. The calculation formula is as follows:

$$u_i = \tanh\left(W_i \cdot h_i + b_i\right). \tag{8}$$

where h_i is the output of bidirectional LSTM, W_i is the weight matrix, b_i is the offset value, and h_i is calculated by the simple neural network to get its hidden state u_i.

In the scenario of multiclassification steganalysis, the secret carrier generated by the unknown type of steganography algorithm can be regarded as an abnormal sample [18]. The purpose of anomaly detection is to detect the sample points in the dataset that do not meet expectations. CNN networks that solve classification problems often set the Softmax activation layer at the end, and its output can be regarded as the prediction probability $p(m^k)$ of each classification category. From this, we calculate the entropy value $H(m)$ of the output prediction probability of the multiclassification network:

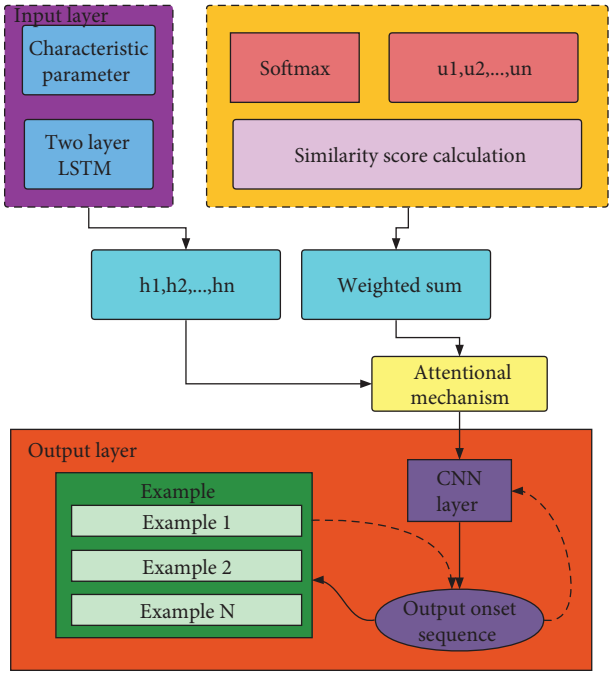

Figure 3: The network model based on attention mechanism.

$$H(m) = -\sum_{k=0}^{M-1} p\left(m^k\right)\log p\left(m^k\right). \tag{9}$$

According to the knowledge of information theory, the entropy value $H(m)$ gets the maximum value $\log M$ when the output probability is evenly distributed. We use formula (10) to set the normalized confidence C:

$$C = \frac{H(m)}{H(m)_{\max}} = \frac{H(m)}{\log M}. \tag{10}$$

It is easy to push $C \in [0, 1]$. Thereby setting the empirical confidence threshold CT. When the confidence is $C \geq CT$, it is considered that the prediction probability distribution of the network output is uniform, and at this time, the network is not sure about the prediction probability of each type of the algorithm.

4. Result Analysis

Western musicians usually think that using the pentatonic scale in their works can reflect the oriental charm. Mahler's Song of the Earth and Stravinsky's Song of the Nightingale all use pentatonic scales with Chinese characteristics. Puccini also created some melodies written in pentatonic scale in Turandot, in order to better create a Chinese atmosphere.

In addition, Turandot contains many pentatonic motives, which either exist or are created by composers. We found that Puccini especially likes to use G flat major and E flat minor in the creation of exotic music. This is obviously because the piano's black keys present pentatonic notes and are easy to play. This section will comprehensively analyze Turandot in the dimensions of speed, rhythm, and beat by using several most important visual analysis methods.

Table 1: Single sample k-s test results.

Emotion	Jitter	Shimmer
Sad	0.436	0.814
Joy	0.028	0.236
Fear	0.302	0.015
Anger	0.556	0.529
Neutral	0.149	0.566
Disgust	0.759	0.584
Surprise	0.501	0.238

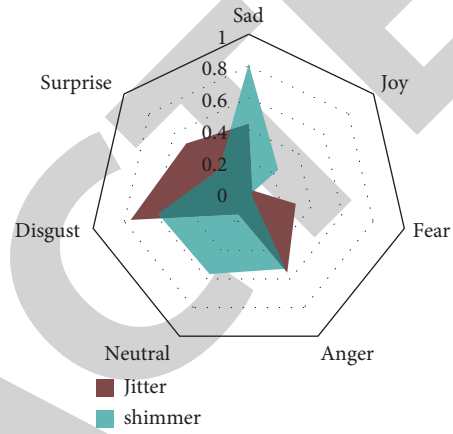

Figure 4: Radar chart of detection result.

In order to study the significant differences in parameters under different emotions, it is necessary to perform one-way ANOVA for each parameter. To ensure the availability of data, we first perform the one-sample k-s test on the data samples before performing one-way ANOVA. The test results of Jitter and shimmer are shown in Table 1 and Figure 4.

It can be seen that the significance of the jitter parameter in happy emotion and the shimmer parameter in scared emotion is less than 0.05, and the sample does not conform to the standard normal distribution. The two-tailed significance of the samples of other parameters is greater than 0.05, so there is no reason to reject the original hypothesis.

In order to better fit the residual error, this study also uses the GBRT (gradient boosting regulation tree) method. The more important super parameters are the maximum number of leaf nodes, the learning rate, the number of forests, and the ratio of positive and negative samples. The idea of combination feature selection in this study is to adopt the principle similar to the greedy method, that is, to add possible binary combination features. If the f-score effect is improved on the basis of cross-validation, this feature is considered to be better, see Table 2.

In this study, the information of word vector word2vec is especially considered as the feature. Generally, it is considered a DL model, and word2vec can express each word as a feature of an N-dimensional vector through training. So, the distance between two words can be expressed by the similarity of word vector space (see Figure 5).

Experiments show that, after introducing word vectors as features in the CNN model, the F-score value increases

Table 2: *F*-score results of the LSTM model.

Overfitting parameter	Number of features	Corpus prosodic phrases (%)
1	120	61.24
0.7	120	75.28
0.4	120	66.93
0.1	120	76.18

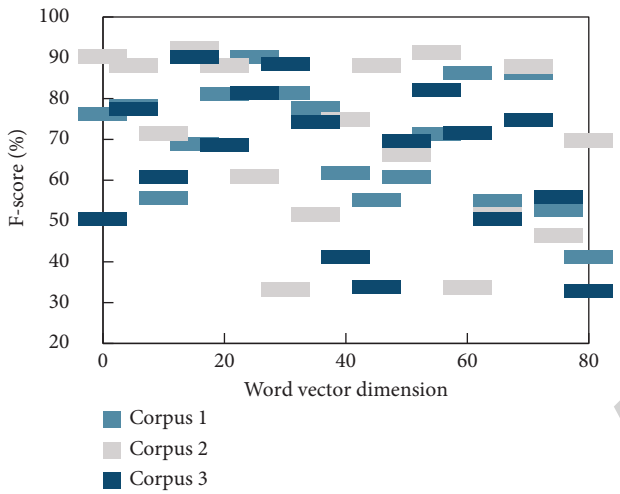

Figure 5: *F*-score results of the CNN model.

from 77% to 80%. The vector feature of this explanatory word can better adapt to the CNN model. However, adding word vectors to the LSTM model will greatly reduce the accuracy of the decision tree model.

As the expressive factors of music performance inevitably require that the performance deviates from the music score, this trend of precision and regularization actually gradually limits the basic space of artistic creation from a scientific point of view. Performance conventions such as over-symbol points, under-symbol points, uneven notes, and arpeggio treatment of chords have a long history and are often part of the complete artistic conception in the creative stage. Therefore, contemporary performance, while sticking to the accuracy of the spectrum, is probably far away from the authenticity of the work and the original intention of the composer.

Sometimes, the research will involve qualitative or quantitative analysis of the rhythm of the whole music, which requires statistical analysis of the microrhythm. In order to avoid the influence of accidental factors, Figure 6 sums up the rhythms of 70 bars of the whole song with 7 beats per bar to get the average value and takes the length of the first tone as a reference to get the relative lengths of other tones. It can be seen that the six beats are not evenly distributed.

The ratio of strong beat to strong beat is very similar, and the extension of the last beat is obvious, while performer 2 emphasizes the overall sense of 5/7 beat rhythm, and each bar is divided into two levels of elastic ups and downs on a microscopic level, thus causing some kind of periodic

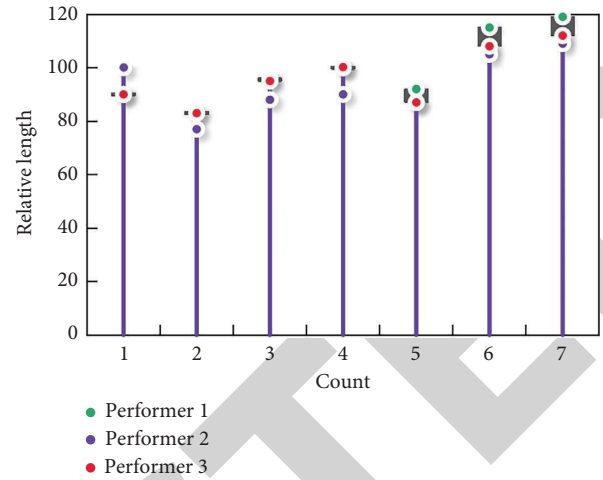

Figure 6: Analysis of rhythm proportion of each bar beat.

fluctuations with a sense of expectation. The tightness difference of this microrhythm is usually closely related to the performer's artistic conception. The unique five-tone style of Chinese music, *G* major, four or four beats, and adagio is adopted, and the melody is very stretched and full of singing. Especially, the repeated use of four-tone groups shows the unique warmth and femininity of oriental women to the fullest.

Puccini endows these characters with character, temperament, and soul with musical melody, treats these ordinary women in life with infinite love, shows them the most essential mentality and angle of human beings, and makes us feel their greatness. Turandot, based on fairy tales, has a strong appeal. However, Puccini, as a school of realism, was not finished when the opera was written for four years. I think it is mainly because the characters are difficult to grasp. Puccini died after writing Liu Er's aria in Turandot. Liu Er's death coupled with Puccini's death virtually added some content to this sad singing.

Figure 7 shows the test set accuracy of the above steganalysis method under different embedding rates of various steganalysis algorithms, in which less than 51% of the results can be considered as steganalysis failure. It can be seen that the CNN proposed in this study has achieved the highest detection accuracy for matching steganography algorithms with different embedding rates.

Similar to the high-order difference filter kernel, the secret information is adaptively embedded in the region with large audio amplitude and drastic changes, so the steganalysis model proposed in this section does not work for it. It shows that the CNN model proposed in this section is more reliable.

As the tested sample contains seven different emotions, it is impossible to judge which parameters have significant differences in which emotional combinations and which emotional combinations have no significant differences only based on the results of the previous section. In order to explore the differences of parameters in specific emotion groups and the difficulty of distinguishing emotion combinations by parameters, this study tests the rank sum of

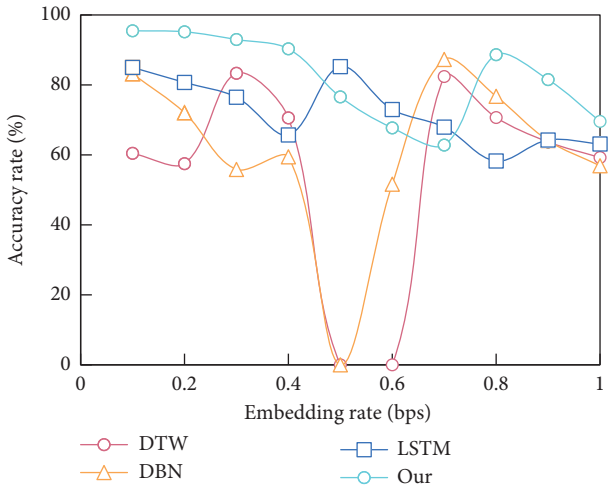

Figure 7: Accuracy of the matching algorithm.

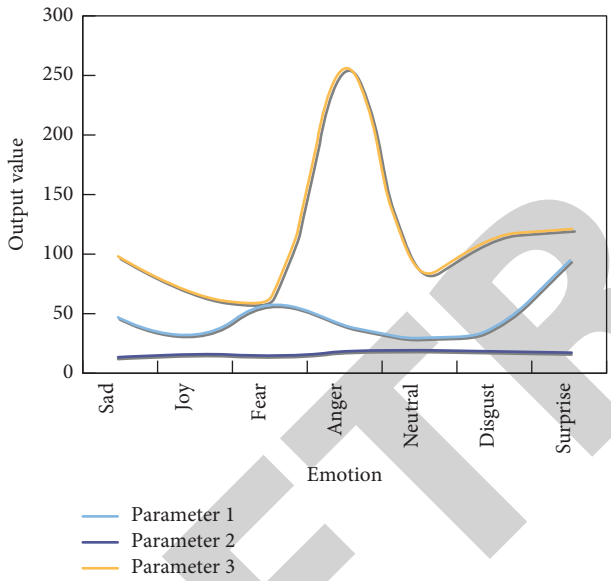

Figure 8: Typical values of vowels correspond to emotions.

emotion pairwise combinations of a vowel. The specific results are shown in Figure 8.

Different voices have different formant modes, and formant is the basic factor that determines the different auditory perception of this voice. Formant speech synthesis mainly depends on formant filters, which are combined in certain rules (usually including series and parallel connection). In the process of simulating vocal tract resonance, many harmonics are usually generated, that is, there are many resonance frequencies. Studies have shown that, at least five resonance frequencies should be selected to make the synthesized speech have high intelligibility and naturalness [17].

Puccini has broken the tradition that creators do not participate in dance design and put forward specific ideas and suggestions on the collection and display of props and the costumes of characters. At the premiere, behind the stage was a huge palace, in front of which was a street lined with shops. The Little Square was on the left and the execution ground was on the right, which was properly arranged. Puccini, according to his own understanding, made great efforts to render the "Chinese style" from the aspects of rhythm, harmony, and band orchestration. This external secondary factor has become an indispensable factor in the overall style of opera, and judging from the actual effect of music performance, Puccini's success is beyond doubt.

5. Conclusion

Vocal music teaching is a complex systematic project, and the sound spectrum analysis technology can provide powerful help for vocal music teaching. Through the analysis method of computer visual sound parameters, we can intuitively "see" the dynamic change process of the various parameters they care about, such as pitch, timbre, intensity, and speed. In this study, through the case analysis of Turandot's classic vocal music works, the basic idea of computer visual method for auxiliary analysis of common elements in vocal performance and teaching, such as timbre, intonation, vibrato, intensity, rhythm, beat, and speed, is fully demonstrated. The CNN DL method is used to predict the prosodic boundary. Through training and testing on the test data set and repeatedly adjusting the model, better model parameters are obtained. The experiment shows that the F-score value of the CNN model increased from 77% to 80% after introducing word vectors as features. The vector feature of this explanatory word can better adapt to the CNN model.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. Yue, B. Zhang, Y. Wu et al., "Simultaneous and signal-to-noise ratio enhancement extraction of vibration location and frequency information in phase-sensitive optical time domain reflectometry distributed sensing system," *Optical Engineering*, vol. 54, no. 4, Article ID 047101, 2015.

[2] Z. Marinkovic, N. Ivkovic, O. Pronic-Rancic, V. Markovic, and A. Caddemi, "Analysis and validation of neural network approach for extraction of small-signal model parameters of microwave transistors," *Microelectronics Reliability*, vol. 53, no. 3, pp. 414–419, 2013.

[3] T. Aubin and J. C. Bremond, "The Process of Species-specific Song Recognition in the Skylark Alauda arvensis. An Experimental Study by Means of Synthesis:pecific song recognition in the skylark alauda arvensis. an experimental study by means of synthesis," *Zeitschrift für Tierpsychologie*, vol. 61, no. 2, pp. 141–152, 2010.

[4] E. N. W. Senevirathna and L. Jayaratne, "Audio music monitoring: analyzing current techniques for song

recognition and identification," *GSTF Journal on Computing*, vol. 4, no. 3, pp. 15–344, 2015.

[5] K. Zhang, H. Song, and L. Zhang, "Active contours driven by local image fitting energy," *Pattern Recognition*, vol. 43, no. 4, pp. 1199–1206, 2010.

[6] Z. Gao, J. M. Song, H. Zhang, A. A. Liu, Y. B. Xue, and G. P. Xu, "Human action recognition via multi-modality information," *Journal of Electrical Engineering and Technology*, vol. 9, no. 2, pp. 739–748, 2014.

[7] A. Vall, M. Dorfer, H. Eghbal-Zadeh, M. Schedl, K. Burjorjee, and G. Widmer, "Feature-combination hybrid recommender systems for automated music playlist continuation," *User Modeling and User-Adapted Interaction*, vol. 29, no. 2, pp. 527–572, 2019.

[8] B. Mcfee, J. W. Kim, M. Cartwright, J. Salamon, R. M. Bittner, and J. P. Bello, "Open-source practices for music signal processing research: recommendations for transparent, sustainable, and reproducible audio research," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 128–137, 2019.

[9] D. Bisharad and R. H. Laskar, "Music genre recognition using convolutional recurrent neural network architecture," *Expert Systems*, vol. 36, no. 4, pp. e12429–e12429.13, 2019.

[10] X. Hu, C. W. Cheong, S. Zhang, and J. S. Downie, "Mood metadata on Chinese music websites: an exploratory study with user feedback," *Online Information Review*, vol. 42, no. 6, pp. 864–879, 2018.

[11] J. Zhang, "Music feature extraction and classification algorithm based on deep learning," *Scientific Programming*, vol. 2021, no. 2, 9 pages, Article ID 1651560, 2021.

[12] K. Zhang, "Music style classification algorithm based on music feature extraction and deep neural network," *Wireless Communications and Mobile Computing*, vol. 2021, no. 4, 7 pages, Article ID 929865, 2021.

[13] Y. V. S. Murthy and S. G. Koolagudi, "Content-based music information retrieval (cb-mir) and its applications toward the music industry: a review," *ACM Computing Surveys*, vol. 51, no. 3, pp. 1–46, 2018.

[14] A. Lerch, A. Xambó, and J. Freeman, "Music information retrieval in live coding: a theoretical framework," *Computer Music Journal*, vol. 42, no. 4, pp. 9–25, 2018.

[15] G. Tsagkatakis, A. Aidini, K. Fotiadou, M. Giannopoulos, A. Pentari, and P. Tsakalides, "Survey of deep-learning approaches for remote sensing observation enhancement," *Sensors*, vol. 19, no. 18, p. 3929, 2019.

[16] S. Pouyanfar, S. Sadiq, Y. Yan et al., "A survey on deep learning: algorithms, techniques, and applications," *ACM Computing Surveys*, vol. 51, no. 5, pp. 92.1–92, 2019.

[17] J. Kim, J. Urbano, C. C. S. Liem, and A. Hanjalic, "One deep music representation to rule them all? a comparative analysis of different representation learning strategies," *Neural Computing & Applications*, vol. 32, no. 4, pp. 1067–1093, 2020.

[18] X. Meng, H. Chang, and X. Wang, "Methane concentration prediction method based on deep learning and classical time series analysis," *Energies*, vol. 15, no. 6, p. 2262, 2022.

[19] A. Carter and L. Briens, "An application of deep learning to detect process upset during pharmaceutical manufacturing using passive acoustic emissions," *International Journal of Pharmaceutics*, vol. 552, no. 1-2, pp. 235–240, 2018.

[20] M. Güder and N. K. Çiçekli, "Multi-modal video event recognition based on association rules and decision fusion," *Multimedia Systems*, vol. 24, no. 1, pp. 55–72, 2018.