Hindawi
Scientific Programming
Volume 2022, Article ID 6893128, 10 pages
https://doi.org/10.1155/2022/6893128

Research Article

Construction of Vocal Timbre Evaluation System Based on Classification Algorithm

Ying Duan

Wuhan Conservatory of Music, Wuhan, Hubei 430060, China

Correspondence should be addressed to Ying Duan; 10362@whcm.edu.cn

Received 8 March 2022; Revised 12 May 2022; Accepted 16 May 2022; Published 6 June 2022

Academic Editor: Sheng Bin

With the continuous development of communication technology, computer technology, and network technology, a large amount of information such as images, videos, and audios has grown exponentially, and people have started to be exposed to massive multimedia contents, which can easily and quickly access the increasingly rich music resources, so new technologies are urgently needed for their effective management, and automatic classification of audio signals has become the focus of engineering and academic attention. Currently, music retrieval can be achieved by selecting song titles and singer names, but as people's living standards continue to improve, the spiritual realm is also enriched. People want to be able to select music with different types of emotional expressions with their emotions. It mainly includes the basic principles of audio classification, the analysis and extraction of music emotion features, and the selection of the best classifier. Two classification algorithms, hybrid Gaussian model and AdaBoost, are used to classify music emotions, and the two classifiers are combined. In this paper, we propose the Discrete Harmonic Transform (DHT), a sparse transform based on harmonic frequencies. This paper derives and proves the formula of Discrete Harmonic Transform and further analyzes the harmonic structure of musical tone signal and the accuracy of harmonic structure. Since the timbre of musical instruments depends on the harmonic structure, and similar instruments have similar harmonic structures, the discrete harmonic transform coefficients can be defined as objective indicators corresponding to the timbre of musical instruments, and thus the concept of timbre expression spectrum is proposed, and a specific construction algorithm is given in this paper. In the application of musical instrument recognition, the 53-dimensional combined features of LPCC, MFCC, and timbre expression spectrum are selected, and a nonlinear support vector machine is used as the classifier. The classification recognition rate is improved by reducing the number of feature dimensions.

1. Introduction

Music, as the most important form of multimedia, has received widespread attention in the field of computer research [1]. In recent years, with the rapid growth of digital music data, the audio information retrieval of music signals has received widespread attention in the field of computer research [2]. In terms of commercial applications of music information retrieval, music software and search engines can be simply implemented at present, but the retrieval of such information is still essentially based on the existing textual tag information of the music signal, such as song name, artist name, and song style, or around the behavioral characteristics of the user to make music recommendations, and the combination of the characteristic factors of the music itself with these technologies is not enough [3]. The characteristic information of music itself (such as timbre, melody, and other musical information) is still to be explored. In essence, it is still a traditional text search, where the text information corresponding to music files can only be obtained by manual annotation, which is not only costly in terms of labor and time, but also an almost impossible task in the face of the extremely large number of multimedia files [4–6]. At the same time, the textual annotation of music files cannot represent the complete information of music, especially the information that reflects the characteristics of music signal itself, such as timbre, melody, pitch, and intonation [7]. The loss of such information can seriously affect the accuracy of music retrieval results and cause inefficient retrieval [8].

Audio information retrieval of musical signals includes several research directions: instrument recognition, singer recognition, humming retrieval, automatic beat detection, and sentiment analysis [9]. Automatic musical instrument timbre recognition is one of the important research contents, which involves the principle of musical sound generation and the perception mechanism of human ear, and is of great significance for the mining and application of the feature information contained in musical signals [10–12]. Musical instrument recognition has similarity with the problem of speaker recognition in speech signal processing in that both of them determine the sound source of the signal based on the timbre characteristics of the audio signal [13]. However, the concept and perception of musical timbre have always been vague and mysterious, and in fact its definition is not clearly defined either in psychology, musicology, or computer science [14]. The complexity of timbre is reflected in the following aspects: timbre is a subjective property of sound perception rather than a purely physical property; timbre is a multidimensional property; no subjective scale has been found suitable for judging timbre; there is no unified set of musical tone signal criteria for researchers to test the developed computational models of timbre. Music is an indispensable seasoning in people's daily life, and people choose to listen to different music in different environments, different moods, and different occasions [15]. For example, in a cafe, soft and quiet songs with a slow rhythm are often played; in a dance party, happy and exciting songs with a fast rhythm may be played [16]; when people have difficulty sleeping at night, they may choose to listen to some calm songs to help them sleep [17]; restaurants may choose some music to enhance the customer's dining experience, and even medical treatment requires the selection of music according to emotions. Even medical treatment requires the selection of appropriate music for psychotherapy [18]. The above examples show that music has become an integral part of people's leisure life and even in the medical field [19–21]. Since emotion is the main content of music, emotion-based music retrieval has also become an important research area.

This paper focuses on music emotion classification techniques. It mainly includes the basic principles of audio classification, the analysis and extraction of music emotion features, and the selection of the best classifier. Two classification algorithms, hybrid Gaussian model and AdaBoost, are used to classify music emotions, and the two classifiers are combined. In this paper, the model parameters of GMM are used as the training data of AdaBoost algorithm, and a music emotion classification system based on the statistical characteristic parameters of GMM and AdaBoost is established. The music emotion classification problem is essentially a pattern recognition process, so the main objectives are to achieve the following:

(1) Preprocessing. The preprocessing of audio signal includes preemphasis, frame splitting, windowing, silent frame discrimination techniques, and endpoint detection.

(2) Extract the audio features that can express the music emotion. The extraction and selection of features are the most important part of the pattern recognition system. In this paper, we select two major types of features, timbre features and rhythm features, which can better reflect the music emotion to distinguish the music of four emotional categories.

(3) Selecting classifiers. A machine learning approach is used for automatic classification of music signals. Gaussian mixture model GMM is a widely used statistical learning algorithm, while AdaBoost is an iterative algorithm with stable performance. In this paper, firstly, the effect of different orders of Gaussian mixture model on classification is investigated, and the optimal order is selected as the system parameter. Secondly, the AdaBoost algorithm is used to construct a weak classifier group for the GMM model.

2. Construction of Vocal Singing Timbre Evaluation System

2.1. Statistical Learning of Classification Algorithms. Traditionally, when designing machine learning methods, the ultimate goal of machine learning is to minimize the empirical risk by adopting the principle of empirical risk minimization. However, in practice, the use of ERM principle to replace the learning objective of expected risk minimization is only an intuitive idea without sufficient theoretical basis, and the experience of neural network design in this regard can fully illustrate that the small error of training samples sometimes does not yield good prediction results but makes the designed network structure weak in generalization. Practical research proves that using multiple complex models to fit a finite number of samples often makes the generalization ability of the trained machine much weaker. In statistical learning theory, the VC dimension occupies an important position and is defined as follows in pattern recognition: for the VC dimension of the set of functions $Q\ (Z, \alpha)$, $\alpha \in A$, the maximum number of vectors that can be scattered by the set of functions can be divided into two different classes using all possible ways in this set of functions. In other words, if there is a sample set of m samples that can be broken by the function set, and there is no $m + 1$ sample set that is intended by the function set, then the VC dimension of the function set is m. Although the VC dimension largely reflects the machine learning ability of the function set, however, there is no theory about calculating the VC dimension of an arbitrary function set, and only the VC dimension of certain functions is known, for those having more complex learning. It is extremely difficult to determine the VC dimension for more complex learning machines (e.g., neural networks). In general, the larger the frame length N, the more pronounced the peak of the autocorrelation function, and the more the data available for analysis. In contrast, the longer the frame, the weaker the transient response when calculating the ACF of the long-time signal.

When the worst distribution scenario is encountered, according to statistical learning theory, the risk in experience

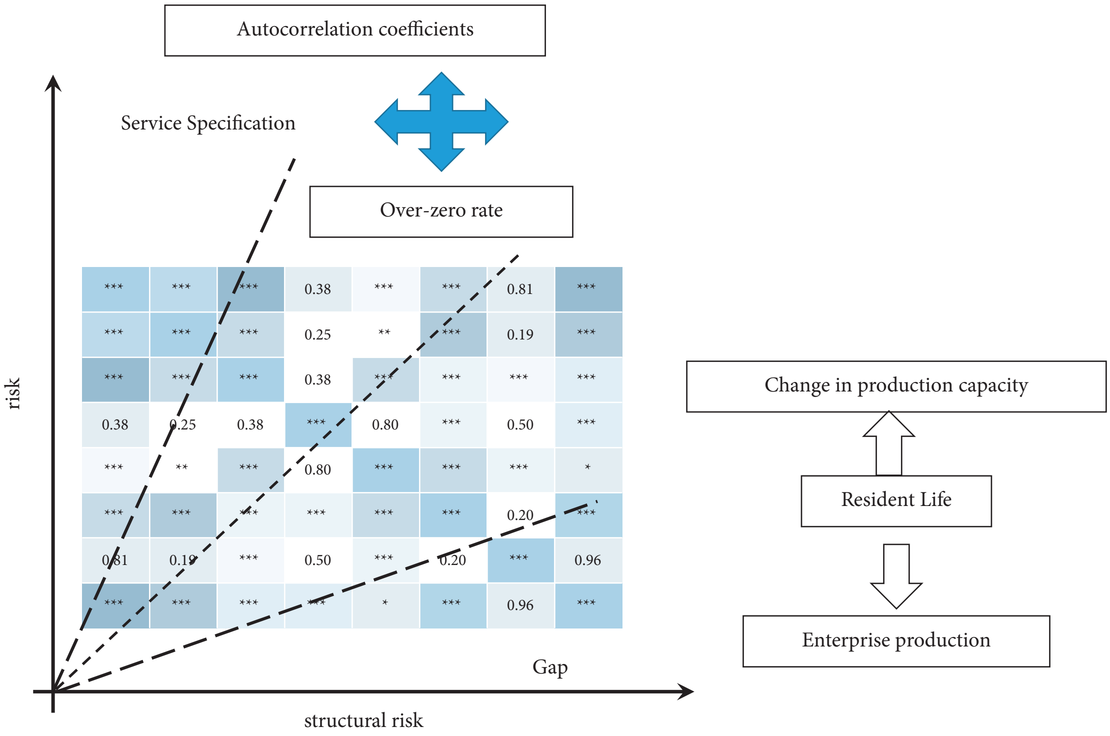

Figure 1: Classification of structural risk.

and the risk in practice both satisfy the following probability relationship $(1-\eta)$:

$$R(W) \leq R_{em}(W) + \sqrt{\frac{\ln(2n1/m) + \ln\,\eta}{n+1}}, \tag{1}$$

$$R(W) \leq R_{em}(W) + \alpha.$$

From the above equation, the confidence interval (VC confidence interval) and the empirical risk together form the actual risk. The VC dimensional confidence interval is a function of the number of machine training samples (N) and the VC dimension of the function set m and is mainly influenced by $1-\eta$ (confidence level). As the number of samples N decreases, or the number of VC dimensions increases, ϑ increases. As shown in Figure 1, usually in practical classification applications, the number of samples (N) is certain, and if a larger number of VC dimensions mean a higher complexity of the classifier, obviously, the corresponding confidence interval range is also larger, which will widen the gap between real and empirical risks. Therefore, when we design a classifier, we should not only consider the problem of minimizing the empirical risk, but also make the number of VC dimensions as small as possible to narrow the confidence interval, so that the expected risk can be minimized. This idea is the structured risk minimization criterion, also known as the SRM principle.

The autocorrelation coefficients and over-zero rates are time domain features calculated directly from the audio signal.

(a) Autocorrelation coefficients is used to represent the spectral distribution of the signal (t_n) in the time domain, which has been shown to provide a good description for classification. We keep only the first 12-dimensional autocorrelation coefficients $(c \in \{1, \ldots, 12\})$, denoted as

$$\mathrm{Xcor}(C) = \frac{\sum_{i=-1}^{n} a(n+c)}{\mathrm{Xcor}(0)}, \tag{2}$$

where Ln is the window length, and c is the time lag. In the experiment, we find the mean and variance of the 12-dimensional autocorrelation coefficients for all frames to obtain the 24-dimensional features.

(b) Over-zero rate is the number of times the value of signal $s(tn)$ crosses the zero axis. This value tends to be small for periodic sounds and large for noisy sounds. To calculate this descriptor, the local DC offset of the signal at each frame is first subtracted, and then the value of the crossing rate at each frame is normalized by the window length Ln. In our experiments, we find the mean and variance of the over-zero rate for all frames.

(c) Log Attack Time: in order to estimate the start (t_{st}) and end (t_{end}) times of musical tones, many algorithms rely on applying to the signal energy envelope (t_{nn}) and the threshold value. The logarithmic initiation time is defined as follows:

$$\mathrm{LAT} = \log\left(t_{\text{end}} - e_{st}\right). \tag{3}$$

In the experiments, the estimated start time (Attack), decay time (Decay), release time (Release), and logarithmic start time of the musical tones were extracted.

(d) Attack slope: defined as the average time slope of the energy in the attack phase.

(e) Decrease slope: defined as the average time slope of the energy in the decay phase, it is a measure of the rate at which the signal energy decreases and distinguishes between nonsustained sounds and sustained sounds.

(f) Temporal centroid: it is defined as the moment when the center of mass of the signal energy envelope is located. The percussion is distinguished from the sustained sound by this feature with the following equation:

$$t_c = t_n \frac{\sum_{i=1}^{n} \exp\left(t_n\right)}{\sum_{n=1}^{i} t_n}, \tag{4}$$

where n_1 and n_2 are the first and last values of n.

The time domain features include 26-dimensional features extracted directly from the music signal and 12-dimensional features extracted from the energy envelope of the music signal, totaling 38 dimensions; the frequency domain features include 44-dimensional common audio descriptors extracted from the fast Fourier transform energy and power spectra, and the mean and variance are obtained for all frames of the signal; the cepstrum domain features include 12-dimensional linear cepstrum prediction coefficients (LPCC). The cepstrum domain features include 12-dimensional linear cepstrum prediction coefficients (LPCC) and 23-dimensional Mel frequency cepstrum coefficients (MFCC) and Mel difference cepstrum coefficients (ΔMFCC). The extracted time-frequency cepstrum-based musical features with their Chinese and English cross-referenced names, abbreviations in the experiments, and the dimensionality of the features are shown in Table 1.

2.2. Harmonic Structure Accuracy Analysis. In the actual music signal, multiple notes are sounded at the same time, and the complexities of the superimposed note spectrum greatly increase the difficulty of harmonic structure extraction, and the following issues need to be considered. The sound of the instruments in this experiment is finally the audio of the instruments with a uniform sampling rate of 44.1 k Hz and a precision of 16 bit, sampled in the uniform format of Wav.

2.2.1. Dissonance Coefficient. The strings of many stringed instruments, including pianos, have a certain degree of hardness, and when the strings vibrate, part of the response force comes from the hardness of the strings themselves. This leads to a spectrum in which the frequency interval of the harmonics gradually increases as the number of harmonics rises; that is, there is inharmonicity, at which point the harmonic frequencies can be approximated as

Table 1: Time-frequency inverse spectral domain-based timbre correlation features.

Musical signal characteristics	Specific gravity	Feature dimension
Start time	3	2
Recession time	4	4
Release time	3	2
Logarithmic start time	6	4
Slope	9	5
Downward slope	8	7
Time focus	4	6
Effective duration	5	
Amplitude modulation	2	
Frequency modulation	1	

$$f_m = B\frac{m \times f_0}{\left(m^2 - 1\right)}, \tag{5}$$

where f_0 is the fundamental frequency; m is the harmonic number; and B is the dissonance coefficient. Its value is related to the instrument and the note.

2.2.2. Frequency Overlap. In contrast to fundamental frequency loss, most instruments have a spectrum of notes in the upper register in which the fundamental frequency is the dominant component, and harmonics occur with low or no harmonic amplitude. Many musicians use notes at frequencies that differ from international standards, and instruments are subject to errors in sounding frequencies due to temperature, humidity, intensity of use, and other factors. For a note with a standard frequency of f_0, the actual frequency can be expressed as

$$Q_{f_m}^R = B\alpha\left(\frac{m \times f_0}{\left(m^2 - 1\right)}\right), \tag{6}$$

where α is the frequency error coefficient of the note.

From the above questions, combined with the frequency of the mth harmonic in the notes of the actual musical signal from the formula,

$$Q_{f_m}^R = \frac{B \times \alpha\left(m \times f_0/\left(m^2 - 1\right)\right)}{\alpha + 1}. \tag{7}$$

If the coefficients (α, B) are given, the actual position of each harmonic in the frequency domain can be determined from the above equation, and thus, the actual harmonic structure of the instrument can be constructed by extracting the amplitude of each harmonic from the equation. However, the coefficients (α, B) are unknown quantities and can only be estimated by testing the input signal of the instrument.

In this paper, a fast Fourier transform is first performed on the actual monophonic signal to obtain the frequency-amplitude spectrum of the signal, and then the coefficients of the note are estimated in the frequency domain using a bandpass filter set with the center frequency of each harmonic frequency f_{mR}. According to the sampling theorem, when the sampling frequency is greater than twice the signal bandwidth, the original signal waveform can be reconstructed without distortion using an ideal filter. Each sample must be quantized by taking sample points, and the more the bits of quantization, the higher the accuracy, usually by taking 8 or 16 bits of digital quantity.

In instrument recognition, the training of instrument models and the recognition of instruments are based on the selected timbre-related feature parameters. In order to make the extracted features more effective, the instrument signal is first analyzed and processed. Preprocessing of the musical signal is an extremely important stage for instrument recognition and classification. The basic process of preprocessing of musical signal is divided into sampling and quantization, removal of silent segments, preemphasis processing, and framing and windowing, and the process is shown in Figure 2.

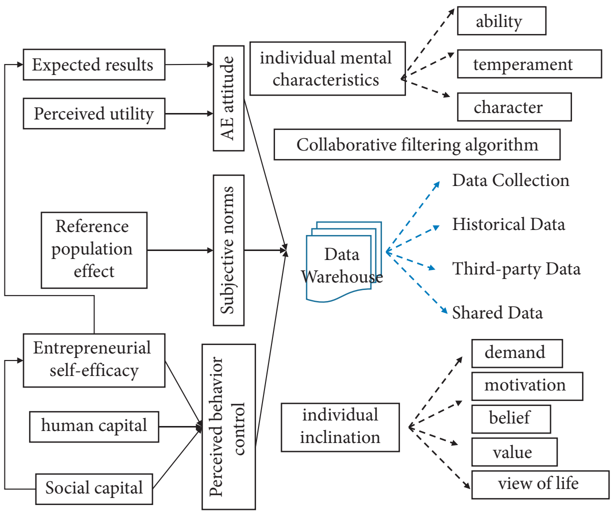

Figure 2: Preprocessing flowchart of musical signal.

When a musical instrument is played, the sound emitted by the instrument is a continuously changing analog signal, but what is needed for processing in a computer is a digital signal, so the input sound must be digitized before the signal of the instrument can be analyzed. Digitization of analog signals is generally achieved by sampling and quantization.

Data preprocessing includes removing DC components, amplitude normalization, and removing the mute segment; this paper uses a short-time energy based double threshold endpoint detection party to remove the mute segment, the preprocessed signal is shown in Figure 3, and the specific steps are as follows.

The first step is to divide the signal into frames, find the short-time average energy, and then compare and judge it based on the threshold value frame by frame.

A coarse judgment is made based on a higher threshold α_1 selected on the short-time energy envelope of the speech, that is, above which the speech must be, and the music start and end point should be located outside the time point corresponding to the intersection of this threshold and the energy envelope.

Determine a lower threshold α_2 on the average energy, search from the previous two end intersection points to the left and right, respectively, and find the two points where the short-time energy intersects with the threshold α_2, which is the location of the start and end point of the music section determined by the double threshold method.

Consider that there may be a minimum length between notes of the music signal to indicate a pause; that is, the end of the music segment is judged only after less than the threshold α_2 satisfies such a minimum length, which actually corresponds to an extension of the coda length.

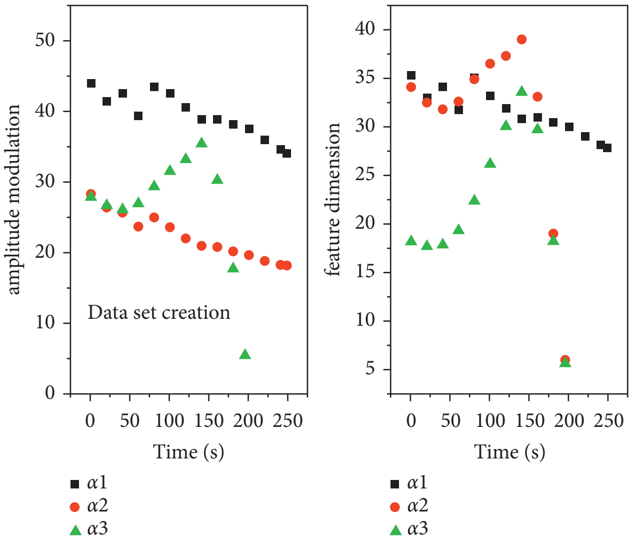

Figure 3: De-mute segment diagram.

2.3. Tone Test. The main principle of the short-time autocorrelation function method of fundamental detection is mostly to use these characteristics of the short-time autocorrelation function to determine the fundamental period by comparing the similarity between the original signal and its delayed signal. If the delay is equal to the fundamental period, then the two signals have the maximum similarity or directly find the distance between the two maxima of the short-time autocorrelation function, which is the initial value of the fundamental period. The calculated value of the autocorrelation function of the piano A3 monophonic one-frame signal is given in Figure 4, and it can be seen that the autocorrelation function has a distinct period, from which the fundamental period of the musical signal can be estimated.

3. Results and Analysis

3.1. Base Tone Detection Based on Classification Algorithm. The unit of the delay amount of the normalized autocorrelation function is the number of sample points, and when the sampling frequency is fs, the delay amount of each sample point is $1/fs$. When using the correlation function method, the delay amount is found between $T_{\text{min}} \sim T_{\text{max}}$, the maximum value of the normalized autocorrelation function is found, and the amount of delay corresponding to the

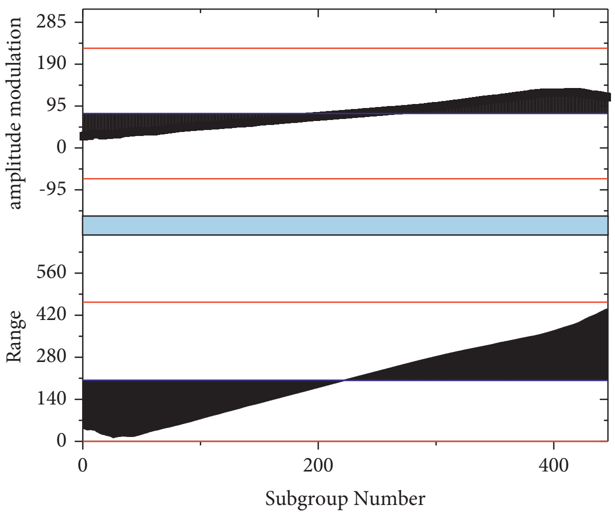

Figure 4: Autocorrelation function of A3 single tone frame signal for piano.

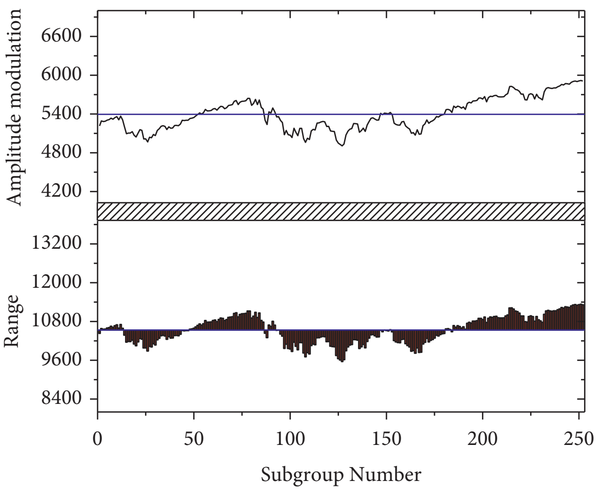

Figure 5: Piano A4 monophonic fundamental cycle estimation chart.

maximum value is the fundamental period. For music signals, common note frequencies lie between 27.5 Hz and 4186.0 Hz, and $T_{\min}$ and $T_{\max}$ can be set accordingly, and the estimated fundamental period is shown in Figure 5.

If the pitch shift is very fast, the use of large length frames will cause errors in the pitch detection results due to missing transient features. Therefore, there is a need to maintain a balance between transient response and accuracy of fundamental frequency calculation. The method to improve the transient response in the autocorrelation function-based fundamental tone detection algorithm is to dynamically change the frame length. The basic idea of the algorithm is to determine the maximum frame length at the lowest fundamental frequency, and if the number of samples in the calculated period is small, the maximum frame length is no longer needed, but a shorter frame length is used. Two notes with rational number relationship in fundamental frequency: they have many harmonics that will overlap; that is, they share the same frequency. Frequency overlap is a major difficulty in the extraction of multiharmonic structures. Loss of fundamental frequency means that, in the spectrum of a note, there is no fundamental component, or the amplitude of the fundamental component is much smaller than that of the other harmonics. Most of the fundamental frequency loss occurs in the bass region notes.

Since the timbre of musical instruments depends on the harmonic structure, similar instruments have similar harmonic structures, and thus, the harmonic structure can be defined as an objective indicator corresponding to the timbre of musical instruments. Based on the existing discrete harmonic transform, the steps of timbre feature extraction based on harmonic structure are as follows: firstly, the music signal is divided into frames with frame length 0.5 s and frame shift 0.25 s, and for signal duration less than 0.5 s, it is not divided into frames; according to the extraction method of harmonic structure, the harmonic structure information of music signal is obtained and normalized to obtain harmonic coefficients; from the discrete harmonic transform coefficients, first-order differential, the discrete harmonic transform coefficients, first-order differential discrete harmonic transform coefficients, and second-order differential discrete harmonic transform coefficients form the timbre expression spectrum. However, the number of incorrect songs in the fast-paced category is higher, and the accuracy rate is not high. This leads to a significant increase in the number of errors in the second level when classifying the pleasant and exciting categories.

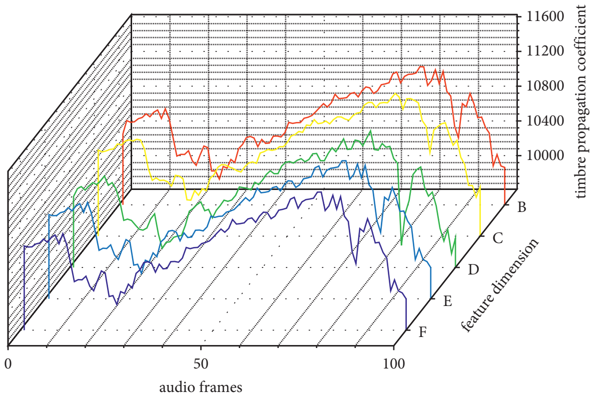

Figure 6: Tone expression spectrum of a single tone.

For A4 monophonic of piano, the fundamental frequency is 440.0 Hz, and as the adopted rate is 44.1 kHz, the window length of discrete harmonic transform is 100 one sample; set the window shift as 1/3 of the window length, for each frame of the audio signal of 1 s long, and the timbre expression spectrum with the highest harmonic number is 10, calculated respectively, among which 50~150 frames of characteristic values are shown in Figure 6. It can be seen that the first few frames of the timbre expression spectrum contain a large amount of the first few frames of the timbre expression spectrum contain a large amount of audio information, and the appropriate highest harmonic number

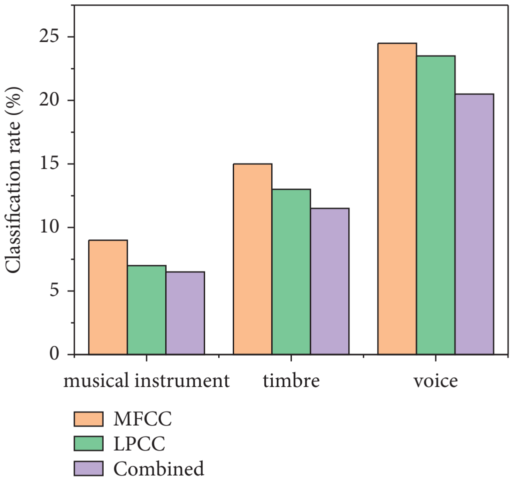

Figure 7: Results of single-tone instrument classification based on combinatorial features.

should be selected when conducting instrument classification experiments.

The recognition rate of musical instruments based on frequency domain features is about 40%–60%, which shows that the frequency domain features of musical signals are not sufficient to respond to the timbre of musical instruments; the recognition rate of musical instruments based on LPCC is between 60% and 75%, considering that LPCC features are only 12-dimensional, which shows that LPCC of musical LPCC of the musical signal has a certain effect on the recognition of musical instruments. In addition, the recognition rate within the woodwind family is generally low, because it contains ten instruments, and the flute, alto flute, and bass flute are more similar, and the Eb clarinet, Bb clarinet, and bass clarinet are more similar. In this paper, the 53-dimensional combination of LPCC, MFCC, and timbre expression spectrum was selected to classify the monophonic files of musical instruments in the above dataset again, and the results are shown in Figure 7.

Figure 8 shows the classification accuracy of the four sentiment categories for the whole test set. The table shows that, for the same training and test sets, the number of misclassified songs is significantly higher when the SVM algorithm is used than when the AdaBoost classification algorithm is used. The data in the table shows that the number of misclassified songs is higher for the pleasant and exciting songs than for the calm and sad songs, which is not consistent with the classification accuracy of each model. The AdaBoost classification algorithm, however, did not have a significant negative impact on the classification of music in the second level because the accuracy of the first level was 0.912, and the number of incorrectly classified songs was very low.

From the results, it can be seen that the number of classification errors is higher for the calm category and the sad category because the songs in these two categories may also get different classification results for different people, and it is certainly a difficult task for the computer to distinguish between these two categories with more overlapping areas. From the results of the above two cases (i.e., with and without singer's gender), we can see that the performance of the AdaBoost classification algorithm is better than that of the SVM classification algorithm. This is because the SVM method directly finds an optimal hyperplane to distinguish the two categories of music and uses it as the final classifier. However, classifying emotions in music is inherently ambiguous, and artificial emotion labeling of music is highly subjective, and the same music fragment may be classified as different emotions. It is difficult to find a strong classifier with high classification accuracy for the ambiguity of music emotion classification, and the advantage of AdaBoost algorithm is to combine multiple weak classifiers to generate a strong classifier, which greatly improves the final classification performance. In order to verify the effectiveness of the two-layer classification system in this paper, we compare the experimental results of using the system structure in the literature with that of this paper and find that the two-layer classification system structure in this paper has better classification results.

According to the experimental results, the combined features consisting of LPCC, MFCC, and timbre expression spectrum were selected as timbre features, and the support vector machine was chosen as the classifier to extract the combined features for ten types of musical signals played by different instruments in the IRMAS dataset, and the recognition rate was 71.14%. The confusion matrix of the classification results is shown in Figure 9. The recognition rate of musical instruments based on time domain features is generally low, which shows that the time domain features of musical signals are not sufficient to reflect the timbre characteristics of musical instruments.

In addition, the recognition rate of a single instrument on this dataset based on deep convolutional neural network is 63.3%, and the recognition rate of this paper is 7.8% higher than that of the SVM-based classification in this paper. In addition, the recognition rate of a single instrument on this dataset is 63.3% based on deep convolutional neural network, and the recognition rate of this paper is improved by

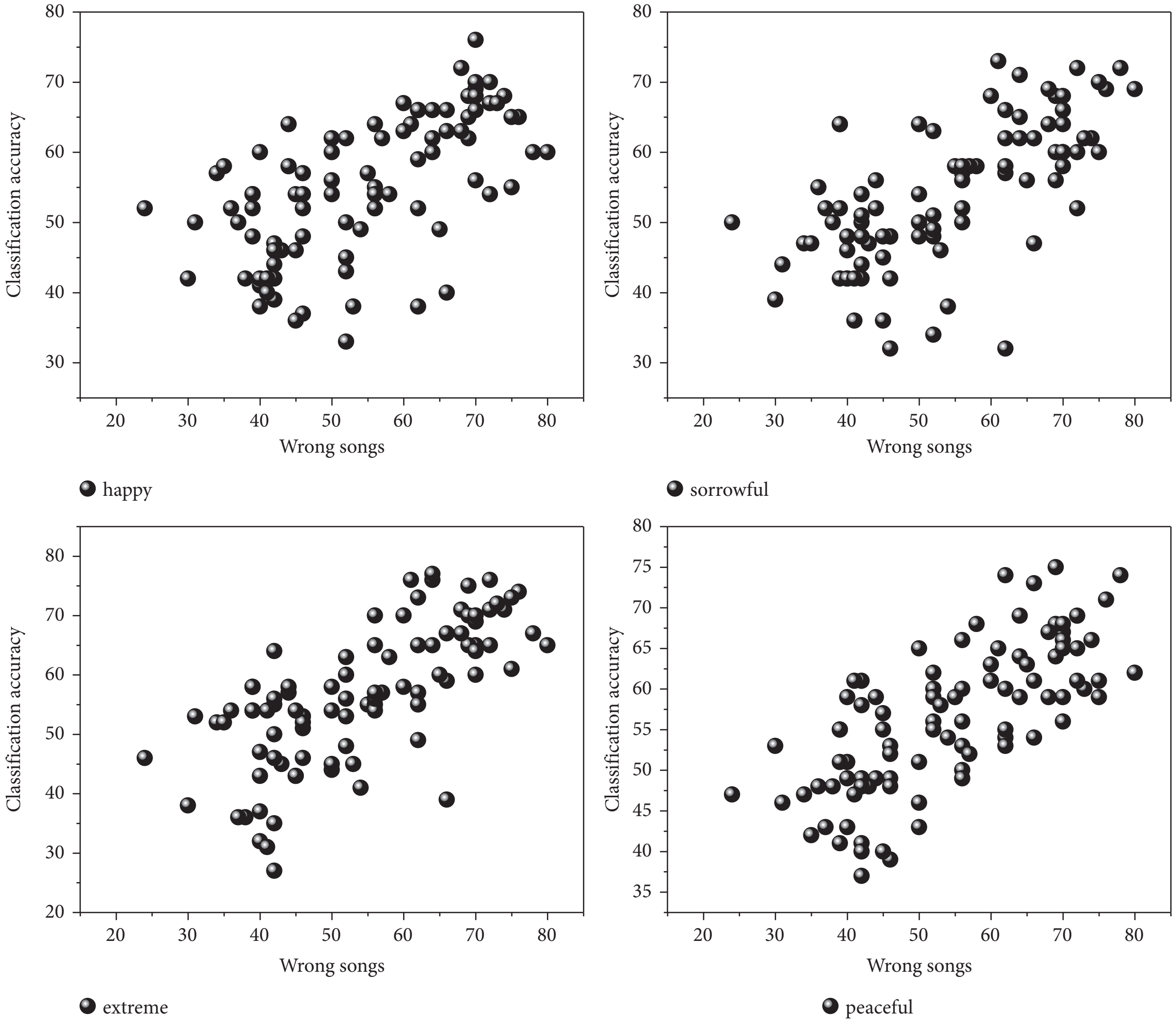

Figure 8: Classification accuracy of the four sentiment categories for the test set.

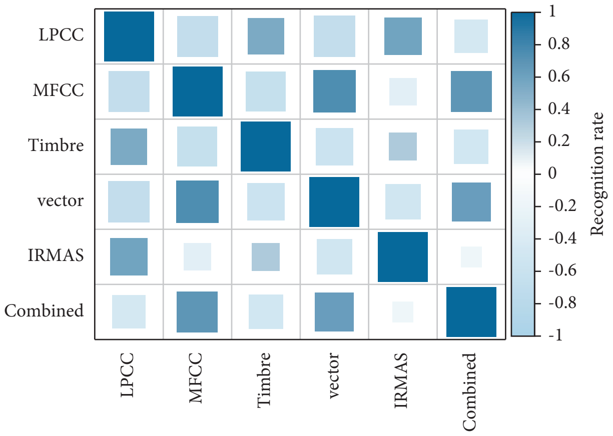

Figure 9: IRMAS instrument classification results based on combined features.

7.8% compared to that. In this paper, we designed and implemented an automatic musical instrument classification recognition system and conducted classification recognition experiments for single notes and musical sections, respectively. Firstly, the experimental database is introduced in detail, including the instrument monophonic timbre database based on the University of Iowa instrument samples and the instrument music segment dataset based on the IRMAS dataset. The selection of timbre expression spectral parameters, different timbre-based features, and instrument recognition classification results based on different classifiers are discussed for the classification of musical instrument monophones. Finally, based on the existing monophonic recognition results, the optimal feature set and classifier are selected to classify the musical fragments, and a recognition rate of 71.14% is achieved for ten types of musical instruments in the IRMAS dataset.

4. Conclusion

In order to better identify musical instruments, it is necessary to find the features that are most directly related to the timbre of musical instruments. In this paper, we focus on the study of timbre-related features of musical instruments, find the features with higher degree of timbre-relatedness and their extraction methods, and use them as the basis for an in-depth analysis and research on the recognition of single tones and sections of musical instruments. Inspired by the constant Q-transform, this paper proposes the Discrete Harmonic Transform (DHT), a sparse transform based on harmonic frequencies, considering the physical significance of the harmonics of musical signals on timbre. In this paper, the formula of the discrete harmonic transform is derived and proved, and the harmonic structure of the musical signal and the accuracy of the harmonic structure are further analyzed. Since the timbre of musical instruments depends on the harmonic structure, and similar instruments have similar harmonic structures, the discrete harmonic transform coefficients can be defined as objective indicators corresponding to the timbre of musical instruments, and thus, the concept of timbre expression spectrum is proposed, and a specific construction algorithm is given in this paper. In the application of musical instrument recognition, the 53-dimensional combined features of LPCC, MFCC, and timbre expression spectrum are selected, and a nonlinear support vector machine is used as the classifier. The classification recognition rate is improved by reducing the number of feature dimensions. In the future, it is chosen as the classifier to extract the combined features for ten types of musical signals played by different instruments in the IRMAS dataset, and the recognition rate was 71.14%.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the project of General Project of Science and Technology Research Plan of Hubei Education Department (Research on Virtual Technology Application in Performing Arts, No. B2021200), Provincial Teaching Team Project of Hubei Education Department (Jazz Teaching Team), and Key Discipline Construction Project of Wuhan Conservatory of Music (Exploration and Research on the Path of "Creation, Editing and Performance" of Original Musical, No. XK2021Y02).

References

[1] F. T. Al-Dhief, M. M. Baki, N. M. a. A. Latiff et al., "Voice pathology detection and classification by adopting online sequential extreme learning machine," *IEEE Access*, vol. 9, no. 2, pp. 77293–77306, 2021.

[2] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part I: review of concepts and an insight to the state of the art," *Biomedical Signal Processing and Control*, vol. 51, no. 3, pp. 181–199, 2019.

[3] P. Sun, "Analysis and recognition of cello timbre based on deep trust network model," *Journal of Physics: Conference Series*, vol. 1533, no. 2, Article ID 022015, 2020.

[4] A. Ivry, B. Berdugo, and I. Cohen, "Voice activity detection for transient noisy environment based on diffusion nets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 254–264, 2019.

[5] R. Jahangir, Y. W. Teh, N. A. Memon et al., "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, no. 6, pp. 32187–32202, 2020.

[6] M. Lataifeh, A. Elnagar, I. Shahin, and A. B. Nassif, "Arabic audio clips: identification and discrimination of authentic Cantillations from imitations," *Neurocomputing*, vol. 418, no. 8, pp. 162–177, 2020.

[7] K. K. Lella and A. Pja, "Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: cough, voice, and breath," *Alexandria Engineering Journal*, vol. 61, no. 2, pp. 1319–1334, 2022.

[8] S. S. Mahmoud, A. Kumar, Y. Tang et al., "An efficient deep learning based method for speech assessment of Mandarin-speaking aphasic patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3191–3202, 2020.

[9] P. Mouawad, T. Dubnov, and S. Dubnov, "Robust detection of COVID-19 in cough sounds," *SN Computer Science*, vol. 2, no. 1, pp. 34–13, 2021.

[10] Y. Srinivasa Murthy and S. G. Koolagudi, "Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS)," *Expert Systems with Applications*, vol. 106, no. 9, pp. 77–91, 2018.

[11] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, "Deep learning for audio-based music classification and tagging: teaching computers to distinguish rock from bach," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 41–51, 2019.

[12] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67745–67755, 2020.

[13] Q. W. Oung, H. Muthusamy, S. N. Basah, H. Lee, and V. Vijean, "Empirical wavelet transform based features for classification of Parkinson's disease severity," *Journal of Medical Systems*, vol. 42, no. 2, pp. 29–17, 2018.

[14] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, 2020.

[15] J. Park, H. Son, J. Lee, and J. Choi, "Driving assistant companion with voice interface using long short-term memory networks," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 582–590, 2019.

[16] M. S. R. Sajal, M. T. Ehsan, R. Vaidyanathan, S. Wang, T. Aziz, and K. A. Al Mamun, "Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis," *Brain Informatics*, vol. 7, no. 1, pp. 12–11, 2020.

[17] C. O. Sakar, G. Serbes, A. Gunduz et al., "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Applied Soft Computing*, vol. 74, no. 5, pp. 255–263, 2019.

[18] A. A. Salih and A. M. Abdulazeez, "Evaluation of classification algorithms for intrusion detection system: a review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 31–40, 2021.

[19] H. C. Tunc, C. O. Sakar, H. Apaydin et al., "Estimation of Parkinson's disease severity using speech features and extreme gradient boosting," *Medical, & Biological Engineering & Computing*, vol. 58, no. 11, pp. 2757–2773, 2020.

[20] W. Wen, G. Liu, Z.-H. Mao et al., "Toward constructing a real-time social anxiety evaluation system: exploring effective heart rate features," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 100–110, 2020.

[21] R. Li, Y. Zhou, Z. Shao et al., "Enhanced coloration/bleaching photochromic performance of WO_3 based on PVP/PU composite matrix," *ChemistrySelect*, vol. 4, no. 33, pp. 9817–9821, 2019.