Hindawi
Mobile Information Systems
Volume 2023, Article ID 9835156, 1 page
https://doi.org/10.1155/2023/9835156

Retraction

Retracted: Music Timbre Extracted from Audio Signal Features

Mobile Information Systems

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 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This article has been retracted by Hindawi following an investigation undertaken by the publisher [1].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] Y. Mo, "Music Timbre Extracted from Audio Signal Features," *Mobile Information Systems*, vol. 2022, Article ID 1349935, 13 pages, 2022.

Hindawi
Mobile Information Systems
Volume 2022, Article ID 1349935, 13 pages
https://doi.org/10.1155/2022/1349935

Research Article

Music Timbre Extracted from Audio Signal Features

Ying Mo

Music College, Wenzhou University, Wenzhou 325035, Zhejiang, China

Correspondence should be addressed to Ying Mo; 00071027@wzu.edu.cn

Received 15 April 2022; Revised 18 May 2022; Accepted 3 June 2022; Published 16 June 2022

Academic Editor: Yanyi Rao

Among the basic elements of music, timbre is one of the most important elements of sound, and it is also the main basis for distinguishing one pronunciation from another. People usually have the ability to "listen and argue" because everyone's pronunciation is different. However, the existing audio extraction technology has low efficiency and low accuracy. Therefore, this paper aims to discuss the algorithm that can make music timbre feature extraction more accurate and efficient. For audio signal feature extraction, this paper proposed an audio feature based on harmonic components to describe the harmonic structure information in the audio signal spectrum. The algorithm in this paper extracts timbre features from the sound data of Western musical instruments and national musical instruments and analyzes the recognition accuracy. The experimental results showed that the classification accuracy of the four feature extractors is above 92%, among which B has the worst effect, with an accuracy of 92.42%, and *D* has the best classification effect, with an accuracy of 99.15%, which shows that the feature extraction algorithm designed in this paper combined with the traditional feature extraction algorithm can achieve better results.

1. Introduction

Sound is one of nature's most common signals. Sound is produced by the vibration of objects. Its existence even predates the existence of living things. As long as there is the vibration and transmission medium of objects, there is the generation of sound signals. Since the birth of human beings, sound has always occupied the most important part of people's lives. For example, natural sounds convey the information of nature to people: if you hear the wind, it means the wind is coming; if you hear the sound of rain, it means it is raining; if you hear the sound of water, it means there are rivers and oceans nearby. Language is the most important communication tool in human society, as a signal for transmitting messages between people. It is convenient, natural, and efficient, and can accurately convey various messages. As for music, its content has risen to the height of human art. People can use different musical instruments to play music in various poses and sounds, and express their emotions such as happiness, anger, sorrow, and music with the help of music.

In the subject of sound, the component analysis of audio signals has always occupied the mainstream of audio signal analysis technology. Various components of the audio signals determine the theme emotion expressed by the music signals. By analyzing the characteristics of various components of the audio signal, the characteristic parameters of the music can be extracted, and the music signal can be classified, identified, and retrieved. This is of great significance for establishing a high-performance and high-accuracy music retrieval database and implementing music classification algorithms based on music style, music content, and musical instruments.

The main innovations of this paper are as follows: (1) feature extraction—this part mainly analyzes the characteristics of the audio signal on the basis of preprocessing and then extracts the characteristic curve of the audio signal to pave the way for the subsequent audio melody matching. (2) Audio feature library construction—this part mainly studies the music in MIDI file format and uses the improved contour algorithm to build the audio feature library, which is used as the data source for melody matching. This plays an

important role in the timbre analysis of various musical instruments in the text, and he can distinguish the timbres of various musical instruments.

2. Related Work

The timbre analysis is of great significance not only to music and musical instruments, but also to the current simultaneous interpretation, speech recognition, etc., so there are many studies on timbre analysis. Kim et al. have done research on classifying musical instruments from polyphonic music, which he believes is a challenging but important task in music information retrieval [1]. Tatar et al. introduce latent timbre synthesis, a new approach to audio synthesis using deep learning. This synthesis method allows composers and sound designers to interpolate and extrapolate between the timbres of multiple sounds in the latent space of audio frames [2]. Banerjee et al. believe that the analysis of sound signals in the linear deterministic approach reached a new dimension and developed many well-equipped software to precisely measure and control the basic parameters of sound, such as pitch, intensity, and rhythm [3]. Rossetti and Manzolli believe that analyzing electroacoustic music is a challenging task that can be solved by different strategies. He proposed to use representations of complex dynamical systems (such as phase space graphs) in music analysis to reveal the timbre characteristics that arise in acoustic music based on particle techniques [4].

It can be seen that most of the timbre analysis uses audio signal extraction technology and deep learning technology. There are also many studies on the extraction of audio signal features. An overview and benchmarking of Sharan et al. test various audio signal representation techniques for classification using CNNs, including methods for processing signals of different lengths and combining multiple representations to improve classification accuracy [5]. Santosh et al. believe that speech, music, and audio signals are essential for communication (e.g., sharing information) and entertainment. This automatic processing of signal processing reduced expert and/or human intervention [6]. Baxter believes that audio metering and monitoring can be described as the ability to audibly or visually determine certain characteristics of an audio signal. For example, loudness measurements are accumulated over a period of time, and it may be difficult for a mixer to detect changes in audio by ear through control room speakers, but the changes obtained through LKFS meters are noticeable [7]. However, it can also be clearly found that most of the current audio signal extraction technology is limited to signal extraction, and the accuracy in removing noise and identifying timbre is not high enough, so this article will conduct in-depth research on this.

3. Music Tone and Its Principle

3.1. Principle of Hearing. The process of human perception of audio information is carried out through the hearing of the human ear. The process of hearing includes from sound vibrations to changes in electrical potential and the release of chemicals, to the emergence of nerve impulses, and finally to central information processing. Therefore, to understand the human perception and cognitive mechanism of audio information, it is necessary to start from the physiological structure of the human ear and the process of auditory perception.

The ear is an important human sense. The function of the ear is to first receive external sounds and then convert the received sounds into neural signals that humans can recognize. Sound perception refers to the internal processing of the received sound through the brain and finally converts it into semantics that humans can understand [8, 9]. The human ear consists of three parts: the outer ear, the middle ear, and the inner ear. The cochlea of the ear is the human auditory sense; the inner ear is responsible for position determination and balance; the outer ear is responsible for sensing external sounds; the middle ear is responsible for transmitting external sounds to the inner ear; the inner ear is responsible for converting the incoming sound energy into human nerve stimulation. The structure of the human ear is shown in Figure 1.

The perception process of the human ear is inseparable from the brain, which is a very sensitive organ. The eardrum vibrates when sound waves from the external environment travel through the air in the external auditory canal to the eardrum. The vibration of the tympanic membrane is transmitted by the ossicles through the middle ear. During the transmission process, the ossicles vibrate the cochlear fluid, which in turn causes the basilar membrane to vibrate and finally generates traveling waves. Different sounds produce different traveling waves. When the organ of Corti with the basement membrane vibrates, electrical potentials appear on the hair cells. The characteristic of this potential is that the frequency and waveform are the same as the external stimulus sound. This potential also stimulates the nerve fibers in the lower part of the hair cells, generating action potentials. This action potential causes a change in the electrical potential between the auditory nerve and the hair cell, which in turn creates a chemical reaction. A certain substance produced by the chemical reaction will stimulate the nerve endings and finally transmit the excitement generated by the nerve endings to the nerve center.

3.2. Tone. The basic knowledge of music theory is essential, and all music production is closely related to this knowledge. These music theory knowledge can help us better understand the elements in music. Music melody includes many elements: melody, rhythm, beat, speed, dynamics, range, pitch, sound intensity, sound value, timbre, etc. [10]. The common physical quantities of sound, such as loudness, pitch, and timbre, are shown in Figure 2.

Timbre: the timbre of the sound. The difference in timbre is mainly due to the different inherent properties of objects. Different timbres can be combined into a different very nice music, and the timbre can also determine the image of the music.

From a disciplinary point of view, the concept of timbre contains multiple attributes such as physical and

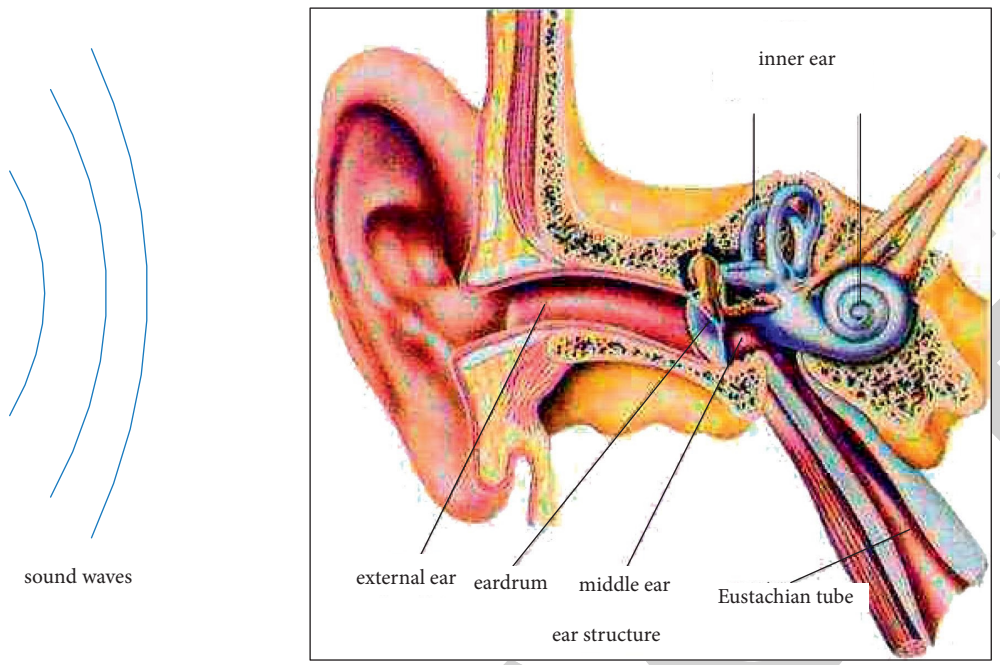

Figure 1: The principle of human hearing.

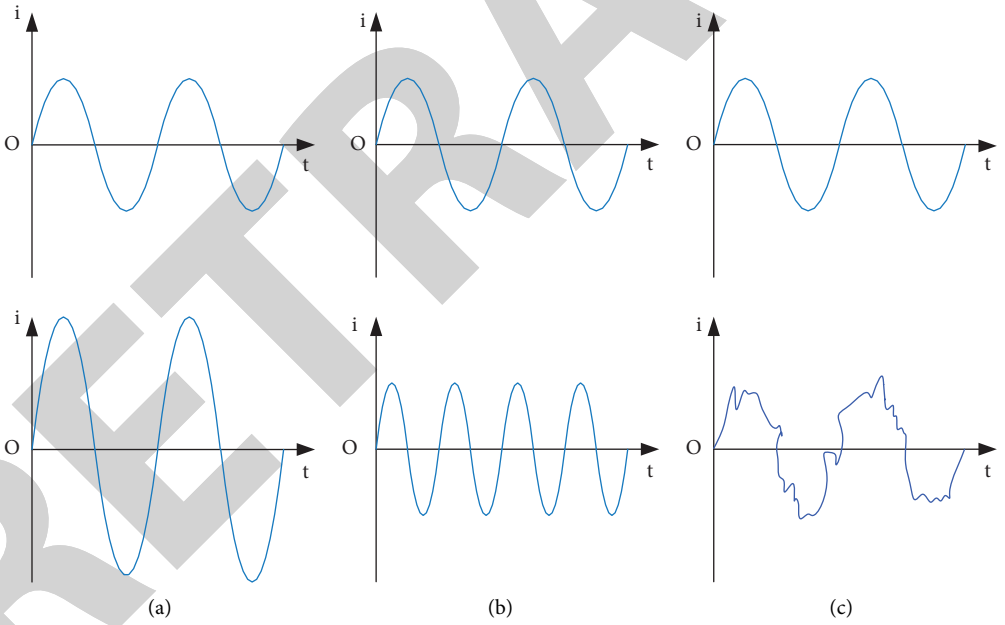

Figure 2: Volume, tone, and timbre. (a) Volume level. (b) Pitch high and low. (c) The timbre is different.

psychological. In different academic fields, the definition of timbre and its impact will also be different. From the point of view of physical acoustics, timbre is a certain property of sound produced by hearing, and it is a kind of vibration wave propagating in the medium. The timbre usually heard is a composite tone composed of the fundamental tone and an overtone produced by the sounding body. Therefore, the structure, material, shape of the sounding body, the number of overtones above the fundamental sound generated by these factors of the sounding body, and the amplitude of each overtone are the fundamental reasons that affect the timbre. Therefore, in physical acoustics, the difference in timbre depends on the distribution and intensity of each overtone [11, 12].

From a psychological point of view, timbre is a key factor in creating an auditory linkage between performers and listeners. The performer can convey the musical mood and musical image of the musical work to the audience through the timbre, so that the audience can produce synesthesia in the auditory through the unique timbre perception. This synesthesia effect directly affects people's emotions and subjective cognition, which is an important basis for music

therapy. Music therapy is one of the important methods and means of psychotherapy, which is to relieve emotions and cure diseases through music [13, 14]. It uses various forms of musical activities, including listening, singing, playing, rhythm, and other means to stimulate and hypnotize people and stimulate physical responses to sound, so that people can achieve health goals. Among them, the choice of different timbres and different sound qualities has an essential and even decisive influence on the treatment. It is mentioned in "Huangdi Neijing" that the melodious, quiet, honest, and solemn tone is as broad and firm as "earth" and can enter the spleen.

In the theory of musicology, timbre is an auditory property, and the timbre of a musical tone refers to the sound property produced by a musical instrument when a note is played. When different musical instruments play notes of the same pitch, strength, and time value, the listeners will perceive the sound perception of different attributes, that is, the difference in timbre.

3.3. Tone Classification

3.3.1. Tone Classification of a Single Instrument. There are many kinds of musical instruments, and understanding each musical instrument and its classification will help us learn and train timbres. This paper discusses the classification of musical instrument timbres into the two categories of Chinese national musical instruments and Western musical instruments. In the classification of traditional Chinese national musical instruments, the "Introduction to Ethnic Art" of the Chinese Academy of Arts divides musical instruments into wind instruments, stringed instruments, plucked instruments, and percussion instruments according to the performance of the instruments themselves and the differences in their playing methods.

In the orchestration method, according to the principle of orchestration, various Western musical instruments divide the modern symphony orchestra into different musical instrument groups according to their own materials, structure, sounding methods, and performance skills: woodwinds, brass, percussion, plucked, bowed, and keyboards. Some of the Western and ethnic musical instruments are shown in Figure 3.

To sum up, to have a clearer understanding of the classification of musical instrument timbres, the author divides the common Western musical instruments in the table through first-level classification, second-level classification, and third-level classification. The sound classification of each musical instrument is shown in Tables 1 and 2.

3.3.2. Tone Classification of Musical Instrument Combinations. In this paper, according to the classification of a single instrument timbre, the instrument combination timbre is defined as the combination of two or more musical instrument timbres. The definition of instrument combination is extremely simple, but the form of instrument combination tone is very rich, there are two instruments playing in unison at the same time, and there are also band formations formed by a combination of multiple instruments. According to the division of musical instrument timbre types in this section, the timbres of musical instruments can also be divided into two types: national orchestra and Western orchestra. The national orchestra includes four categories of wind instruments, stringed instruments, plucked instruments, and percussion instruments. Western orchestras are also known as symphony orchestras. Western musical instruments can also be divided into four groups according to the origin and timbre of the instruments—woodwind, brass, strings, and percussion. The combination forms of the two bands are extremely rich. With the development of composition technology and the development of cultural exchanges, Chinese and Western instruments also appear in many musical works at the same time. Due to the wide variety of Chinese and Western musical instruments, the combinations are endless. In this article, the author takes Western musical instruments and Western orchestras as the main content and conducts research and discussion on musical instrument timbre perception training [15, 16].

3.3.3. Classification of Human Voices. In the classification of vocal timbre, according to the bel canto method, the timbre difference of human voice can be divided into children's voice, male voice, and female voice according to age and gender; according to the division of the range, it can be divided into soprano, alto, and bass; in the field of mixed chorus, it can be divided into soprano, mezzo-soprano, tenor, and bass. After synthesizing various vocals, it is divided into children's voices, male voices, and female voices as a whole. To sum up, most scholars divide human voices into children's voices, male voices (tenor, baritone, bass), and female voices (soprano, mezzo-soprano, and alto) after synthesizing the above classification basis.

4. Audio Feature Extraction Based on Harmonic Components

This chapter proposes an audio feature based on harmonic components to describe the harmonic structure information in the audio signal spectrum. This chapter first introduces the commonly used frequency-domain features and then introduces the human brain's perception characteristics of harmonic signals. In particular, the research results in psychoacoustics on the use of fundamental frequency, spectral peaks, and timbre in the harmonic signal by the human brain to distinguish different audio signal types. On this basis, a harmonic dictionary is proposed, which describes the harmonic structure in the spectrum by constructing harmonic atoms composed of fundamental frequency, formant, and overtone energy decay rates in the frequency domain.

4.1. Frequency-Domain Features of Audio. Sound is a mechanical wave that can be transmitted by vibrations in solids, liquids, and gases. Most commonly, it is airborne to the human ear. There are cilia of different lengths in the inner

tribal instrument western musical instruments

Figure 3: Some ethnic and Western musical instruments.

Table 1: Classification of Western musical instruments.

Stringed instrument	Bowed string instrument	Violin, viola, cello, double bass
	Plucked instruments	Harp
Wind instrument	Woodwind	Flute, clarinet, oboe, bassoon
	Brass instruments	Trumpet, horn, trombone, large
Percussion	Has a fixed pitch	Timpani, xylophone, carillon, row bell, etc.
	No fixed pitch	Snare drum, bass drum, triangle, cymbal, tambourine, gong, etc.
Keyboard instrument	—	Piano, celesta
Electric musical instrument	—	Electronic organ, electric piano, etc.

Table 2: National instruments.

Wind instrument	Whistleless musical instrument	Flute, pan flute, etc.
	Whistle instrument	Pipe, suona
	Reed instrument	Sheng
Stringed instrument	—	Gaohu, jinghu, erhu, banhu, zhonghu, zuihu, etc.
Plucked instruments	Play musical instrument	Pipa, ruan, sanxian, yueqin, etc.
	Flat musical instrument	Zither, guqin
	Stringed instrument	Dulcimer
Percussion	Drums	Drums, war drums, waist drums, long drums, etc.
	Gong class	Big gong, small gong, cloud gong, etc.
	Cymbals	Large cymbals, small cymbals, cymbals, etc.
	Bang class	Board, clapper, wooden fish, etc.

ear of the human ear. Through the resonance of the cilia, the mechanical vibrations are converted into nerve impulses. After the impulses reach the brain, they "hear" the sound through a series of higher-level perceptions [17, 18]. Since ciliary resonance reflects that the human ear can convert time-domain signals into frequency-domain signals, researchers have proposed a variety of frequency-domain features that reflect spectral characteristics. The discrete Fourier transform (DFT) can transform a sequence of time-domain samples of an audio signal into the frequency domain, converting the audio signal x_m of the mth frame length of N sampling points to the frequency domain; X_m can be calculated by the following formula:

$$X_m(k) = \sum_{n=1}^{N} x_m(ne)^{-2i\pi nk/N}. \quad (1)$$

Among them, n ($1 \le n \le N$) and k ($1 \le k \le K$) are the time-domain and frequency-domain sampling indices, respectively, and K is the DFT length. On this basis, there are various spectral features used to describe the distribution characteristics of the spectrum:

(1) Bandwidth. The bandwidth of an audio signal describes whether the frequency distribution of the signal is more dispersed or relatively concentrated and is defined as follows:

$$BW_m = \sqrt{\frac{\sum_{k=1}^{K} (k - SC_m)^2 |X_m(k)|^2}{E_m}}. \quad (2)$$

The bandwidth of speech is generally 300~3400 Hz, and the bandwidth of music is usually much larger

than that of speech, which can be as high as 22 KHz [19].

(2) Sub-band energy ratio. The spectral distribution of audio signals generated by different sound sources is different. For example, the spectral energy of speech signals is mainly concentrated in the low-frequency part, while the spectral distribution of music signals is relatively average. Therefore, by dividing the frequency band into several sub-bands, and separately calculating the energy ratio of each sub-band to the entire spectrum, the distribution characteristics of the spectral energy can be roughly described.

$$SER_{m,i} = \frac{\sum_{k=L_i}^{H_i} |X_m(k)|^2}{E_m}. \tag{3}$$

Among them, H_i and L_i, respectively, represent the upper and lower frequency of the ith sub-band, and the sub-band bandwidth $H_i - L_i$ can be divided into equal or unequal lengths.

(3) Sub-band spectral flux. Sub-band spectral flow refers to the cumulative change of the corresponding intensity of adjacent frequencies in each sub-band of the spectrum, which can be used to detect the frequency components of sudden changes in each sub-band, and is defined as

$$SF_{m,i} = \frac{1}{H_i - L_i} \sum_{k=L_i}^{H_i} \left| \widehat{X}_m(k+1) - \widehat{X}_m(k) \right|. \tag{4}$$

Among them, $\widehat{X}_m(k)$ refers to the normalized spectral signal. The normalization is to avoid the scale difference between different frames due to different energies. It is achieved by converting the spectral energy into a decibel scale and normalizing it to unit energy, that is,

$$\widehat{X}_m(k) = \frac{10\log_{10} X_m(k)}{\sqrt{\sum_{k=1}^{k} \left| 10\log_{10} X_m(k) \right|^2}}. \tag{5}$$

(4) Spectral roll-off point. The spectral roll-off point is defined as at frequency RP_m, the sum of the spectral amplitudes less than this frequency accounts for 85% of the sum of the entire spectral amplitudes, that is,

$$\sum_{k=1}^{RP_m} X_m(k) = 0.85 \sum_{k=1}^{K} X_m(k). \tag{6}$$

The spectral roll-off point describes the energy ratio of the low-frequency part and the overall shape of the spectrum.

(5) Linear prediction coefficient (LPC) and linear spectrum pair (LSP). Linear prediction coefficient refers to the method of describing the vocal tract model that produces speech using linear prediction analysis [20], as shown in Figure 4.

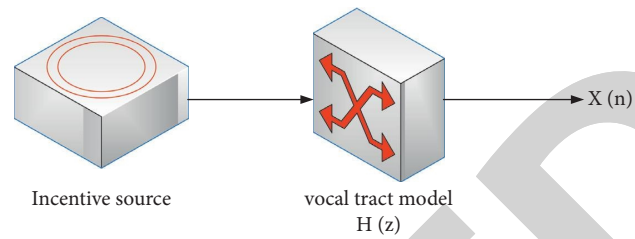

Figure 4: Linear prediction model for speech production.

Usually, an all-pole model is used to describe the vocal tract model, as

$$H_p(z) = \frac{G}{A_p(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}}. \tag{7}$$

Among them, G is the gain, $A_p(z)$ is the transfer function of the p-order linear filter, and a_i, $i = 1, \ldots, p$ represents the linear prediction coefficient, that is, the filter parameter. The prediction coefficient is usually obtained by minimizing the prediction error, and the solution algorithms include covariance method, autocorrelation method, and Levinson–Durbin grid method. Among them, the Levinson–Durbin algorithm is the most commonly used [21, 22].

Line spectrum pair is a feature based on linear prediction and a deduction parameter of linear prediction coefficient. In harmonic signals, this feature describes the distribution characteristics of formants (i.e., peaks of spectral envelope). The line spectrum pair feature treats the channel as a cascade of $p+1$ resonant cavities, which represent the resonant frequencies of the resonators when the excitation energy is at a local minimum or a local maximum, respectively. In speech signal processing, it corresponds to the resonant frequency of the vocal tract when the glottis is fully closed or fully opened.

The recurrence relation of the transfer function can be obtained by the Levinson–Durbin algorithm,

$$A_{p+1}(z) = A_p(z) - k_{p+1} z^{-(P+1)} A_p\left(z^{-1}\right). \tag{8}$$

Among them, the reflection coefficients $k_{p+1} = 1$ and $k_p = -1$ correspond to the boundary conditions when the glottal door is closed and opened, respectively, and $P(z)$ and $Q(z)$ are used to represent $A_{p+1}(z)$ when $k_{p+1} = 1$ and $k_p = -1$ are, respectively,

$$P(z) = A_p(z) - z^{-(P+1)} A_p\left(z^{-1}\right), \tag{9}$$

$$Q(z) = A_p(z) + z^{-(P+1)} A_p\left(z^{-1}\right), \tag{10}$$

$$A_p(z) = \frac{1}{2}\left[P(z) + Q(z)\right]. \tag{11}$$

The roots of equations (9) and (10) both lie on the unit circle of the z-plane and alternate, representing $P(z)$ and $Q(z)$ in the form of factorization, respectively,

$$P(z) = \left(1 - z^{-1}\right) \prod_{i=1}^{p/2} \left(1 - 2\cos w_i z^{-1} + z^{-2}\right),$$
$$Q(z) = \left(1 + z^{-1}\right) \prod_{i=1}^{p/2} \left(1 - 2\cos \theta_i z^{-1} + z^{-2}\right), \tag{12}$$

where roots w_i and θ_i satisfy

$$0 < w_1 < \theta_1 < \ldots < w_{p/2} < \theta_{p/2} < \pi. \tag{13}$$

The factorization coefficients w_i and θ_i appear in pairs and reflect the spectral resonance frequencies, so they are called line spectral pairs. Because of its good quantization and interpolation properties, it is widely used in the research on vocoders for speech coding.

4.2. Harmonic Dictionary. Humans can hear sounds with frequencies between 20 and 20 KHz. The objective description indicators of sound waves include frequency (fundamental frequency), harmonic components, sound pressure, and amplitude. Human perception of sound includes loudness, pitch, and timbre. Harmonic means that a signal can be decomposed into a fundamental frequency sine wave plus several other higher frequency sine waves, and each higher frequency is an integer multiple of the fundamental frequency. The frequency components of these octaves are usually called overtones (Overtones), for example, the 2 times the frequency components of the fundamental frequency are called the second harmonic or the first overtone. Usually, the generation of harmonic signals is due to the resonance phenomenon when the excitation source passes through a resonant cavity. The frequency of the excitation source corresponds to the fundamental frequency of the harmonic signal, and the resonance frequency of the resonant cavity is reflected in the frequency spectrum as the formant frequency. Taking speech as an example, the opening and closing cycle of the glottis determines the fundamental frequency, while the vocal tract as a resonant cavity determines the formant frequency. More commonly, many musical instruments such as violins, pianos, and guitars generate scales by resonating strings and resonating boxes with different vibration frequencies. The vibration frequency of strings is determined by the length, thickness, and material of the strings. Instruments such as violins and guitars control their fundamental frequency by changing the length of the part of the string that can vibrate by pressing a finger on the string. Harmonic characteristics can be used to distinguish sound sources with and without resonant cavities. Sound sources with resonant cavities can generate harmonic signals, such as the resonant cavities of speech and music, which are the vocal tract and the resonance box, respectively; sound sources without resonant cavities produce nonharmonic signals, such as the sound of brakes and the sound of a river, and some researchers have proposed algorithms to distinguish the two types of signals. While DFT is capable of converting a time-domain audio signal to the frequency domain, this method of extracting frequencies is different from how humans perceive them. According to the research results of psychoacoustics, when people hear a sound with harmonic structure, they do not perceive the frequency of each single overtone in turn, but perceive the signal as a fundamental frequency as a whole. The number of overtones, energy size, and overtone energy decay rate is perceived as timbres. This perceptual fusion phenomenon is due to the brain's ability to use harmonic relationships to organize complex acoustic environments into independent acoustic targets. An intuitive example of this brain function is when two speakers have a difference in fundamental frequency, even if they speak at the same time, a person can easily distinguish two speakers.

Fundamental frequency refers to the greatest common divisor of each octave in a harmonic signal and is an objective measure. Pitch refers to the pitch perceived by people, which mainly depends on the fundamental frequency, intensity, and subjective feelings of people. The fundamental frequency detection algorithm can be performed in the time domain or the frequency domain, usually using the method of directly finding the peak and trough positions, the autocorrelation function method, and the comb filter. The methods in the frequency domain include cepstral method, maximum similarity method, and methods based on wavelet transform. Timbre involves many fields related to psychology and is related to human perception and various characteristics of sound, including the nature, material, and shape of the sound source, the number of overtones, the rate of energy decay, and the shape of the spectral envelope. As shown in Figure 5, the time-domain waveforms and power spectrograms of the clarinet and violin in a short time frame (16 kHz sampling rate, 32 ms frame length) are shown. The overtone energy of the black tube attenuates quickly, the number of overtones is small, and its second formant appears around 1718 Hz. The violin has more overtones, and the second formant appears around 3125 Hz. These factors together determine the timbre characteristics of the clarinet's low sound and the violin's clearer sound.

In this section, the signal spectrum is decomposed using the proposed harmonic dictionary and matching pursuit algorithm. Harmonic spectral component extraction technology is usually used in music signals, such as multi-fundamental frequency detection by decomposing the power spectrum of the signal into a series of base vectors.

The power spectrum s of a short-duration audio frame can be represented as a linear combination of a set of basis vectors.

$$s = \sum_{(f_{\max}, w, \sigma) \in A} \delta_{f_{\max}, w, \sigma} d_{f_{\max}, w, \sigma} + r. \tag{14}$$

Among them, $d_{f_{\max}, w, \sigma}$ are the atom in the dictionary, $\delta_{f_{\max}, w, \sigma}$ are the scale corresponding to the atom, and the $f_{\max}, w, \sigma$ parameters, respectively, represent the fundamental frequency, the center frequency, and the side lobe decay rate. A is the set of parameters for the selected atoms in the sparse representation, and r is the residual.

The solution process of the sparse representation adopts an iterative process based on the matching pursuit algorithm. In the ith iteration, a basis vector $d_{(f_{\max}, w, \sigma)^i}$ is chosen,

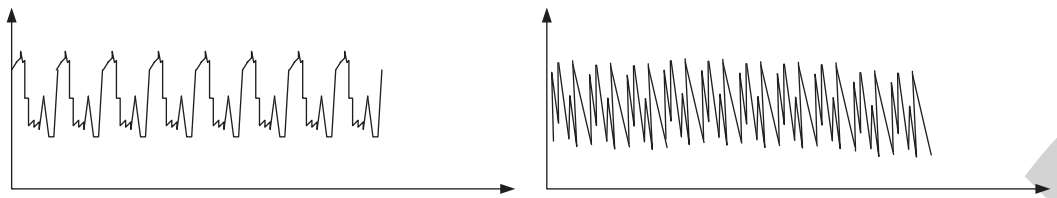

Figure 5: Black pipe and violin waveforms. (a) Black pipe time-domain waveform. (b) Violin time-domain waveform.

multiplied by scale $\delta_{(f_{\max},w,\sigma)^i}$ and subtracted from the residual s^{i-1} of the spectrum.

$$s^i = s^{i-1} - d_{(f_{\max},w,\sigma)^i}\delta_{(f_{\max},w,\sigma)^i}. \tag{15}$$

This iterative process involves two key issues: one is how to choose the basis vector $d_{(f_{\max},w,\sigma)^i}$, and the other is how to find its scale. For the first problem, since the harmonic structure is essential for audio classification, each harmonic basis vector can be represented as a linear sum of a set of instantaneous basis vectors. Therefore, the harmonic components should be extracted from the spectrum first, and then the nonharmonic components should be extracted. In addition, since the basis vector with the highest correlation with the residual spectrum reflects the most significant structure in the residual spectrum, in each iteration, the optimal basis vector should have a high correlation with the residual spectrum. Therefore, in each step of selection, a basis vector with a harmonic structure and a high correlation with the residual spectrum is selected.

To prevent the spectral aliasing phenomenon of the signals obtained after framing, there will be several repetition points in the two frames before and after the music signal, and only a small section in the middle is a different signal. At the same time, the signal needs to be windowed during the framing process. Moreover to prevent the occurrence of spectral aliasing, the window function can choose rectangular windows and Hamming windows. The framing process is shown in Figure 6.

A sparse representation of the audio spectrum can be obtained through the MP algorithm; that is, a set of selected basis functions and their scales are used to characterize the power spectrum of the original signal. Among them, the parameters of the basis function characterize the fundamental frequency, center frequency, and frequency multiplication attenuation rate of the spectrum, and the scale indicates the proportion of the basis function in the signal power spectrum. Thus, the mean value of each parameter value weighted by the scale indicates the mean property of the signal power spectrum. In addition, the variance of each parameter can characterize the distribution range of each parameter. Therefore, by using the combination of the weighted mean and variance, the distribution of each parameter can be roughly depicted, reflecting the characteristics of the signal power spectrum.

It is assumed that the parameter set of the basis function selected by the sparse representation is $(f_{\max}, w, \sigma)^i, 1 \le i \le I$, and I is the number of basis functions selected in the sparse representation. The scale set is δ^i, $1 \le i \le I$. Then, the weighted parameter mean values are, respectively,

$$\begin{aligned}
\overline{f}_{\max} &= \frac{\sum_{i=1}^{I}\delta^i f^i_{\max}}{\sum_{i=1}^{I}\delta^i},\\
\overline{w} &= \frac{\sum_{i=1}^{I}\delta^i w^i}{\sum_{i=1}^{I}\delta^i},\\
\overline{\sigma} &= \frac{\sum_{i=1}^{I}\delta^i \sigma^i}{\sum_{i=1}^{I}\delta^i}.
\end{aligned} \tag{16}$$

The variances of the parameters are, respectively,

$$\begin{aligned}
\Delta f_{\max} &= \frac{1}{I}\sum_{i=1}^{I}\left(f^i_{\max} - \overline{f}_{\max}\right)^2,\\
\Delta w &= \frac{1}{I}\sum_{i=1}^{I}\left(w^i - \overline{w}\right)^2,\\
\Delta \sigma &= \frac{1}{I}\sum_{i=1}^{I}\left(\sigma^i - \overline{\sigma}\right)^2.
\end{aligned} \tag{17}$$

The audio feature vector is represented as $[\overline{f}_{\max}, \overline{w}, \overline{\sigma}, \Delta f_{\max}, \Delta w, \Delta\sigma]$.

5. Extraction Experiment

This chapter mainly conducts simulation, verification, and analysis on the application performance of different classification models in musical instrument classification scenarios. Using the feature parameter extraction method in the previous chapter, the timbre feature parameter set of the existing music data source is extracted. Combined with pattern recognition technology, different classifiers are used for classification, training, and cross-validation of timbre feature parameter sets. And it compares and analyzes the classification results of different classification models, compares the advantages and disadvantages of each classifier, and lays the algorithm foundation for the design and implementation of the timbre analysis system. The overall flow of simulation analysis is shown in Figure 7.

After preparing the data source for the experimental simulation, first use the content of Chapter 2 to extract the timbre feature parameters of the data source in MATLAB and then convert the extracted timbre feature parameter

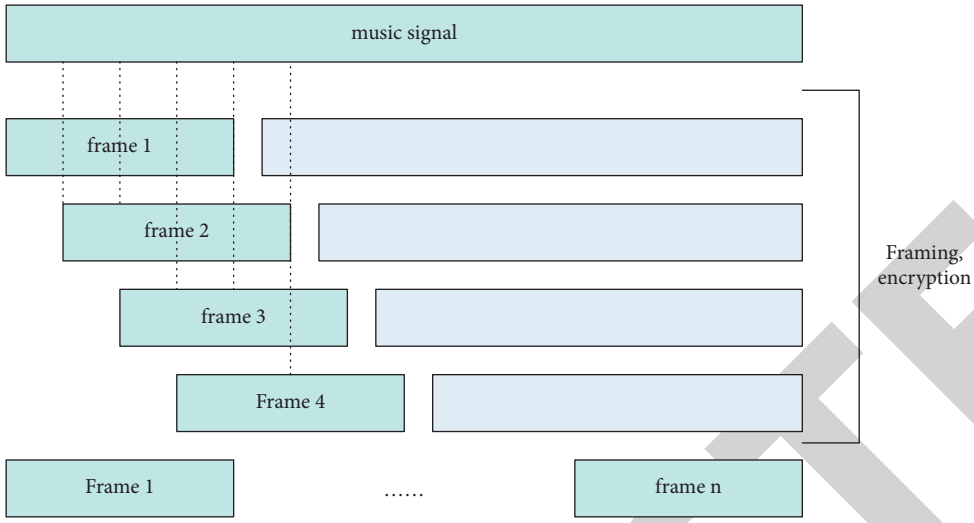

Figure 6: Framing process.

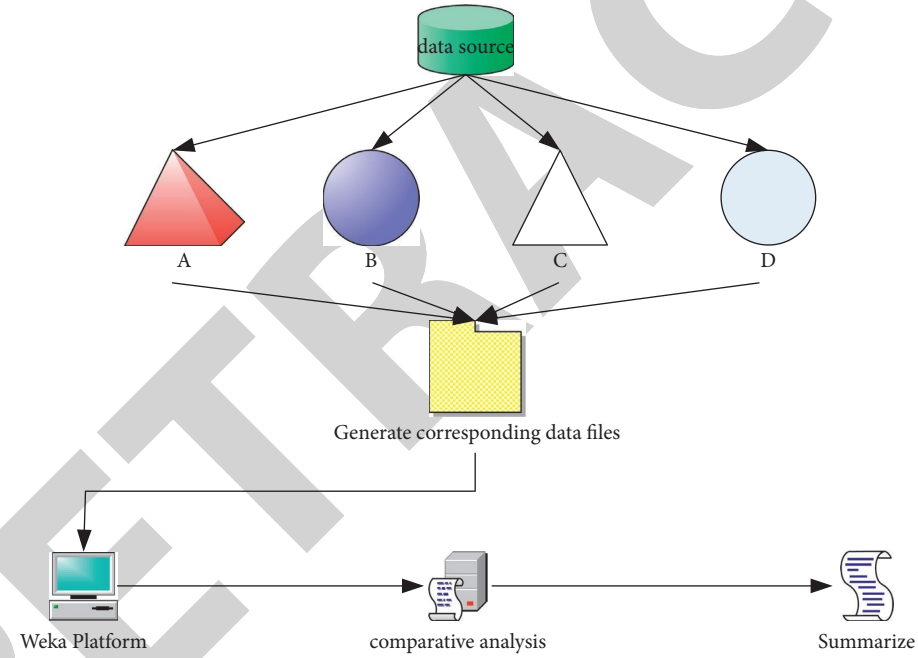

Figure 7: Overall flowchart of experimental simulation.

vector into the arf format that can be recognized by the Weka data mining tool text file. Based on the Weka platform, the feature parameter vector is classified, and some test data are provided to calculate the accuracy of the classification results. Finally, it is necessary to compare and analyze the classification results, summarize the advantages and disadvantages of different classification models, and lay the algorithm foundation for the design and implementation of the timbre analysis system in the next step.

5.1. Tone Feature Data Collection. To obtain comprehensive and accurate simulation results, the data sources used as learning samples and test samples in the simulation come from the Internet, recording equipment and musical instrument sound effects synthesized by software. All pieces are solo pieces played by a single instrument. To cover a wide range of music, this experiment selected eight instruments as sound sources, four of which were Western instruments: piano, violin, saxophone, and guitar; four oriental instruments are pipa, guzheng, flute, and erhu. In addition, among the eight musical instruments, violin, piano, guitar, pipa, guzheng, and erhu are percussion instruments, while flute and erhu are wind instruments, so that we can compare and analyze the timbre of musical instruments with similar pronunciation principles.

In the selection of music repertoire, the influence of performance style is also taken into account, and a considerable number of repertoires with fast, medium, and slow rhythms are selected for the repertoire played by each type of

instrument. In addition, to simplify the experimental data, this paper does not strictly distinguish the test data and the training learning samples, but inputs all data into the Weka tool platform for cross-validation.

The audio file format selected in this article is the audio in MP3 format, which is the most common audio file and has a suitable size. In addition, due to the limitation of practical conditions, this paper adopts the method of dividing a track into music segments ranging from 10 s to 30 s to expand the database, which has two advantages: one is to verify the effect of playing time on the classification results; the other is to eliminate the influence of other factors except timbre on the experiment. Of course, the method of segmenting music may have a certain impact on the accuracy of the experimental results, so there may be a certain deviation between the data obtained by the classification accuracy rate and the actual accuracy rate at the end. However, this does not affect the horizontal comparison between the various classifiers, nor does it affect the algorithm selection of the timbre analysis system design.

When slicing music files, it is necessary to eliminate the silent sections in the track, because these silent sections do not contain data, so it does not make any sense for this experiment. These silent segments need to be filtered, and this experiment uses endpoint detection to eliminate silent segments. The specific processing process of the endpoint detection algorithm is shown in Figure 8.

(A) Preprocess the music signal. The specific process is introduced in detail in the second chapter, namely, noise reduction, frame separation, etc.

(B) Calculate the short-term energy E of each frame.

(C) Calculate the zero-crossing rate (ZCR) of each frame.

(D) Set the threshold of short-term energy and ZCR, and the short-term energy and ZCR exceeding the threshold value are judged as valid signals.

(E) Delete the silent segment to obtain the output audio signal.

After deleting the silent segment, we use the MP3 cutting tool to cut the audio, and every 10 s-30 s there is a music segment. After cutting, the timbre, feature extraction can be performed.

A total of 950 pieces of music were used in this experiment, each piece being a solo piece for one instrument. A total of 8 musical instruments were collected in the experiment, including 4 Western musical instruments: piano, violin, saxophone, and guitar; 4 oriental musical instruments are pipa, guzheng, erhu, and dizi. Among the 950 pieces of music, there are 98 pieces for flute, 158 pieces for erhu, 122 pieces for piano, 122 pieces for guzheng, 100 pieces for guitar, 146 pieces for pipa, 132 pieces for saxophone, and 72 pieces for violin. The data distribution is shown in Figure 9.

The simulation analysis is carried out on the Weka data mining platform. The minimum distance classifier, decision tree classifier, SVM classifier, and BP neural network implemented by Weka are used to train and learn the timbre feature parameter vector set to establish different classification models. The analysis method of the classification model is carried out by means of cross-validation.

5.2. Simulation Based on Weka. Weka is a completely open-source data mining work platform, designed and implemented by the University of Waikato based on Java language. As a data mining work platform, Weka collects a large number of machine learning algorithms capable of data mining tasks, which can be used directly or invoked in their own Java code. Weka includes tools for data preprocessing, classification, regression, clustering, rule association, and visualization. Weka is also suitable for secondary development of machine learning algorithms on its basis.

The datasets that Weka can handle are datasets in the form of two-dimensional tables. The row of the two-dimensional table represents the instance. In this experiment, one row represents the timbre feature vector (29 dimensions) of a piece of music; the columns of the two-dimensional table represent attributes, and Weka mines the relationship between attributes. The file format that Weka can handle is ARFF (Attribute-Relation File Format) file. In this experiment, after the timbre feature parameter vector set is extracted, it needs to be converted into the corresponding ARFF file that Weka can recognize.

The timbre feature parameter of a piece of music is a 29-dimensional feature vector, so when it is converted into an ARFF file, each feature vector has a total of 30 columns, that is, 30 attributes, because it is necessary to add a one-dimensional column of instruments to the feature vector. In this way, the ARFF file finally generated in this experiment is a $950 * 30$ two-dimensional dataset. The latest Weka3-6 requires running in a JVM with jdkl. 7 or above, and a Java environment needs to be configured before simulation with Weka.

5.3. Summary of Simulation Results. The experimental simulation of training and learning is carried out on the musical instrument timbre dataset, and four different audio feature extraction methods are used to classify musical instrument timbres, namely, MFCC and spectral features (SF). The HCE features proposed in this paper and the spliced MFCC + HCE features are denoted as A, B, C, D for convenience in the text. The two groups of feature parameters with different dimensions are compared and analyzed. Group A is the sample containing only the MFCC parameters, and group B is the sample containing all feature parameters. The simulation results show that the classification accuracy of group B data is higher than that of group A, which shows that the parameters selected in this paper are all effective feature parameters. In this section, the simulation results of group B will be analyzed in detail. The accuracy of the cross-validation results of the classifiers implemented by different algorithms is shown in Table 3.

Analyzing the overall classification situation, it can be seen from Figure 10 that the classification accuracy of the four feature extractors is above 92%. Among them, B has the worst effect, with an accuracy rate of 92.42%, while D has the best classification effect, with an accuracy rate of 99.15%. For

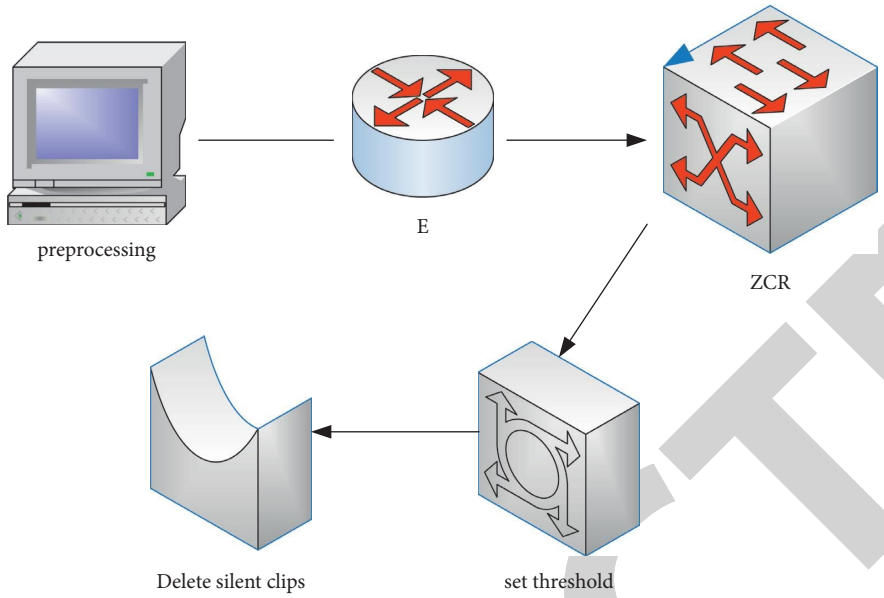

Figure 8: Silence removal process.

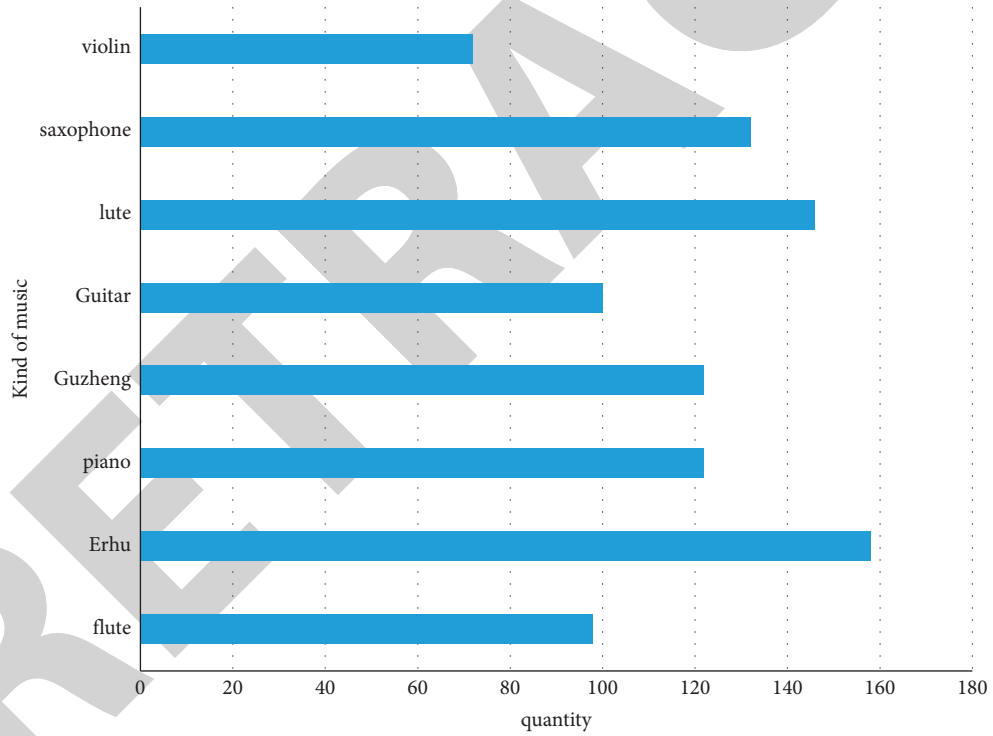

Figure 9: Distribution map of timbre feature dataset.

a single instrument, all 100 guitar samples are correctly classified in the four classifiers, indicating that the guitar timbre has obvious characteristics and is easy to distinguish; there are a few errors in the classification of piano and flute timbres in the four classifiers, as can be seen from the confusion matrix of each classification result in the previous section. Most of the wrong types of piano timbres are classified as guzheng, and the wrong classification of guzheng is also classified as piano, which shows that the timbres of piano and guzheng are relatively similar and easy to be confused. The error situation of the flute is more complicated, indicating that the sound of the flute and valid information have not been found. The violin and saxophone are also easy to be confused, indicating that the tone of the violin and the saxophone has something in common, and the two can bring a similar auditory experience.

Different feature extraction models have different accuracy rates, which shows that the selection of the classification model has a great influence on the classification results. As can be seen from Table 3 and Figure 10, *D* has the

Table 3: Implementation of different feature extraction algorithms.

Musical instrument		A	B	C	D
Flute	Correct	91	86	89	94
	Mistake	7	12	9	4
Erhu	Correct	154	153	155	158
	Mistake	4	5	3	0
Piano	Correct	116	110	111	118
	Mistake	6	12	11	4
Guzheng	Correct	121	112	111	122
	Mistake	1	10	11	0
Guitar	Correct	100	100	100	100
	Mistake	0	0	0	0
Lute	Correct	144	123	137	146
	Mistake	2	23	9	0
Saxophone	Correct	132	128	132	132
	Mistake	0	4	0	0
Violin	Correct	66	66	72	72
	Mistake	6	6	0	0
Summary	Correct	924	878	279	942
	Mistake	26	72	141	8

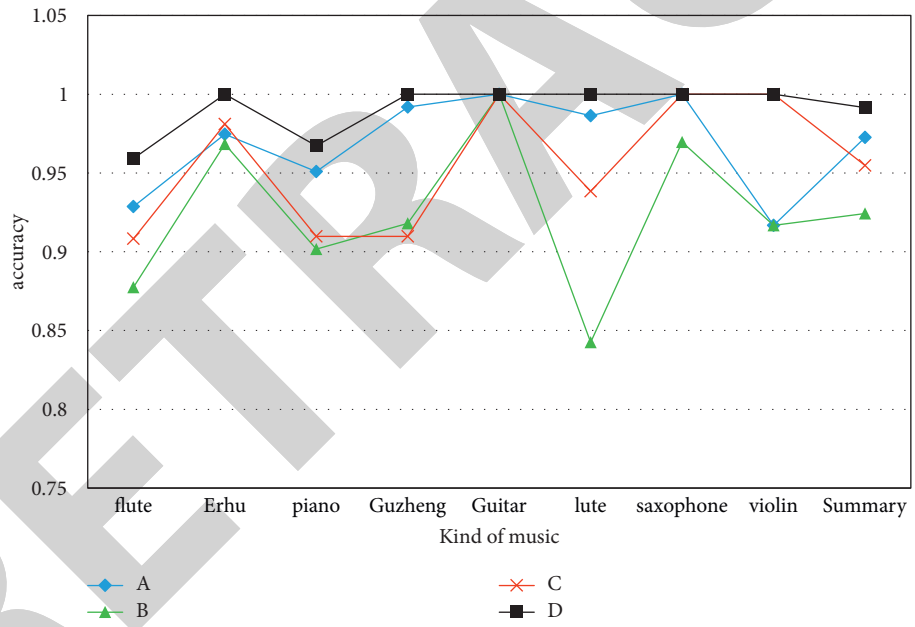

Figure 10: Accuracy achieved by different feature extraction algorithms.

best classification effect, with an accuracy rate of 99.15%, but its time consumption is relatively long, while the second algorithm A has certain advantages in time consumption. For these two classification algorithms, after a trade-off comparison, the classifier implemented by the *D* algorithm is the most suitable for timbre classification. There are two main reasons: first, although the time overhead of the MFCC + HCE feature extraction algorithm is relatively small, its *K* value is an empirical value, and there is currently no scientific method for selecting the *K* value. In the process of training and learning, repeated parameter adjustment is required to ensure the accuracy of the classification model, and the *C* algorithm is not practical when the number of classification categories increases. Although the MFCC + HCE feature extraction algorithm takes a long time, the parameter adjustment of this algorithm is mainly to adjust the weight of the feature parameter vector itself, so it has a similar value method for the same mode, which is more stable than the C algorithm. Second, the generation of the classification model itself can be calculated as an offline method, the result is more important than the process, and the offline calculation is not sensitive to the time overhead. Considering the above two points, the classifier implemented by the MFCC + HCE feature extraction algorithm is the most suitable for musical instrument classification.

6. Conclusions

Musical instrument recognition is one of the important issues in the field of audio information retrieval, which mainly includes feature extraction and classification algorithms. From the perspective of timbre, this paper studies the identification of Western musical instruments. The main work can be summarized as follows: this paper introduces four kinds of mathematical models commonly used in musical sound signal research. The excitation source filter model is based on the sounding mechanism of the musical instrument, and the musical tone signal is modeled as the convolution of the excitation signal and the resonant body, which greatly simplifies the complex musical tone performance process. The sine-plus-noise model utilizes the spectrum analysis characteristics of human hearing and regards the spectrum of the musical sound signal as the superposition of the sine component and the noise component. The FM-AM model successfully describes the subtle changes in musical tone during performance. After summarizing and analyzing the existing models, this paper finds that an appropriate mathematical model can provide not only a concise and intuitive description for musical tone signals, but also an important basis for timbre research.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] D. Kim, T. T. Sung, S. Cho, C. B. Sohn, and G. Lee, "A single predominant instrument recognition of polyphonic music using CNN-based timbre analysis," *International Journal of Engineering & Technology*, vol. 7, no. 3, pp. 590–593, 2018.

[2] K. Tatar, D. Bisig, and P. Pasquier, "Latent timbre synthesis," *Neural Computing & Applications*, vol. 33, no. 1, pp. 1–18, 2021.

[3] A. Banerjee, S. Sanyal, S. Roy, D. Ghosh, and S. Nag, "A novel study on perception cognition scenario in music using deterministic and non-deterministic approach," *Physica A: Statistical Mechanics and Its Applications*, vol. 567, no. 6, Article ID 125682, 2020.

[4] D. Rossetti and J. Manzolli, "Analysis of granular acousmatic music: representation of sound flux and emergence," *Organised Sound*, vol. 24, no. 2, pp. 205–216, 2019.

[5] R. V. Sharan, H. Xiong, and S. Berkovsky, "Benchmarking audio signal representation techniques for classification with convolutional neural networks," *Sensors*, vol. 21, no. 10, p. 3434, 2021.

[6] K. C. Santosh, S. Borra, A. Joshi, and N. Dey, "Preface: special section: advances in speech, music and audio signal processing (articles 1-13)," *International Journal of Speech Technology*, vol. 22, no. 2, pp. 293-294, 2019.

[7] D. Baxter, "For audio monitoring, seeing is believing," *Tv Technology*, vol. 37, no. 8, pp. 12-13, 2019.

[8] A. Thuring, K. Källén, K. J. Brännström, T. Jansson, and K. Maršál, "Doppler audio signal analysis as an additional tool in Evaluation of umbilical artery circulation," *Ultraschall in der Medizin*, vol. 38, no. 5, pp. 549–555, 2017.

[9] S. Chakraborty, "Music similarity and retrieval: an introduction to audio-and web-based strategies," *Computing Reviews*, vol. 58, no. 5, p. 270, 2017.

[10] S. Joshi and P. Sensarma, "Hybrid controller for mid-power audio application," *IET Power Electronics*, vol. 10, no. 10, pp. 1200–1207, 2017.

[11] W. Liang, H. Tong, B. Li, and Y. Li, "Feasibility research on break-out detection using audio signal in drilling film cooling holes by EDM," *Procedia CIRP*, vol. 95, no. 4, pp. 566–571, 2020.

[12] J. Zhihao, "Simulation of ocean surface temperature based on audio signal collection and accuracy of trade English translation," *Arabian Journal of Geosciences*, vol. 14, no. 16, pp. 1–15, 2021.

[13] N. Engebretsen, "Minding the gap: conceptualizing "perceptualized" timbre in music analysis," *Leonardo Music Journal*, vol. 30, no. 1, pp. 14–17, 2020.

[14] E. Thoret, B. Caramiaux, P. Depalle, and S. McAdams, "Human dissimilarity ratings of musical instrument timbre: a computational meta-analysis," *Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1745-1746, 2018.

[15] K.-Y. Sung, "An analysis of timbre comparison between jeongak daegeum and sanjo daegeum," *Journal of the Korea Entertainment Industry Association*, vol. 14, no. 3, pp. 229–236, 2020.

[16] Z. Wallmark, "A corpus analysis of timbre semantics in orchestration treatises," *Psychology of Music*, vol. 47, no. 4, pp. 585–605, 2019.

[17] P. N. Simes, D. Lüders, M. R. Jose, C. M. D. Araujo, and G. Romanelli, "Musical perception assessment of people with hearing impairment: a systematic review and meta-analysis," *American Journal of Audiology*, vol. 30, no. 1, pp. 1–16, 2021.

[18] L. Velardi, J.-P. Hermand, and R. D'Autilia, "On timbre in urban soundscapes: the role of fountains," *Journal of the Acoustical Society of America*, vol. 141, no. 5, p. 4017, 2017.

[19] H. C. Song and S. Kim, "Exploration of perceptual differences of virtually enhanced sound fields using timbre toolbox," *Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2499-2500, 2017.

[20] T. I. Minaeva, E. L. Ovchinnikov, and S. S. Yashin, "Tone timbre as hearing quality: visualization OF the condition and dynamic pattern," *Science and Innovations in Medicine*, vol. 3, no. 1, pp. 59–65, 2018.

[21] C. Chauhan, P. M. SiNgRu, and R. Vathsan, "Vibro-acoustic modeling, numerical and experimental study of the resonator and its contribution to the timbre of Sarasvati veena, a South Indian stringed instrument," *Journal of the Acoustical Society of America*, vol. 149, no. 1, pp. 540–555, 2021.

[22] G. Homer, "An empirical study of the timbre differences between gut core and metal core violin "A" strings," *American String Teacher*, vol. 23, no. 3, pp. 36-37, 2017.