



Research on sound classification based on SVM

Pengcheng Wei¹ · Fangcheng He² · Li Li¹ · Jing Li¹

Received: 20 January 2019 / Accepted: 29 March 2019 / Published online: 16 April 2019

© Springer-Verlag London Ltd., part of Springer Nature 2019

Abstract

Sound is a ubiquitous natural phenomenon that contains a wealth of information that constantly enhances our understanding of the objective world. With the continuous development of computer network technology and communication technology, audio information has become a very important part. Audio is a non-semantic symbolic representation and an unstructured binary stream. Because the audio itself lacks the description of content semantics and structured organization, it brings great difficulty to the audio classification work. The research of digital audio classification will become more and more important with the increasing number of digital audio resources in the network. Digital audio classification technology is the key technology to solve this problem. It is the key to solve the problem of audio structure and extract audio structured information and content semantics. It is a research hot spot in the field of audio analysis. It has important application value in many fields, such as audio retrieval, video summary and auxiliary video analysis. This paper studies the structure of audio, the analysis and extraction of audio features, the digital audio classifier based on support vector machines (SVM) and the audio segmentation technology based on BCI. SVM is an important achievement of machine learning research in recent years. As a new machine learning method, SVM can solve practical problems such as small sample, nonlinearity and high dimension, so it has become a new research hot spot after the study of neural network. Experiments show that the SVM-based audio classification algorithm has good classification effect, and the smoothed audio segmentation results are more accurate. With the further development of the research, the research results will be well applied in practice.

Keywords Support vector machine · Audio segmentation · Audio classification · Audio signal preprocessing

1 Introduction

Today's human society has entered the era of digitalization. With the continuous development of computer technology, network technology and communication technology, multimedia information such as images, video and audio have gradually become the main form of information media in the field of information processing. Among them, audio has a very important position. Audio is an important part of multimedia. Compared with images and video, audio not only has distinctive features, but also has a small amount of audio data and fast processing speed, which has attracted people's attention. There are many forms of audio expression, which meet people's needs in life, work, entertainment, etc. The audio data resources on the Internet continue to grow at an unparalleled speed. If people want to quickly and efficiently obtain and process the effective information they need from the massive audio data on the Internet, it is a good and convenient way to

✉ Pengcheng Wei
gxwpc@cque.edu.cn
Fangcheng He
wpc75@cque.edu.cn
Li Li
cqwpc@cque.edu.cn
Jing Li
weipc@cque.edu.cn

¹ School of Mathematics and Information Engineering, Chongqing University of Education at Nanshan, No. 1 Chongjiao Rd. Nanshan Avenue, Nanan District, Chongqing, China

² College of Foreign Languages Literature, Chongqing University of Education at Nanshan, No. 1 Chongjiao Rd. Nanshan Avenue, Nanan District, Chongqing, China

analyze, classify and retrieve the data. How to effectively organize and manage these audio resources making it easy for people to find the required audio clips has become an urgent need.

Now, research on audio classification problems has more than just classification of music and speech. The categories of classifications will change with people's needs, facilitating people's work and life. In general, the most basic objects of audio classification are voice, music and mute, further divided into five categories: pure voice, music, ambient sound, voice with background sound and mute. Audio classification is the basis of audio information deep processing, the core technology of audio structure and an important means to extract audio structure and content semantics. It actually divides the audio data into different categories according to the perceived characteristics or the content of the expression and can also play an important role in voice retrieval, content-based audio segmentation and audio supervision. On the one hand, it can be used as an initialization process for continuous speech recognition, which can prohibit non-voice streams in the audio stream from entering the speech recognizer, improve the accuracy of speech recognition and shorten the recognition time. On the other hand, it is also the first step in the classification of music types. For a given piece of audio, we can classify and segment it by audio classification. After the judgment, different processing is performed for different types of audio data for the judgment result. This can reduce the processing time and space consumption and can also improve the processing accuracy. At present, research in this field mainly focuses on three aspects: audio feature analysis and extraction, classifier design and implementation and audio segmentation methods.

The classification of audio can be said to be a process of pattern recognition. Its research focus usually includes two basic aspects: audio feature analysis and extraction, design and implementation of classifier. The essence of audio classification is actually the pattern recognition process, which mainly achieves the following: (1) Pretreatment. Before processing the audio file, we need to preprocess it, which is to divide the audio stream into smaller units. Audio files are classified by classifying these shorter length audio units. The preprocessing of the audio signal includes pre-emphasis, framing and windowing. (2) Extract audio features for classification. The selection and extraction of features are the most important part of the pattern recognition system, and of course the most important part of audio classification. (3) Feature screening. Multi-class audio classification, multi-level two classification, in order to better distinguish the two types of audio data of each level, the feature selection method will be used to select the feature set that is most suitable for each level classification. (4) Select the classifier. Using machine learning to

automatically classify audio signals not only reduces manpower, but also reduces time and efficiency. The implementation of commonly used audio classifiers is mainly divided into two categories: threshold-based and statistical-based models.

In the field of audio classification, the early implementation of the classifier implementation method is based on thresholds. This classification method requires a large amount of training data, and since the thresholds selected in different applications are generally different, it is not universal, and the threshold judgment method can only realize the classification on the audio coarse level (such as classification music, mute, voice, etc.), and it cannot realize the fine classification of audio data (such as recognition of applause, shouting, explosion sound). Therefore, in order to overcome these shortcomings, people proposed audio classification based on statistical models. There is no threshold in this classification method, which is a classification model obtained through data training on the basis of statistical theory. It not only recognizes audio data on the coarse level, but also recognizes fine-level audio data.

Many researchers have done a lot of work in this field and proposed different audio features and classification methods. There are two main problems: First, most of these studies use relatively simple features, and the classification problem is also relatively simple, usually only the classification of speech and music. The classification accuracy is satisfactory in simple classification, but if the classification object is increased, such as adding environmental sounds, non-pure speech or taking smaller windows, only simple features are used for classification, and the precision is very low. Second, the conventional audio classification algorithm mainly adopts a rule-based classification algorithm, that is, determines the category to which the audio belongs according to one or several audio features and their threshold values. However, this method has some shortcomings. For example, the decision rules and classification order are not necessarily optimal; the upper layer decision errors will accumulate to the next layer and form a "snowball" effect; the classification error is large and requires human test analysis, in particular, the determination of the threshold. Therefore, rule-based classification algorithms are difficult to meet different applications under different conditions.

In the statistical model, there are also the distinction between the supervised model and the unsupervised model. In the early days, people often used supervised data analysis and classification methods, such as support vector machine (SVM). SVM is a new machine learning method based on statistical learning theory [1, 2], which is suitable for processing classification and reflects the differences between categories to a greater extent. The SVM method fully demonstrates its effectiveness in many

applications. However, the effectiveness of the SVM method has a strong dependence on the quality and quantity of training data. A good classifier determines a high classification accuracy, and a classifier adapted to the target according to the classification target of the classified audio data contributes to an improvement in the classification accuracy. The statistical model has very good ability to simulate the spatial distribution of features of sound and has good robustness. Therefore, in recent years, support vector machine (SVM) has been widely used in audio classification.

Audio segmentation, also known as hopping point detection, as the name implies, is to find the hopping point in the audio sequence to be tested by some means. So what kind of point is called a jump point? In general, when the human ear receives a continuous stream of audio signals, different signals give different senses. From a perceptual point of view, when the human ear feels a signal change, this point is called a jump point, also called a dividing point. From a signal perspective, this change can be referred to as a change in the auditory characteristic, as a certain characteristic of the corresponding signal must change with this change. The process of segmenting out audio segments of varying lengths is known as audio segmentation.

In the current multimedia information processing, audio occupies a very important position, but due to the characteristics of the media source itself and the constraints of the prior art, the further analysis and utilization of the audio information is limited. The audio classification and segmentation technology can solve this problem well, providing a solid foundation for audio structuring and deep analysis and utilization of audio information.

2 Proposed method

2.1 Audio signal preprocessing

The audio signal preprocessing is divided into two steps: Firstly, the original audio signal is preprocessed and the main purpose is to unify the audio format, perform pre-emphasis, divide the audio signal into audio segments and perform windowing and framing for each audio segment; Secondly, the extracted audio frames and audio segments are extracted, and the extracted features are merged. The main purpose is to obtain the final required audio feature vectors. Preprocessing raw audio data, including pre-emphasis, segmentation and windowing.

(1) Pre-emphasis processing

Combined with the human ear hearing mechanism, the audio frequency range that can be heard by the human ear

is 60 Hz–20 kHz. When audio signal processing is performed, the audio signal is pre-emphasized, and its purpose is to eliminate low-frequency interference, especially 50 Hz or 60 Hz power–frequency interference. Pre-emphasis is generally implemented by digitizing the audio signal with a pre-emphasis digital filter, which is typically a first-order high-pass digital filter:

$$H(z) = 1 - \mu z^{-1} \quad (1)$$

In terms of time domain, if the processed signal is $y(n)$, then $y(n)$ can be expressed as:

$$y(n) = x(n) - \mu * x(n - 1) \quad (2)$$

In the formula, the pre-weighting coefficient μ is taken as [0.97, 0.98].

Where $x(n)$ is the original signal sequence and $y(n)$ is the pre-emphasized sequence.

The first-order high-pass digital filter, as shown in Fig. 1.

By pre-emphasis processing, the effect of sharp noise can be reduced, and the high-frequency portion of the signal can be boosted, which makes the spectrum of the signal flat, and the pre-emphasis coefficient is usually about 0.97 or 0.98. The signal that is pre-emphasized by the filter needs to be normalized.

(2) Windowed framing

After the pre-emphasis digital filtering process is performed, the windowing and framing processing is performed next. The audio signal characteristics change very slowly over a short period of time, so the extracted audio features remain stable during this slow transition. Thus, when processing an audio signal, the discrete audio signal is first divided into a unit of length for processing; that is, the discrete audio sample points are divided into audio frames. This method is a signal “short-time” processing method. Generally, a “short-time” audio frame has duration of about several to several tens of milliseconds. According to the length of the divided audio unit, we can divide the audio unit into audio frame, audio clip, audio shot, audio high-level semantic unit. Although the framing can adopt the method of continuous segmentation, the method of overlapping segments is generally adopted, in order to make a smooth transition between frames and frames and maintain its continuity. The overlapping portion of the previous frame and the next frame is called frame shift, and the frame shift is often taken as half of the frame

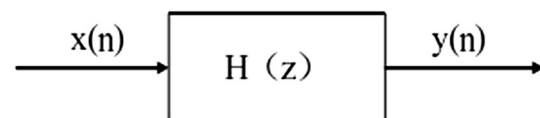


Fig. 1 Schematic diagram of the pre-emphasis filter

length. Framing is implemented by weighting a finite length window that can be multiplied by $y(n)$ with a certain window function $w(n)$ to form a windowed audio signal $y_w(n) = w(n) * y(n)$. The signal in the time domain is multiplied, which is equivalent to the convolution calculation in the frequency domain. Therefore, the windowing calculation can also be expressed as follows:

$$Y_w(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} Y(e^{j\theta}) W(e^{j(\omega-\theta)}) d\theta \quad (3)$$

where Y and W represent the spectrum, respectively.

It can be seen that the window function $w(n)$ not only affects the waveform of the original signal in the time domain, but also affects the shape of its frequency domain. The two most commonly used window functions are the rectangular window and the Hamming window.

$$\text{Rectangular window: } w(n) = \begin{cases} 1, & 0 \leq n \leq (n-1) \\ 0, & n = \text{else} \end{cases} \quad (4)$$

Hamming window:

$$w(n) = \begin{cases} 0.54 - 0.46 \cos[2\pi n/(N-1)], & 0 \leq n \leq (n-1) \\ 0, & n = \text{else} \end{cases} \quad (5)$$

The choice of the shape and length of the window function $w(n)$ has a great influence on the characteristics of the short-term analysis parameters. Therefore, an appropriate window should be selected to make the short-term parameters better reflect the characteristic changes of the speech signal. The rectangular window has better spectral smoothness, but the high-frequency component is lost, the waveform detail is lost, and the rectangular window will cause leakage. The Hamming window can effectively overcome the leakage (Gibbs) phenomenon and has the widest application range. If the window length N is large, it is equivalent to a very narrow low pass filter. When the audio signal passes, the high-frequency portion reflecting the waveform details is hindered, and its short-time energy changes little with time. This does not truly reflect the amplitude variation of the speech signal. Conversely, if N is too small, the passband of the filter becomes wider, and the short-term energy changes sharply with time, and a smooth energy function cannot be obtained. Therefore, the length of the window should be chosen appropriately, generally with a duration of 15–30 ms. After the above processing, the audio signal has been divided into short-time signals of a frame-by-frame plus window function, and then, each short-term audio frame is regarded as a smooth random signal, and the digital signal technology is used to extract the audio characteristic parameters.

2.2 Audio feature analysis extraction

Audio signals contain a wealth of information, and there are many interfering signals and redundant information. How to extract the most representative information of the audio signal in the audio signal is crucial for audio classification. Audio features are the basis of audio classification, and the extracted audio features are to reflect the salient features of the audio to the greatest extent possible. At the same time, the impact on the environment should reflect good robustness, while eliminating the signal characteristics that cause recognition ambiguity [3]. The parameters extracted by the feature are used as input to the classification processing method in the form of vectors. Therefore, the independence between vector parameters should be considered, and the computational complexity should be minimized while ensuring the accuracy of the results. It has the characteristics of including as much information as possible, but the amount of data is as small as possible. Feature extraction of audio can be based on feature analysis and extraction of audio frames and feature analysis and extraction based on audio segments. The characteristics of the audio frame are analyzed by the audio frame, and the feature analysis and extraction of the audio segment is based on the characteristic parameters of the audio frame. The characteristics of audio include three aspects: time Domain features, frequency domain features and perceptual features.

- (1) *Time Domain Features*: There are two aspects of time domain features. The main indicators we use in audio frames are short-time energy and zero-crossing rate. The indicators used in the audio segment are mainly three indicators: mute ratio, low frequency energy ratio and high zero-crossing ratio.
- (2) *Frequency Domain Features* There are two aspects to obtaining frequency domain features after Fourier transform. The indicators used in audio frames are frequency domain energy, sub-band energy distribution, frequency centroid, bandwidth, pitch frequency, MFCC coefficient (Mel-frequency cepstrum coefficients). In the audio segment, we use indicators such as sub-band energy ratio mean, spectrum centroid mean, bandwidth mean, spectrum transition and MFCC coefficient mean.
- (3) *Perceptual Features*: Perceptual Features mainly have pitch in audio frame features, and the main features in audio segment are basic audio frequency standard deviation [4]. In this paper, the acoustic characteristics do not reflect the class characteristics of the audio well in the operation process, so we will not adopt it.

2.2.1 Audio time domain feature analysis and extraction

The audio time domain feature refers to a vector parameter representing a time domain feature extracted by analyzing an audio signal in units of frames on a time domain waveform.

Zero-crossing rate (ZCR) refers to the ratio of the number of points where the signal values of two adjacent sampling points in the discrete points of the audio signal are different from each other to the number of all sampling points in a frame. The zero-crossing rate [5] shows the frequency of signal zero crossings, and the zero-crossing rate is also a common audio feature.

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} |\text{sgn}[x(m+1)] - \text{sgn}[x(m)]| \quad (6)$$

$x(m)$ is the processed discrete audio signal.

Short-term average energy Short-term average energy is one of the commonly used audio characteristic parameters. It is a relatively intuitive feature that reflects the change of audio energy, which is directly related to the selection of window length N . If the value of N is too long, the change of the whole energy is relatively smooth while the difference is not reflected, but a window that is too narrow does not have a smooth energy function. Therefore, the choice of window is more important. In this paper, the Hamming window is chosen to have a good balance between the two. The short-term average energy can be calculated using the formula as shown in (7):

$$E_m = \frac{1}{N} \sum_m [x(n)w(n-m)]^2 \quad (7)$$

where $x(n)$ represents the n th signal value in the m th frame of the audio signal, and $w(n)$ is the window function previously described in the text. The short-time energy can be set to a threshold, and below the threshold, it can be judged as silent, so the short-time energy is mainly used to determine whether the audio signal is muted. The short-term energy rate can be used to judge whether the audio signal is a voice, music and noise category.

2.2.2 Audio frequency domain feature analysis and extraction

The frame is the smallest unit of the audio signal we process, calculates the feature value of each frame and then calculates the feature value at the slice level. There are usually several typical audio features at the frame level.

- (1) The MFCC coefficient: Mel-frequency cepstral coefficient [6] is an acoustic feature derived from the human auditory mechanism. Humans follow an approximate linear relationship to the perception of

the sound frequency range below 1000 Hz. The perception of the sound frequency range above 1000 Hz does not follow a linear relationship, but follows an approximate linear relationship on the logarithmic frequency coordinates. The Mel scale describes the nonlinear characteristics of the human ear's perception of frequency. The MFCC is a cepstrum parameter extracted in the Mel-scale frequency domain. This feature has a high recognition rate and good noise robustness.

The MFCC is derived from the research results of two auditory systems [7]. Firstly, human perception of a single tone is approximately proportional to the logarithm of the pitch frequency. The so-called Mel-frequency scale, whose value generally corresponds to the actual frequency logarithmic distribution relationship. In the Mel-frequency domain, people's perception of tones is linear. The relationship between the Mel frequency and the actual frequency can be approximated by the following formula:

$$\text{Mel}(f) = 2595 \log \left(1 + \frac{f}{700} \right) \quad (8)$$

Secondly, when two tones with similar frequencies are emitted at the same time, one can only hear one tone. Critical bandwidth refers to such a bandwidth boundary that makes the subjective feeling abrupt. When the frequency difference between the two tones is less than the critical bandwidth, the person will hear the two tones as one, which is called the shielding effect. The critical bandwidth calculation formula is as follows:

$$\text{BW}_c = 25 + 75 \left[1 + 1.4 \left(1 + \frac{f_c}{700} \right)^2 \right]^{0.69} \quad (9)$$

where f_c is the center frequency.

Therefore, a critical band filter bank can be constructed to mimic the perceptual characteristics of the human ear. The Mel frequency cepstral coefficient [8] (MFCC) is calculated by using the filter bank method in the spectrum. The audio frequency is divided into a series of triangular filter sequences. This set of filters is of equal bandwidth at the Mel coordinates of the frequency.

- (2) The frequency domain energy: the frequency domain energy formula is as follows:

$$E = \log \left(\int_0^{\omega_0} |F(\omega)|^2 d\omega \right) \quad (10)$$

where $F(\omega)$ is the FFT transform coefficient of the frame, and ω_0 is half of the sampling frequency. The frequency domain energy E is used to determine the silence frame. If the frequency domain energy of a

certain frame is less than the threshold, the frame is marked as a silence frame, otherwise it is a non-silence frame.

- (3) The sub-band energy ratio divides the frequency domain into four sub-bands, which are $[0, \frac{\omega_0}{8}]$, $[\frac{\omega_0}{8}, \frac{\omega_0}{4}]$, $[\frac{\omega_0}{4}, \frac{\omega_0}{2}]$, $[\frac{\omega_0}{2}, \omega_0]$, and calculate the distribution of energy of each sub-band. The calculation formula is as shown in Eq. (11):

$$D = \frac{1}{E} \int_{L_j}^{H_j} |F(\omega)|^2 d\omega \tag{11}$$

where L_j and H_j are the upper and lower boundary frequencies of the sub-bands. Different types of audio have different energy distributions in each sub-band interval. The frequency domain energy of music is relatively evenly distributed in each sub-band interval. In speech, energy is mainly concentrated in the 0th sub-band, about 80% or more.

- (4) Zero-crossing rate, in the case of discrete-time signals, adjacent samples with different algebraic symbols are said to have zero crossings. The zero-crossing rate is a speed that describes zero crossing and is a simple measure of the amount of signal frequency. The formula is given by Eq. (12):

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} |\text{sgn}[x(m+1)] - \text{sgn}[x(m)]| \tag{12}$$

where $x(m)$ is the discrete audio signal. ZCR is a more common audio feature.

- (5) Frequency centroid: the brightness of a frame is measured by the frequency centroid in a frame, and the calculation method is as shown in Eq. (13):

$$FC = \frac{\int_0^{\omega_0} \omega |F(\omega)|^2 d\omega}{\int_0^{\omega_0} |F(\omega)|^2 d\omega} \tag{13}$$

- (6) Bandwidth: bandwidth is an indicator of the frequency range of the audio. Calculate as in Eq. (14):

$$BW = \sqrt{\frac{\int_0^{\omega_0} (\omega - FC)^2 |F(\omega)|^2 d\omega}{\int_0^{\omega_0} |F(\omega)|^2 d\omega}} \tag{14}$$

- (7) The pitch frequency: The pitch frequency is a unit that measures the pitch. Pitch period detection methods can be broadly divided into three categories: time domain methods, frequency domain methods, and methods for summing the time domain and

frequency domain characteristics of signals. In general, the pitch period is estimated using a simpler peak clipping algorithm for the center clipping short-term autocorrelation function. The principle of the autocorrelation method is that the short-time autocorrelation function of speech has a large peak at an integral multiple of the pitch period, and the pitch period can be estimated as long as the position of the largest peak point is found. The steps for calculating the pitch period are as follows:

- (a) Preprocessing: The center clipping function (15) is used to clip the audio to reduce the effect of the formant. The clipping threshold L is determined by the peak amplitude of the speech signal, generally taking 60–70% of the maximum signal amplitude.

$$y(n) = C(x(n)) = \begin{cases} x(n) - L, & x(n) > L \\ 0, & |x(n)| \leq L \\ x(n) + L, & x(n) < -L \end{cases} \tag{15}$$

- (b) Computation of $y(n)$ and $y'(n)$ cross-correlation: In order to overcome the problem of large amount of short-term autocorrelation calculation, the autocorrelation of $y(n)$ after the center clipping of Eq. (15) is replaced by the cross-correlation of two signals. One signal is only $y(n)$, and the other signal is three-level quantization of $y(n)$ only to produce $y'(n)$, namely

$$y'(n) = C(x(n)) = \begin{cases} 1, & y(n) > 0 \\ 0, & y(n) = 0 \\ -1, & y(n) < 0 \end{cases} \tag{16}$$

Calculate cross-correlation using the following formula:

$$R(k) = \sum_{n=0}^{N-1-k} y(n)y'(n+k) \tag{17}$$

- (c) Find the pitch period: Select the maximum value of $R(k)$, which is recorded as R_{\max} . If $R_{\max} < c \cdot R(0)$ (c is the threshold), it is judged as unvoiced, so that its pitch period is 0; otherwise, the pitch period is the value of k when $R(k)$ is the maximum value R_{\max} , namely

$$N_p = \arg \max_{k_1 \leq k \leq k_2} R(k) \tag{18}$$

- (d) Post-processing: Some scattered pitch periods deviate significantly from the pitch period trajectory due to the presence of sound interference, bias in the pitch period estimation, etc. For the accuracy and convenience of post-processing, the median filtering technique is generally used to smooth the original

curve. Median filtering is a nonlinear process. It uses a sliding window to select a piece of data from the data sequence and then replaces the data with the median value of the data. As the window continually slides along the data sequence, it constantly draws a median value, which is the result of the filtering.

2.2.3 Feature analysis and extraction based on audio segments

The audio segment is larger than the audio frame unit. One audio segment generally contains several audio frames. Its characteristic source is to statistically divide the audio frames. The general calculation method is to calculate their mean, variance and standard deviation for the audio frames contained in the audio segment. The main audio segment features used in this chapter are

- (1) The mute ratio sets a threshold in the frequency domain energy. When the energy of the sample frame is less than this threshold, we call the frame a silence frame, otherwise it is a non-silence frame. Based on the audio segment, the proportion of the mute frame is the mute ratio, which can be expressed by the following formula (19).

$$r = \frac{M}{N} \tag{19}$$

The parameter M represents the number of silence frames in the audio segment, and the parameter N represents the number of all audio frames contained in the audio segment.

- (2) The sub-band energy ratio mean: Sub-band energy ratio mean [9] is the audio segment feature calculated by the sub-band energy ratio parameter, that is, the average value of the energy ratio of each frame sub-band in an audio segment. This feature is widely used in the research of signals.
- (3) The bandwidth mean and the spectral centroid mean: bandwidth mean are the average of the bandwidth of each frame in the audio segment, and the average value of the spectrum centroid is the average of the audio brightness of each frame in the audio segment.
- (4) The high zero-crossing ratio: The zero-crossing rate of speech is higher than the music. If a threshold is set, the proportion of audio frames in the audio segment that exceeds this threshold can be calculated. This ratio is called the high zero-crossing ratio (high ZCR ratio). The threshold is generally 1.5 times the average of the zero-crossing rate in the audio segment. The calculation formula of its

eigenvalue is as shown in (20):

$$\text{HZCRR} = \frac{1}{2N} \sum_{n=0}^{N-1} [\text{sgn}(\text{ZCR}(n) - 1.5\text{avZCR}) + 1] \tag{20}$$

The parameter N represents the total number of audio frames in the audio segment, and ZCR(n) represents the zero-crossing rate of the nth frame in the audio segment.

- (5) The low frequency energy ratio sets an energy threshold in an audio segment. Below this energy is called the low frequency energy frame. The ratio of the low frequency energy frame in an audio segment can be calculated. This ratio is called the low frequency energy ratio, referred to as LFER [10], is obtained by the following formula (21).

$$\text{LFER} = \frac{1}{2N} \sum_{n=0}^{N-1} [\text{sgn}(0.5\text{av}E - E(n) + 1)] \tag{21}$$

The parameter N is the total number of audio frames in the audio segment, and E(n) represents the frequency domain energy of the nth frame in the audio segment. The threshold in the formula is 0.5 times the average value of the energy in each frame in the audio segment.

- (6) The spectral transition spectrum transition is used to describe an average parameter of the spectral difference of each adjacent audio frame in an audio segment. The calculation formula is as shown in (22):

$$\text{SF} = \frac{1}{(N - 1) \times K} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} |\log(|\text{DFT}(n + 1, k)|) - \log(|\text{DFT}(n, k)|)|^2 \tag{22}$$

- (7) Base audio rate standard variance: In an audio segment, the pitch frequency of each frame is first calculated, and then their standard deviation is calculated using these pitch frequency parameters, which is a feature used to describe the range of the pitch frequency.
- (8) The constituent vector set of the feature vector set is divided into two parts, a 24-dimensional MFCC vector, and an 11-dimensional feature vector extracted by the audio segment. Because the difference between the feature vectors is relatively large, it needs to be normalized. However, after the normalization of the MFCC vector set, the experimental results are not improved well. Therefore, only the

segment features are normalized and processed. As shown in (23):

$$x'_i = (x_i - \mu_i) / \beta_i \tag{23}$$

Parameter x_i needs to be normalized input feature vector, μ_i is mean, β_i is variance, and x'_i is the feature obtained after normalization.

2.3 Audio classification method

Audio classification technology is essentially a pattern recognition technology [11]. The statistical learning method has the advantages of a solid theoretical foundation and a simple implementation mechanism and thus is adopted by most current audio classification systems. The statistical learning method requires a batch of training samples with category markers to be given in advance, and the classifier is generated through guided learning training, and then the samples to be classified of the test sample set are tested to measure the classification performance. Typical audio classification methods include minimum distance method [12], support vector machine, neural network [13], hidden Markov model [14] and decision tree [15].

2.3.1 SVM-based classification algorithm

Support vector machine (SVM) is a machine learning method based on VC dimension theory and structural risk minimization proposed by Cortes and Vapik [16] in 1995, and its performance is very good. It solves small samples and nonlinearities. And high-dimensional pattern recognition and other issues can show its own unique advantages. Simply put, the purpose of the support vector machine approach is to find an optimal classification hyperplane that can completely separate the two types of data at maximum intervals. SVM can have a good learning effect regardless of the two-category or multi-classification problem. The SVM method was originally used to solve the two-category problem. The basic principles in the second classification are explained in detail below (Fig. 2).

The training sample set is $X = \{x_1 \cdots x_n\}, X \in R^d$. The corresponding category is labeled $\{y_1 \cdots y_n\}, y_i \in \{1, -1\}$. Let the dimension of the training sample feature vector be d and the number of samples be n .

(1) Linear support vector machine

For linearly separable problems, the dichotomous problem can construct a classification hyperplane so that positive and negative samples can be completely separated. As shown in Fig. 3. The solid sample points on the left represent positive samples, and the hollow sample points on

the right represent negative samples. There are several classification planes between H_1 and H_2 , all of which are able to completely separate the positive and negative samples. If one of the classification faces can not only completely separate the positive and negative samples, but also maximize the geometric spacing, then this classification line is called the optimal classification hyperplane. The so-called geometric spacing is the distance between H_1 and H_2 . H is the classification plane, and H_1 and H_2 are straight lines parallel to H and simultaneously passing through the two types of samples closest to the distance H . The sample points that happen to fall on H_1 and H_2 are the support vectors we are talking about. It is these support vectors that together build the optimal classification hyperplane. Assume that the linear discriminant function is $g(x) = wx + b$. By normalization, $\{x_1 \cdots x_n\}$ satisfies $g(x) \geq 1$, and at this time, the classification interval is $2l/||w||$.

$$y_i[wx_i + b] - 1 \geq 0, \quad i = 1, \dots, n \tag{24}$$

When the formula (23) is established, this classifier can correctly label all samples. Obviously, maximizing the classification interval is actually minimizing $||w||$. Therefore, the optimal classification hyperplane should both satisfy Eq. (24) and minimize $||w||$. The support vector machine is a sample of the formula (25). In summary, the problem of solving the optimal classification hyperplane is equivalent to the following constraint optimization problem:

$$\begin{aligned} \min \quad & ||w||^2 / 2 \\ \text{s.t.} \quad & y_i[wx_i + b] - 1 \geq 0, \quad i = 1, \dots, n \end{aligned} \tag{25}$$

In this way, the solution of SVM is finally transformed into solving the quadratic programming problem, so theoretically the solution of SVM is the globally unique optimal solution. First, construct a Lagrangian function:

$$\begin{aligned} \text{M arg in} \quad & 2l||w|| \\ L(w, a, b) = \quad & \frac{1}{2}||w||^2 - \sum_{i=1}^n a_i y_i (x_i \cdot w + b) \\ & + \sum_{i=1}^n a_i, a_i \geq 0, \quad i = 1, 2, \dots, n \end{aligned} \tag{26}$$

In the formula, a_i is the Lagrangian factor, and then we, respectively, differentiate the w and b in the above formula and make them equal to 0, and get $w = \sum_i a_i y_i x_i$ and $\sum_i a_i y_i = 0$ to convert the original optimization problem into a dual problem:

Fig. 2 Mel-scale filter bank

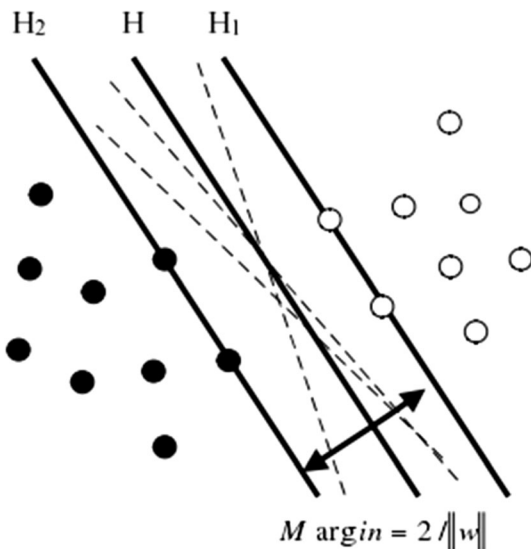
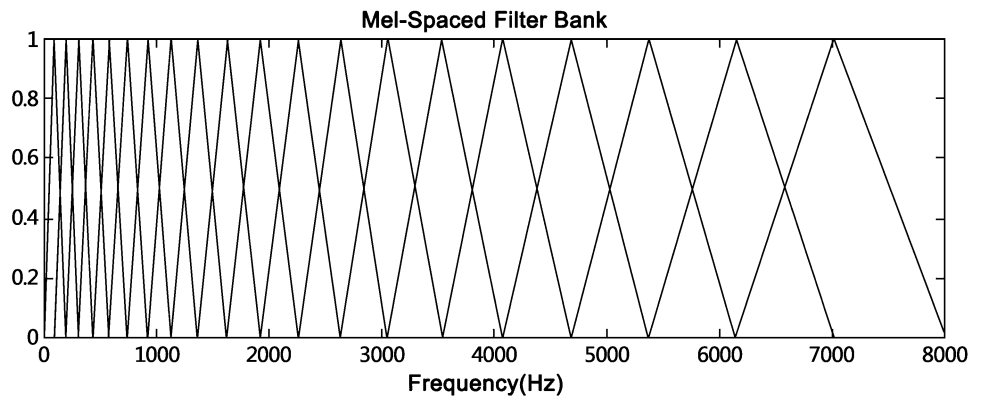


Fig. 3 Support vector machine (SVM) schematic

$$\begin{aligned} \max \quad & W(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j a_i a_j (x_i \cdot x_j) \\ \text{s.t.} \quad & \sum_{i=1}^n y_i a_i = 0 \quad a_i \geq 0, \quad i = 1, 2, 3, \dots, n \end{aligned} \tag{27}$$

Solving the above formula can obtain the corresponding a_i value of each sample, and the obtained solution is the optimal solution of the optimization problem. Only the samples corresponding to a_i that are not 0 are support vectors. Usually only a small part of the samples have a_i not 0. The final classification function discriminant is as follows:

$$f(x) = \text{sgn} \left[\sum_{i=1}^n a_i^* y_i (x_i \cdot x) + b \right] \tag{28}$$

$$b = \frac{1}{2} \left[\sum_{i=1}^n a_i y_i x_i \cdot x_r + \sum_{i=1}^n a_i y_i x_i \cdot x_s \right] \tag{29}$$

The b calculated by the above formula is the skew amount. When a_i^* in the formula is not 0, x_r and x_s represent any pair of support vectors in the two types of samples.

In reality, it is often because of the influence of noise that the classification samples cannot be separated linearly, and thus an uncorrected classification hyperplane cannot be obtained. The noise here can be considered as the rightmost black point in Fig. 4. It is obviously a sample of the negative class. This strange sample makes the linearly separable problem linear and inseparable. Usually this kind of problem is called “Approximate linear separability.” For this kind of problem, our usual treatment method is that the sample point is originally the user who accidentally mislabeled the sample, which is interference, noise, and should be ignored. But its existence does cause the problem to be unsolvable, so for this situation, we deal with a method that allows a small number of sample points to the distance of the classification hyperplane which does not have to meet the original requirements. That is to say, we originally require that all sample points to the classification

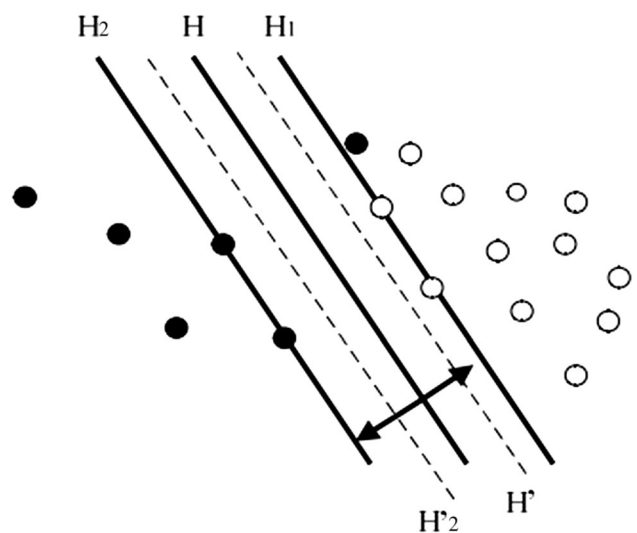


Fig. 4 Singularity

hyperplane should be at least greater than 1 interval. Now add fault tolerance and allow a hard threshold to be added to a hard variable, which allows some sample points to fall within the geometric interval, the expression becomes the following form:

$$y_i[wx_i + b] \geq 1 - \xi_i$$

$$\xi_i \geq 0, \quad i = 1, \dots, n \tag{30}$$

The slack variable is nonnegative; that is, the final result is that the sample interval is allowed to be smaller than 1. When the interval between the sample points is calculated to be less than 1, it means that the classifier gives up the exact classification of these singular points. Although this itself will cause some loss to the classifier, it also allows the classified hyperplane to be moved to these sample points without being affected by these few sample points, resulting in a larger geometric spacing. So there is a need for multiple weightings between the two.

Knowing that $\|w\|^2$ is the objective function, expecting its value to be as small as possible, so the loss should be an amount that makes $\|w\|^2$ larger. There are usually two ways to measure loss, the first is a second-order soft-interval classifier:

$$\sum_{i=1}^n \xi_i^2 \tag{31}$$

The other is a first-order soft-interval classifier:

$$\sum_{i=1}^n \xi_i \tag{32}$$

Adding a loss to the objective function requires a penalty factor, so the original optimization problem can be written as follows:

$$\min \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

$$\text{s.t.} \quad y_i [wx_i + b] \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad i = 1, 2, 3, \dots, n \tag{33}$$

(2) Nonlinear support vector machine

The basic principles of the support vector machine in solving the linear separability problem and the “approximate linear separability problem” are introduced. But in the real world, many times, in the original low-dimensional sample space, the sample is extremely inseparable. No matter how to find the classification hyperplane, there are always many singular points that do not meet the requirements. At this time, it is necessary to map the linearly inseparable sample data in the low-dimensional space to

the high-dimensional space. Although the mapping is not completely linearly separable after the mapping, it is at least “approximate linear separable.” Then with the slack variable to deal with a small number of singular points, you can achieve very good results. Mapping a sample from a low-dimensional space to a high-dimensional space needs to be implemented by means of a kernel function, so that the kernel function is:

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{34}$$

The kernel function itself must satisfy the Mercer condition. Its basic function is to input the vector in two low-dimensional spaces and then calculate the vector inner product value of a transformed high-dimensional space. So the original problem can be converted into the following form:

$$\max \quad W(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j a_i a_j K(x_i \cdot x_j) \tag{35}$$

$$\text{s.t.} \quad \sum_{i=1}^n y_i a_i = 0 \quad 0 \leq a_i \leq C, \quad i = 1, 2, 3, \dots, n$$

The discriminant function becomes

$$f(x) = \text{sgn} \left[\sum_{i=1}^n a_i^* y_i K(x_i \cdot x) + b \right] \tag{36}$$

$$b = \frac{1}{2} \left[\sum_{i=1}^n a_i y_i K(x_i \cdot x_r) + \sum_{i=1}^n a_i y_i K(x_i \cdot x_s) \right] \tag{37}$$

(3) Introduction to the kernel function

The kernel function makes the support vector machine perform well when dealing with nonlinear separable problems. The nonlinear classifiers constructed by different kernel functions are also different. When dealing with practical problems, there is currently no guiding principle for the selection of kernel functions. More needs to be verified by experiments to select the best kernel function. Commonly used kernel functions are as follows:

(a) Linear kernel function:

$$K(x, x_i) = (x_i \cdot x) \tag{38}$$

(b) Polynomial kernel function [17]:

$$K(x, x_i) = [p(x_i \cdot x) + s]^q \tag{39}$$

(c) Sigmoid kernel function [18]:

$$K(x, x_i) = \tan h(\mu(x_i \cdot x) + c) \quad (40)$$

(d) Radial kernel function [19]:

$$K(x, x_i) = \exp\left(-\gamma|x - x_i|^2\right) \quad (41)$$

The most widely used of the above kernel functions is the radial basis kernel function, which has a wide convergence domain and is suitable for various situations such as low dimensional, high dimensional, small sample and large sample. The best performing radial basis kernel function is also selected for audio classification. The value of γ is 8.

2.3.2 SVM multi-classification method

In recent years, SVM multi-class classification algorithms proposed by researchers at home and abroad can be roughly divided into two categories: One is to expand the basic two types of SVM into multi-class classification SVM, and this kind of method solves the optimization problem. There are a lot of variables used in the process, so it is not practical because the computational complexity is too high. The other is to gradually transform the multi-class classification problem into two types of classification problems, that is, to form a multi-class classifier with multiple two-class classification SVM. At present, such methods are widely used, and there are two commonly used classification strategies: One Against One [20] strategy and One Against All [21] strategy.

- (1) One-on-one strategy. This strategy was proposed by Knerr et al. in 1990. The main idea is to construct a classification hyperplane for any two categories when classifying and to separate the N categories. To classify N categories using this strategy, a total of $N * (N-1)/2$ two-class SVM classifiers are required. Then, according to this two-category combination, the classifier training of each two-category classification problem is performed. In the process of identification, each test sample is separately input into the $N * (N-1)/2$ two-classifiers, and the classification result obtained by each classifier is voted to obtain the category with the most votes. The final classification result of the sample, this strategy is called the “voting method.”
- (2) One-to-many strategy. This method was proposed by Bottou et al. in 1994. The main idea is when classifying, for the multi-classification problem of N categories of training samples, first construct a classification hyperplane between the i th class and other $N-1$ classes. Thus, the algorithm constructs a

total of N two types of SVM classifiers. When the i th classifier is trained, the sample of the i th category is $+1$, and the sample points of the other classes are -1 to perform the training of the two-category classification problem. In the identification process, each sample to be identified will enter the trained N classifiers, respectively, and the output values obtained by each classifier are compared to obtain the classification result. The one-to-many strategy requires each classifier to output a probability value of a certain class belonging to the classifier discriminant category, and then all the output probability values are compared, and the class of the classifier with the highest probability is taken as the class of the sample. The output of the support vector machine is a specific classification, and there is no probability value output. Therefore, when applying the one-to-many strategy, we do not find the discriminant category of the SVM, but find the probability output of the SVM. Through this calculation process, each sample has a probability value in each classification, indicating the probability that the sample belongs to a certain classification. Finally, the classifier with the largest output probability value is selected, and the category represented by $+1$ is the final classification result of the sample to be identified. The one-to-many strategy is simple, effective and has a short training time. It is more suitable for large-scale data classification than one-to-one strategy.

2.4 Audio segmentation technology

The purpose of audio segmentation is to use a computer program to intelligently segment the audio stream into segments of different lengths and properties, freeing the time, labor and capital costs of manual segmentation. The so-called uniformity means that the characteristic parameters of the audio segment are the same or similar whether in the time domain or the frequency domain.

2.5 Audio segmentation algorithm based on BIC theory

Audio segmentation based on Bayesian information criterion (BIC) [22] is a widely used method. The BIC criterion generally detects whether a model conforms to the BIC criterion by the difference between the maximum likelihood value of the sample and the complexity of the model. The complexity of the model usually refers to the parameters of the model. In recent years, due to its superior performance, it has been introduced into audio segmentation and clustering problems. Suppose $X =$

$\{x_i : i = 1, 2, \dots, N\}$ is a piece of audio sequence to be tested, N is the signal length, $M = \{m_i : 1, 2, \dots, K\}$ is the candidate model parameter, $L(X, M)$ is the maximum likelihood function of sample data X in model M , and m is the number of parameters of model M , then the BIC criterion is defined as shown in Eq. (42):

$$\text{BIC}(M) = \ln L(X, M) - \frac{1}{2} * \lambda * m * \ln(N) \tag{42}$$

where λ is the penalty factor [23], usually taken as 1.

Suppose the signal X satisfies the multivariate Gaussian distribution, and it has a window length signal $Y = \{y_1, y_2, \dots, y_n\}$, where n is the window length. In order to detect whether there is a split point in Y , it is necessary to detect every point $i(0 < i < n)$ in Y . Suppose that Y is divided into two parts by point i : $Y_1 = \{y_1, y_2, \dots, y_i\}$ and $Y_2 = \{y_{i+1}, y_{i+2}, \dots, y_n\}$, and hypotheses H_0 and H_1 are made on Y , which means that there is no dividing point in Y and there is a dividing point in Y , and the mathematical description is as shown in formula (43):

$$\begin{aligned} H_0 : Y &\sim N(\mu, \Sigma) \\ H_1 : Y_1 &\sim N(\mu_1, \Sigma_1), Y_2 \sim N(\mu_2, \Sigma_2) \end{aligned} \tag{43}$$

The corresponding maximum likelihood ratio can be described as shown in Eq. (44):

$$R(i) = n * \ln |\Sigma| - n_1 * \ln |\Sigma_1| - n_2 * \ln |\Sigma_2| \tag{44}$$

where μ, μ_1, μ_2 are the average values of Y, Y_1, Y_2 , respectively, and $\Sigma, \Sigma_1, \Sigma_2$ are corresponding covariance matrices, and n, n_1, n_2 are corresponding signal lengths.

Compare the H_0 and H_1 models and define the difference between their BIC values as shown in Eq. (45):

$$\Delta\text{BIC} = \text{BIC}(H_1) - \text{BIC}(H_0) = R(i) - \lambda p \tag{45}$$

where $p = 1/2 \times (d + 1/2 \times (d + 1)) \ln(n)$, d is the dimensions of the sample space. If the weighted variance set of the candidate segmentation points of all sequences is greater than 0, it means that there is a segmentation point in Y , and the assumption H_1 is true. The condition description is as shown in Eq. (46):

$$\{\max \Delta\text{BIC}(i)\} > 0 \tag{46}$$

When formula (45) is satisfied, there is a split point in Y , and the moment at which the split point is located can be described as shown in Eq. (47):

$$\hat{i} = \arg \max \Delta\text{BIC}(i) \tag{47}$$

If the formula (46) is not satisfied, then it is assumed that H_0 is established; that is, there is no division point in Y , and a new window Y is formed by amplifying n to perform BIC detection again. For a single split point and multiple split points, Chen proposed their own solutions [24], which is better for short-term clips with more transitions. However, if the column to be tested is too long and the segmentation

point cannot be detected for a long time, it will undoubtedly increase the amount of calculation. In addition, this method is prone to cumulative errors. If the wrong segmentation point appears before, this error will continue and it will not be corrected later.

2.5.1 Improved BIC audio segmentation algorithm

Although there are various defects in BIC-based audio segmentation, its advantages cannot be ignored. It is only necessary to slightly modify various inadequacies to ensure the robustness of the algorithm. The following is a description of some of the higher recognition improvements that have been made by later researchers to address these shortcomings.

A large part of the error accumulation and calculation of the traditional BIC method is due to the increase in the window length, so later researchers proposed a more intuitive improvement method, which is based on the fixed window length sliding mode. For each detected BIC window, the initial window length is constant. If a split point is detected, slide a certain length to the next window. If the split point is not detected, the window length is also increased, but when the window length is increased to a certain extent, the split point is still not detected and then the window keeps the current window length and slides forward until the split point is found to restore the initial window length. Even if the split point is detected, the window length is not increased and it is directly swung backward.

3 Experiments

The experimental test audio data are manually classified into silent/noise, pure speech, mixed voice, music, environmental sound, etc., and is used as a mixture of training samples and test samples.

There are many audio formats, such as wav, mp3, midi. The channels are divided into mono-, dual- and multi-channel. The sampling rate is 44.1 kHz, 32 kHz, 16 kHz, 8 kHz, and the precision is 32 bit, 16 bit and 8 bit. The audio is normalized before the audio experiment, the sampling frequency is 44.1 kHz, the quantization precision is 16 bits, and the audio files are unified to wav, and the uni-channel data are taken for analysis. Audio is divided into clip sequence number is 3600, after manual classification, mute clip760, noise clip630, music clip570, pure voice clip530, voice clip560 with background sound, ambient sound clip550.

4 Discussion

4.1 Classification of mute and noise

Quiet and noise use a rule-based classification method. The experimental design is as follows: Mute and noise domain values are judged for all samples, the correct classification number is recorded, the number of misclassifications (the number of clips that are not muted but judged to be mute) is calculated, and the classification accuracy is calculated. The experimental results are as follows (Table 1).

For other categories, the audio size is obviously different, so the recognition accuracy is high. The misclassification is mainly due to the fact that in a clip, it contains mute and other audio categories, so the energy average may be relatively small. The method of reducing the energy threshold is solved. The accuracy of recognition for noise is 85.87%. The reason for the analysis is that the noise sources appearing in different audio categories are not the same, so the time–frequency characteristics of the noise are also different. The single threshold is used to judge the lack of universality. Therefore, the accuracy of noise judgment in the test is not high, and the false positive rate is high. The environmental sound with small change in energy spectrum is easily misjudged as noise.

4.2 Classification of each audio category

SVM-based classifiers are used to classify pure speech/background sounds and music/environmental sounds. Three trials are performed for each classification (Tables 2, 3).

Table 1 Noise/silent classification results

	Number of correct classification	Number of misclassifications	Classification accuracy (%)
Noise	541	127	85.87
Mute	709	23	93.28

Table 2 Voice-sorted results with pure voice/background sound

	Pure speech classification accuracy (%)	Speech classification accuracy with background sound (%)	Average classification accuracy (%)
Experiment 1	92.76	89.32	91.04
Experiment 2	91.89	89.03	90.46
Experiment 3	94.52	90.17	92.34
Average	93.05	89.50	91.28

It can be seen from the experiment that the classification accuracy of the support vector machine classifier is very high, the average classification accuracy of pure speech and background sound is 91.28%, and the average classification accuracy of music and environmental sound is also 90.77%. It can be seen from the experimental data that the proposed support vector machine classifier has better classification effect and accuracy for audio classification work (Table 4).

4.3 Traditional Δ BIC segmentation method and improved Δ BIC segmentation method

The experiment uses the traditional segmentation method and improved segmentation method, respectively. In order to compare the accuracy of the improved segmentation method with the traditional segmentation method, the rationality of the proposed traditional segmentation criterion and the effectiveness of the improved segmentation method are verified. The classification results were segmented using traditional segmentation methods and improved segmentation methods.

The traditional Δ BIC segmentation method is equivalent to the improved Δ BIC segmentation method, and the number of segmentation results detected is significantly more than the improved method. The reason for the analysis is that the traditional method only smoothes the classification result and then directly combines the same category of audio to obtain the segmentation result. The interaction between adjacent segments is not considered, and the overall optimization of the segmentation is neglected. It is equivalent to relaxing the constraints of the audio lens segmentation, which will improve the accuracy rate, will inevitably lead to an increase in misclassification, resulting in a larger number of detected audio shots. The improved method transforms the segmentation problem into an optimization problem solution. It is a dynamic method, which fully considers the interaction between segments and the overall optimization of segmentation, so the number of false positives is significantly reduced, and

Table 3 Music/environmental sound classification results

	Music classification accuracy (%)	Environmental sound classification accuracy (%)	Average classification accuracy (%)
Experiment 1	91.58	87.74	89.66
Experiment 2	92.49	91.35	91.92
Experiment 3	91.84	89.67	90.75
Average	91.97	89.58	90.77

Table 4 Segmentation test results

Segmentation method	Segmentation result	Correct number/accuracy (%)	Missing judgment	False judgment
Traditional segmentation method	165	127/82.5	27	38
improved segmentation method	148	135/87.6	9	13

the accuracy is also slightly improved. The results reflect that the actual efficiency of the optimized method is significantly higher than the traditional method.

5 Conclusions

Audio classification is the basis of audio information deep processing, the core technology of audio structure, and an important means to extract audio structure and content semantics. It actually divides the audio data into different categories according to the perceived characteristics or the content of the expression, which can play an important role in content-based video segmentation, voice retrieval and audio supervision. Audio classification is also one of the key technologies for audio information processing, audio information retrieval and data management. Although audio classification does not have a long history, researchers have conducted more detailed research in this area, which not only makes the knowledge in this field become a complete system, but also promotes the development of audio information processing technology to a certain extent. In essence, the classification of audio data can be said to be a process of pattern recognition. Its research focus usually includes two basic aspects of audio feature analysis and extraction and the design and implementation of classifiers. An SVM-based audio classification algorithm that classifies audio into six categories: mute, noise, music, background sound, pure speech and speech with background sound. On the basis of classification, a smoothing criterion is proposed, and the classification result is smoothed, and finally the audio stream is segmented by audio category. The experimental results show that the SVM-based classification algorithm has a good classification effect and high classification accuracy. The smoothing processing further improves the

classification accuracy, reduces the misclassification rate, and the segmentation result is more accurate.

Acknowledgements This work was supported by Chongqing Big Data Engineering Laboratory for Children, Chongqing Electronics Engineering Technology Research Center for Interactive Learning, the Science and Technology Research Project of Chongqing Municipal Education Commission of China (No. KJ1601401), the Science and Technology Research Project of Chongqing University of Education (No. KY201725C), Basic Research and Frontier Exploration of Chongqing Science and Technology Commission (CSTC2014jcyjA40019), Project of Science and Technology Research Program of Chongqing Education Commission of China (NO. KJZD-K201801601).

References

1. Vapnik V (1995) The nature of statistical learning theory. Springer, New York
2. Zhang T (2001) Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans Speech Audio Process* 9(4):440–457
3. Kumar M, Mao YH, Wang YH et al (2017) Fuzzy theoretic approach to signals and systems: static systems. *Inf Sci* 418:668–702
4. Zhang WP, Yang JZ, Fang YL et al (2017) Analytical fuzzy approach to biological data analysis. *Saudi J Biol Sci* 24(3):563–573
5. Duda RO, Hart PE, Stork DG (2001) Pattern classification, vol 2. Wiley, New York
6. Molla Md.KI, Hirose K (2004) On the effectiveness of MFCCs and their statistical distribution properties in speaker identification. In: *IEEE international conference on virtual environments, human-computer interfaces and measurement systems*, pp 136–141
7. Picone JW (1976) Signal modeling techniques in speech recognition. *Proc IEEE* 79(4):157–161
8. Zhou B, Hansen JH (2005) Efficient audio stream segmentation via the combined T2 statistic and Bayesian information criterion. *IEEE Trans Speech Audio Process* 13(4):467

9. Seheirer E, Slaney M (1997, April) Construction and evaluation of a robust multifeature music/speech discriminator. In: Proceedings of ICASSP 97
10. Vernstrom T, Gaensler BM, Brown S et al (2017) Low frequency radio constraints on the synchrotron cosmic web. *Mon Not R Astron Soc* 467(4):4914–4936
11. Reynolds DA, Rose RC (1995) Text-independent speaker identification using Gaussian mixture speaker models. In: IEEE Transaction on SAP, pp 72–83
12. Li SZ (2000) Content-Based classification and retrieval of audio using the nearest feature line method. *IEEE Trans Speech Audio Process* 8(5):619–625
13. Feiten B, Frank R, Ungvary T (1991) Organization of sounds with neural nets. In: Proceedings of the 1991 international computer music conference. International computer music association, San Francisco, pp 441–444
14. Liang B, Yaali H, Songyang L, Jianyun C, Lingda W (2004) Feature analysis and extraction for audio automatic classification. In: The International workshop on image, video, audio retrieval and mining, Canada
15. Lu L, Jiang H, Zhang HJ (2001) A robust audio classification and segmentation method. In: Proceedings of the 9th ACM international conference on multimedia, pp 203–211
16. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297
17. Shirvani A, Chegini H, Setayeshi S et al. (2009) Polynomial kernel function and its application in locally polynomial neuro-fuzzy models. In: International CSI computer conference. IEEE, pp 54–59
18. Vapnik VN (1998) Statistical learning theory. Wiley, New York
19. Kim H, Elter D, Sikora T (2005) Hybrid speaker-based segmentation system using model-level clustering. In: Proceedings of the IEEE international conference on acoustics speech, and signal processing, pp 745–748
20. Chen S, Gopalakrishnan PS (1998) Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In: Proceedings of the speech recognition workshop
21. L Lu, H-J Zhang (2002) Real-time unsupervised speaker change detection. In: 6th International conference on pattern recognition, pp 358–361
22. Cheng SS, Wang HM, Fu HC (2008) BIC-based audio segmentation by divide and conquer. In: Proceedings of ICASSP 2008. IEEE Press, Las Vegas, pp 4841–4844
23. Chen S, Gopalakrishnan R (1998) Speaker environment and channel change detection and clustering via the bayesian information criterion. In: Proceedings of DARPA broadcast news transcription and understanding workshop, Lansdowne, VA, USA, pp 127–132
24. Cettolo M, Vescovi M. (2003) Efficient audio segmentation algorithms based on the BIC. In: Proceedings of the international conference on acoustics, speech, and signal processing, Hong Kong, China, pp 537–540

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.