Audio Example Recognition and Retrieval Based on Geometric Incremental Learning Support Vector Machine System

LINYUAN FAN
School of Statistics, Capital University of Economics and Business, Beijing 100070, China
e-mail: fanlinyuan@cueb.edu.cn

This work was supported in part by the Science and Technology Granted by the Beijing Education Committee under Grant KM201810038002, in part by the Key Members of the Outstanding Young Teacher by the Capital University of Economics and Business 2017, and in part by the Top Talent of Youth Teacher by the Beijing Education Committee 2018.

ABSTRACT With the fast development of computer and information technology, multimedia data has become the most important form of information media. Auditory information plays an important role in information location, this comes from the fact that it can be difficult to find useful information. Thus audio classification becomes more important in audio analysis as it prepares for content-based audio retrieval. There is quite a bit of research on the topic of audio classification methods, audio feature analysis, and extraction based on audio classification. Many works of literature extract features of audio signals based on time or Fourier transform frequency domain. The emergence of the wavelet theory provides a time-frequency analysis tool for signal analysis. Wavelet transformation is a local transformation of the signal in time and frequency which can effectively extract information from the signal, and perform multi-scale refinement analysis on functions or signals through operations such as stretching and translation instead of the traditional Fourier transformation. In the time-frequency analysis of the signal, the wavelet analysis captures the local time and frequency characters of the signal which can improve the ability of signal analysis. It can also change certain locals of the signal without affecting other aspects of it. In this paper, the frequency domain features are combined with the wavelet domain features. At the same time that the MFCC features are extracted, the discrete wavelet transform is used to extract the features of the wavelet domain. Then the statistical features are extracted for each audio example, and the SVM model is used to realize the different forms of audio classification identification.

INDEX TERMS Content audio, wavelet transform, audio feature, audio processing.

LIST OF ABBREVIATIONS
DTW Dynamic time warping
VQ Vector quantization
ANN Artificial neural network
SVM Support vector machine

I. INTRODUCTION
People can experience a colorful world from perceiving sounds around them, including the process of continuous recognition. With the advent of the first computer and the rise and development of artificial intelligence, pattern recognition, and speech signal processing technology, computer recognition of audio signals has become one of the core issues in computer intelligence and thus has become a widespread research hotspot of concern [1].

Audio data is an integral part of modern computer and multimedia applications. To increase audio data requires scientific methods for efficient and automatic classification and identification of audio data for user retrieval and browsing. Audio information retrieval is an urgent need for digital information retrieval today. Audio processing is a very broad concept, including audio digital signal processing, psychoacoustics, linguistics, vocal music and speech signal processing technology, computer technology and multimedia database technology. Audio can contain a wealth of information related to the scene it was taken at, making audio-based scene recognition possible [2]–[4]. Audio scene recognition is a process of automatically determining the scene around
the device by extracting the characteristics of the scene audio signal, which can make various portable devices more intelligent. For example, the mobile phone can automatically adjust to various scene modes according to different surrounding scenes [5]–[7]; portable digital devices can provide users with information such as location [8]; various hearing aids can automatically adjust settings for people with hearing impairments as their surrounding scene changes [9]–[12].

With current signal processing techniques, it is difficult to find features that can better characterize individual audio scenes. Therefore, it is a very feasible method to find the optimal feature subset by using existing audio features and thus improve the process of recognition effect. Audio processing has a long history and has achieved many results, mainly in the voice field. In terms of speech recognition, IBM’s Via Voice [13] has matured, the University of Cambridge’s VMR system [14], [15], as well as Carnegie Mellon University’s Infor-media, are excellent audio processing systems [16] that have matured greatly throughout the years. Audio information classification technology is the fundamental point of audio information processing, which has a large range of applications in audio and video interactive processing systems as well as other multimedia applications. The audio pattern matching method is relatively complicated and has entered the practical stage and generally includes three-steps: model library training, pattern matching, and decision making. At present, the model training techniques applied in the field of audio recognition mainly include dynamic time warping (DTW) technology, vector quantization (VQ) technology, artificial neural network (ANN) technology, and support vector machine (SVM).

In simple circumstances, given two discrete sequences (not necessarily related to time), the dynamic time warping (DTW) can describe the similarity or distance between the two sequences, and can also adapt to the extension or compression of two sequences. In this process, the unknown time axis is evenly twisted to align its features with the model’s features [17], [18]. Vector quantization is an important data compression technique, the methods commonly used in speech recognition have time-regulated and memory vector quantization. Both are suitable for small vocabulary recognition with large differences, and the codebook design is based on the main algorithm [19]–[22]. ANN is a relatively new method of speech recognition which came out in the late 1980s [23]. The ANN simulates the basis of human neural activity: with adaptability, parallelism, robustness, fault tolerance and learning characteristics, and has a particularly strong classification and mapping ability, a very attractive trait in the field of speech recognition [24]. However, due to the shortcomings of training models and recognition of samples, it has not been widely used. The Support vector machine is proposed by Batsaikhon et al. [25], Antonanzas-Torres et al. [26], and Liu and Xu [27] and is a general machine learning method. The biggest advantage of this method is that it avoids local optimization, yet it can still maintain a good generalization ability in small sample cases.

One of the best methods for recognizing generalization ability is the Support Vector Machine (SVM) [28]. Although SVM has a solid theoretical foundation and good generalization ability, SVM essentially solves a convex quadratic optimization problem. It requires kernel matrix operations, and the time complexity and space complexity and relativity quotient. Unfortunately, when dealing with large-scale data, the learning efficiency is relatively low [29]. SVM (Support Vector Machine) refers to machine learning algorithm that analyzes data for classification and regression analysis that is used in text categorization, image classification, handwriting recognition and in the other fields of science. Therefore, Liu et al. [30] and Li et al. [31] analyzed the support vector of the sample set and proposed a simple support vector machine incremental learning algorithm. Traditional support vector machine algorithms face dimension disasters as the number of samples increases. For incremental learning SVM, the addition of new samples may have an impact on the original support vector machine and classification results, so how to find new SV integration is an important part of incremental learning SVM. Incremental learning is a method of continuously updating the new sample set. Its thinking fully considers the impact of the new sample set on the original sample set, and at the same time makes full use of the results of historical classification, making the learning process continuous. Incremental learning SVM algorithm can realize online training for new samples. SVM data mining with new learning has become a popular research direction. Incremental learning SVM algorithm has the following advantages: 1. Reduced memory space, because there is no need to save historical data, so the memory usage is greatly improved in the process of incremental learning. When the new sample is trained, the historical training result is used to calculate, so that the training time is significantly smaller. Incremental Learning SVM algorithm is suitable for mining large numbers of data sets, and it can get efficient training results when new training is generated over time.

This paper considers the mechanism of improving the incremental learning algorithm to process audio based on the research from the incremental learning support vector machine. Using the geometric information of the historical sample set, the sample forgetting factor is set, and the boundary vector of the historical sample set is extracted for initial training. The new sample set is entered in batches, and uses the relationship between the KKT condition and the sample, to select the sample of the KKT condition that violates the decision function in the new sample set, adds the incremental learning, and the decision function obtained is then continuously corrected until the end of the incremental learning process to obtain the final decision function. In this algorithm, the training samples are divided into training sub-libraries and analyzed in batches. In each batch, only the support vectors are retained, and the non-support vectors are removed. Compared with the normal and decremental support vector machines, the experimental results show that the incremental learning support vector machine retrieval error rate is 11.8%
under the premise of significantly reducing the training time, which on average can save 47.3% of the traditional time, and which is lower than the traditional support vector machine retrieval error rate of 14.9%.

A. AUDIO SIGNAL PRE-PROCESSING
The audio signals stored in the computer are different in format, sampling rate, the number of bits, or the original audio signal containing sharp noise, which can affect the processing effect. At the same time, the unit of audio processing is in a frame, so before the feature extraction, the original audio data needs to be pre-processed: into a unified format, pre-emphasis, segmentation, and windowing framing. The audio signal pre-processing process is shown in Fig. 1.

![FIGURE 1. The audio signal pre-processing process.](image)

1) UNIFIED FORMAT
There are many audio formats stored in the computer. For example, we separated audio formats into some forms like mp3, wav, midi and so on, whose sample rate is 44.1k Hz, 32 k Hz, 16 k Hz, 8 k Hz, with accuracy is 32bit, 16bit, 8bit. If we do not unify the format, it brings great inconvenience to our experiments. Therefore, the system normalizes the audio before the audio is stored in the library, and the reference literature finally unifies the audio in the library, into wav, mono, 8 k Hz, 16 bit.

2) PRE-EMPHASIS PROCESSING
Pre-emphasis processing reduces the effects of sharp noises and boosts high-frequency signals. We denote \( x(n) \) as the original audio signal, the processed signal is \( y(n) \), as shown in the following formula (1):

\[
y(n) = x(n + 1) - ax(n)
\]

The parameter a is usually 0.97 or 0.98.

3) SEGMENTATION AND WINDOWING
The Haar feature was first proposed by Paul Viola et al. and later introduced by Rainer Lienhart et al. to introduce a 45° tilt feature. Haar features are three categories: edge features, linear, central and diagonal features, combined into feature templates. It is preferred to extract spectral features from a small window and assume that the signal is static in this small window [32], a non-static signal, on the other hand, would be a signal that’s statistical properties change over time.

The spectrum feature extraction is first done by setting a window that is non-zero for some specific time period and zeroes for other periods [33]. Spectral features can only be extracted at non-zero windows.

The window is then described by setting three parameters: the width of the window (in microseconds), the gap between two consecutive windows, and the shape of the window. In this chapter, a speech feature extracted from a window is called a frame, and the number of microseconds in one frame is defined as the frame size, and the number of microseconds between two consecutive windows becomes a frame offset.

In this paper is the classification of the audio clip. There is no way to classify the entire original audio signal, so the pre-emphasized audio signal should thus be segmented. The original audio stream is sliced into audio segment sequences, and then each audio segment is windowed and framed to obtain the minimum unit of audio processing. That is, the window function is multiplied to form a frame, and the adjacent two frames generally have a certain overlapping portion, so that the extracted audio features ensure a certain integrity. A commonly used window in audio sub-processing is the hamming window, which is expressed as follows (where \( N \) is the window width):

\[
w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi i}{N}\right), \quad 0 \leq i \leq N
\]

In this paper, the original audio signal is divided into audio segment sequences with a length of 1 s. Each 32ms is an audio frame with 50% overlap for each audio frame (that is, 256 samples per frame, \( N = 256 \), with 128 points overlapping).

B. AUDIO SIGNAL FEATURE EXTRACTION
Audio theory states that each audio signal is composed of sound waves of different moments, different frequencies and different energy amplitudes. The reason why people can often feel audio signals is that the human ear can feel the result of different energy signals in different frequency bands at different times. Audio classification is primarily based on audio features. Therefore, audio feature selection and extraction has become a focus of audio classification. Audio features mainly
include time domain features, frequency domain features and time-frequency features. The means of audio feature extraction is mainly digital signal processing technology.

According to the theory of short-term processing technology, an audio frame is the smallest unit for processing audio, and the length of a frame in normal audio processing is generally 20 to 40 ms. If it is too short, the information with too fine granularity will be obtained, and the distinguishing characteristics of various types of audio will not be reflected. If too long, it is easy to cause the timing characteristics of the feature after the averaging of the audio features.

1) FREQUENCY SHORT TIME ENERGY IS DEFINED AS FOLLOWS:

\[ STE = \log \left( \int_{0}^{\omega_0} |F(\omega)|^2 d\omega \right) \]  \hspace{1cm} (3)

\( F(\omega) \) is defined as FFT transform coefficient of the frame, with \( \omega_0 \) equal to one half of the sampling frequency. Then frequency domain energy \( STE \) is the determination of the silence frame. Suppose the \( STE \) of a particular frame is no more than the threshold, it can be defined as a silence frame, otherwise it is a non-silence frame. Meanwhile, frequency domain energy is also an effective feature to distinguish between music and mute. In general, speech and ambient sounds contain more muzziness than music, and thus the frequency domain energy variation of speech is much larger than in music.

2) SUB-BAND ENERGY DISTRIBUTION

The domain of frequency is divided into 4 sub-bands, written as \([0, \frac{\omega_0}{8}], [\frac{\omega_0}{8}, \frac{\omega_0}{4}], [\frac{\omega_0}{4}, \frac{\omega_0}{2}], [\frac{\omega_0}{2}, \omega_0] \). The distribution of each sub-band energy is calculated as follows:

\[ D = \frac{1}{STE} \int_{L_j}^{H_j} |F(\omega)|^2 d\omega \]  \hspace{1cm} (4)

\( L_j \) are the upper boundary frequencies of sub-band \( j \), but \( H_j \) the lower ones. All types of audio have separated distributions of energy in each interval. Therefore we got evenly distributed in each sub-band interval of music domain energy; the distribution of environmental sounds has a great relationship with the specific content; while in speech, with the energy concentrated in the first sub-band, which is larger than 80%.

3) ZERO CROSSING RATE

Facing discrete time signals, algebraic symbols of all samples are supposed to be of zero crossings. The zero-crossing rate describes the speed, which is set to be the measure of signal frequency, which can be calculated as follows:

\[ ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} [sgn(x(m+1)) - sgn(x(m))] \]  \hspace{1cm} (5)

where \( x(m) \) is a discrete audio signal, \( m = 1, 2, \ldots, N \). Usually, the speech signal is composed of alternate syllables and unvoiced syllables, and music does not have such a structure. The speech production model shows that the voiced speech energy is concentrated below 3k Hz due to the glottal fluctuation causing the high frequency drop of the spectrum. However, for unvoiced speech, most of the energy appears at higher frequencies. Since high frequency means high zero-crossing rate, low frequency means low zero-crossing rate. Then, there is a strong correlation between the zero-crossing rate and the frequency distribution of energy. Reasonable induction, the voice signal is unvoiced, the zero-crossing rate is high; the voice signal is voiced, and the zero-crossing rate is low. Therefore, for a voice signal, its zero-crossing rate changes more than music.

4) MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Based on Fourier and Cepstrum analysis, Fourier transform is performed on \([K/L]\) sample points in short-term audio to obtain the energy of each short-term audio frame at each frequency. If the audio signal sample rate is 25k Hz, then the sampling theorem shows that the maximum frequency of the audio frame is 12.5k Hz. That is to say, short-term audio frames have energy in the frequency band from 0k Hz to 12.5k Hz, except that the energy at different frequencies is different at each moment. Using the sensing characteristics of the human ear, the frequency band from 0k Hz to 12.5k Hz is divided into several sub-bands. When the entire frequency band is divided into frequency sub-bands, both linear and nonlinear modes can be adopted. If the entire frequency band is to be linearly divided into several sub-bands, each sub-band width can be taken as:

\[ Mef(f) = 2595 \log_{10}(1 + \frac{f}{500}) \]  \hspace{1cm} (6)

The division of the width of each frequency sub-band in the nonlinear partition is much more complicated. Whether it is a linearly divided sub-band or a nonlinearly divided sub-band, if the entire frequency band is divided into \( n \) MFCC coefficients (also called Mel coefficients). If the extracted Mel coefficient is calculated, the corresponding Cepstrum coefficient is also the Mel-Cepstrum coefficient.

C. INCREMENTAL LEARNING SUPPORT VECTOR MACHINE

The key to classification using support vector machines is to find the support vector, because the support vector can be used to determine the decision function. Not all training samples may become support vectors, and only those boundary vectors close to the optimal partition hyperplane may become support vectors. The algorithm of this paper is based on the learning result of pre-extracting the boundary vector.

The boundary vector is not necessarily a support vector, but the support vector must be a boundary vector. Therefore, the training samples can be selected first, and the boundary vectors that may become support vectors are selected, and then the boundary vectors are trained to not only ensure the accuracy of the algorithm, but also improve the speed of the algorithm. For the case where the sample set is linear, the extraction of the boundary vector can be performed in the input space. If the sample set is non-linear, the input space is first mapped to the feature space, and then the boundary
vector is extracted in the feature space. The idea of extracting the boundary vector is based on the distribution information of the two types of samples in the historical sample set, and their respective sample class centers P and N are calculated. From P and N as training sets, the approximate optimal classification function is obtained:

\[ g(x) = \sum \alpha_i^* y_i K(x_i, x) + b^* \]  

(7)

Setting the threshold \( D > 0 \), giving two planes \( g(x) = D \) and \( g(x) = -D \) parallel to the approximate optimal hyperplane \( g(x) = 0 \), and a vector located between these two planes is extracted as a boundary vector. As shown in Fig. 2.

**FIGURE 2. Extracting the boundary vector.**

The implementation process of extracting the boundary vector is as follows:

1. Calculate the class center of two types of samples in the historical sample set. The known historical sample set is \( H = \{x_i^+, y_i^+\}_{i=1}^{N_+}, \{x_i^-, y_i^-\}_{j=1}^{N_-} \}, x_i^+, x_j^- \in \mathbb{R}^n, y_i^+ = 1, y_j^- = -1, \) and the class centres of its positive and negative classes are:

\[ P = \frac{1}{N_+} \sum_{i=1}^{N_+} \phi(x_i^+), \quad N = \frac{1}{N_-} \sum_{j=1}^{N_-} \phi(x_j^-) \]  

(8)

where \( \phi(x_i) \) cannot be directly calculated, but the distance between the sample vectors \( x_i \) and \( x_j \) in the feature space is:

\[ d(\phi(x_i), \phi(x_j)) = \sqrt{K(x_i, x_i) + K(x_j, x_j) - 2K(x_i, x_j)} \]  

(9)

Thus, \( P \) and \( N \) can each be approximated by sample points that have the smallest sum of distances from all sample vectors in their respective classes.

2. Define the forgetting factor of the sample. The forgetting factor of the sample is defined as the distance from the sample to \( g(x) = 0 \), that is:

\[ d(x_k) = y_k \sum \alpha_i^* y_i K(x_i, x_k) + b^* \]  

(10)

3. Extract the boundary vector. Extract a historical sample with a forgetting factor less than a given threshold as the boundary vector, that is:

\[ B = \{x_k | d(x_k) < D, k = 1, 2, \ldots, N_+ + N_- \} \]  

(11)

It can be seen from Fig. 2 that as the threshold \( D \) changes, the scale of the resulting boundary vector also changes, thus controlling the convergence speed of the algorithm.

The boundary vector extraction process of the incremental learning support vector machine is shown in Fig. 3.

**D. RELATIONSHIP BETWEEN KKT CONDITION AND SAMPLE DISTRIBUTION**

In the process of incremental learning of support vector machines, many scholars have proposed using KKT conditions to examine the impact of new samples on historical training data. The support vector machine can attribute the classification problem to the following convex semi-definite programming problem:

\[
\max \left\{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \right\}
\]

s.t. \( \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, l \) (12)

To solve this problem, get the classification function as \( y = \text{sgn}(f(x)) \), where \( f(x) = \sum \alpha_i y_i K(x_i, x) + b \), if and only if each sample point x in the training set satisfies the following KKT condition, namely:

\[
\begin{align*}
\alpha_i & = 0 \Rightarrow y_i f(x_i) \geq 1 \\
0 \leq \alpha_i & \leq C \Rightarrow y_i f(x_i) = 1 \\
\alpha_i & = C \Rightarrow y_i f(x_i) \leq 1,
\end{align*}
\]

(13)

where \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l) \) is the optimal solution for the above plan.

For the decision function \( f(x), (x_i, y_i) \) is a new sample, then this sample violates the KKT condition sample distribution can be divided into the following three types:
(1) \((x_i, y_i)\) is located in the classification interval, and sample on the same side of the boundary can be classified by the classifier, satisfying \(0 \leq y_if(x_i) < 1\);

(2) \((x_i, y_i)\) is located in the interval, and sample of this class is on the opposite side of the boundary, and cannot be classified by the original classifier, satisfying \(-1 \leq y_if(x_i) < 0\);

(3) \((x_i, y_i)\) is outside the classification interval, and is different from the classification boundary of this class of samples, and cannot be correctly classified by the original classifier, satisfying \(y_if(x_i) < -1\);

Therefore, the necessary and sufficient condition for the new sample \((x_i, y_i)\) to violate the KKT condition is \(y_if(x_i) < 1\).

E. PRINCIPAL COMPONENT ANALYSIS

The principal component analysis method is essentially a dimensionality reduction process, that is, mapping the original N-dimensional vector \(x\) into an M-dimensional vector \(y\) \((M < N)\), and storing the information contained by \(x\) as much as possible. It is generally assumed that the mapping is a linear transformation. The commonly used principal component analysis method is a K-L transform. Here, for the dimensionality reduction of the audio vector, the K-L transform is applied to perform the principal component analysis of the audio feature vector.

Given the \(N\)-dimensional vector sequence \(\{x_i\}, i = 1, 2, \ldots, T\), the principal component analysis based on the K-L transform is implemented as follows:

1. The translation coordinate system, the mean vector of the model as the origin of the new coordinate system, that is \(x_i^* = x_i - x_m\), where \(\frac{1}{T} \sum_{i=1}^{T} x_i^*\);  
2. Find the overall autocorrelation matrix \(R: R = \frac{1}{T} \sum_{i=1}^{T} x_i^* x_i^*\);  
3. Find the eigenvalue \(\lambda_1, \lambda_2, \ldots, \lambda_N\) of \(R\) and the corresponding eigenvector \(q_1, q_2, \ldots, q_N\);  
4. Sort the feature values from large to small, such as \(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq \ldots \geq \lambda_N\). Then take the eigenvectors corresponding to the first \(M\) large eigenvalues to form a transformation matrix: \(A = (q_1, q_2, \ldots, q_N), M < N\);  
5. Transform the N-dimensional original vector into a new vector: \(y_i = A^T x_i\).

In the formula, the first dimension component of \(y_i\) is called the first principal component of the original vector \(x_i\), which contains the most information in the original vector, the second dimension component of \(y_i\) becomes the second principal component, and so on.

The flow chart of principal component analysis based on K-L transform is shown in Fig. 4.

In the audio feature set building module, the training audio set is first applied to extract feature vectors, all audio feature vectors form a matrix \(M\), and then the PCM parameter vector \(A\) for dimensionality reduction is calculated by the above method. When there is a new audio input that needs to be classified, the feature extraction is first performed, and the n-dimensional feature vector \(x\) is generated. Then use the PCA’s dimensionality reduction formula \(y_i = A^T x_i\) to calculate the new eigenvector \(y_i\) after dimension reduction. Then substituting it into the LDA equation continues the discriminant analysis and dimensionality reduction operations.

II. RESULTS AND DISCUSSION

A. EXPERIMENTS

The tools and software used in this experiment are ffmpeg (intercept video), MPEG-7 experimentation Model (extract descriptor), MATLAB (extract descriptor), libsvm (open-source SVM library).

There is currently no standard public library for audio classification, so first we build the system audio library for this system. The system’s audio database consists of nearly 40,000 audio segments (one audio segment is 1 second in length). These audios are mainly from broadcast, television and internet, including live reports, article readings, radio programs, world-famous music, pop music, road car sounds, clock sounds, footsteps, water boat sounds, animal sounds, aircraft sounds, noisy markets audio and a variety of nature sounds, for a total length of more than 11 hours. The audio library covers the audio categories mentioned in other previous literatures, and expands the audio library audio types and sample sizes. Its purpose is to be able to train a more general-purpose audio classifier to better meet the needs of the actual application of the system. This article selects the “music”, “advertising”, and “news” audio examples from...
the library as training or test samples, in which each type of audio training sample is about 19,182, and the training sample is 6241. Each audio example has a length of 3 to 10 s, and the audio sample has a sampling rate of 16 kHz.

Before training, the video segmentation and frame extraction are performed on these video segments. The algorithm is used to extract the features of each shot and the extracted video frames in the video. Then, the normalization process is performed, and the obtained feature vector is used as the feature data of the segment video. The training data is then divided into three training data sets by different types. In constructing the DAG classification model, this paper follows the principle of selecting the most easily separated conditions to construct the current node, assigning numbers 1 to 3 to music, advertisements, and news respectively.

B. DISCUSSION

This article will identify three different types of audio examples of “music”, “advertising” and “news”. Since these three types of audio can occur in different situations, the accompanying noise is different in different occasions. Therefore, in order to make the recognition correct rate high, it is necessary to collect as many samples as possible for training, and the training sample library is large.

The test sample is a sample used to detect the performance of the classifier. The unit is one, and the training sample is 6241. That is to say, the sample that trains the classifier has 6241 audio segments of 1 second in length. The training correct rate is the percentage of the correctly classified samples obtained by training the classifier obtained after training. The same test correctness rate is the percentage of the correctly classified samples obtained by the trained classifier to distinguish the test samples. From the above table we can see the effectiveness of the support vector machine in audio classification.

1. The comparison of the recognition results of the audio examples in three different ways is shown as:

It can be seen from Fig. 5 that in the single-class audio, the accuracy of the incremental learning support vector machine in this paper is 96.4% for advertising, 98.5% for music, 97.8% for news, and the overall audio recognition accuracy rate is 97.6%. The recognition rate of the general support vector machine is 81.9%, and the recognition rate of the reduced learning support vector machine is 70.8%. It can be seen that the recognition rate of the incremental learning support vector machine is more accurate. In the training process, the reduced learning support vector machine repeatedly trains the same data, which leads to over-learning problems (as shown in Fig. 2, the correct rate of recognition of the advertising audio example is very low).

Among the three types of videos, the classification accuracy of advertisements is the lowest. This is because the source of the advertisement is rich and the expressions are diverse, making it often indistinguishable from other types of videos. For example, music, animations, and sports scenes are often interspersed in advertising videos, so that its video features are similar to other types of video features, causing misclassification. The more the video with different expressions, the more likely it is to cause such misunderstandings. This is also the problem faced by video auto-classification.

2. The machine audio example retrieval is performed by three different training methods, including incremental support vector machine, common support vector machine and decrement learning support vector machine. The performance comparison is shown in Fig. 6. It should be noted that the recognition error rate listed in Fig. 6 is obtained by the person who judges that the result of the retrieval is correct. For example, in one of the 10 search results returned, the person judges that there are two error return audio examples, and the search error rate is 20%. Fig. 6 shows a comparison of the average error rates during the retrieval process.

Among them, the incremental learning support vector machine error rate is 11.8%, the ordinary support vector machine error rate is 14.9%, and the decrement learning support vector error rate is 23.8%. It can be seen that the incremental learning support vector machine has a lower error rate and a higher recognition rate.
Three different recognition methods training time comparison shown in Fig. 7. The time required for the three different training methods to generate the support vector machine is shown in Fig. 7 (the support vector machine training is completed under Matlab, and the training time does not include the audio example feature extraction).

As can be seen from Fig. 7, the incremental support vector machine has a promising application for the identification of large sample training library audio examples, because it can effectively reduce the training time and reduce the training time by 47.3% compared with the ordinary support vector machine.

In the Internet era, it is necessary to process massive amounts of data, and a large number of audio files are added every day. The traditional batch learning method cannot adapt to this demand, and only through incremental learning can effectively solve this demand. This is the continuous increase of the audio sample library, which leads to the larger the sample size as the number of samples increases, the more samples are tested. In order to verify that the increase probability of the sample is higher, the advertisement, music, news and audio are subjected to 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000 sample accumulation test, and the experimental verification is shown in Fig. 8.

It can be seen from Fig. 8 that the larger the sample size, the smaller the probability of checking the mistakes will be, and the higher the probability verification is, the closer the probability is to 100%. It shows that the statistical characteristics of the sample are consistent with the overall characteristics. The closer the results obtained when statistical analysis is performed on more and more samples, the closer the probability of the first type of error to the second type of error. This proves once again that the incremental learning support vector machine is better for audio recognition and retrieval.

III. CONCLUSION

Audio data is an indispensable part of multimedia information, and research on content-based audio retrieval technology is booming and has achieved considerable results. Based on the work of predecessors, this paper attempts to solve the problems of audio feature analysis, classifier design and speech information retrieval. The main contents of this paper are:

(1) Audio feature extraction and analysis are the fundamental point of audio classification, and content-based audio retrieval belongs to pattern classification. Therefore, in audio feature analysis, the most critical and high-dimensional independent features of audio are extracted, which improves the separability of features.

(2) The audio hierarchical structure is clarified, and the definitions of different levels of audio structural units are cited. According to the theory of pattern recognition, the audio classification is essentially reduced to a pattern recognition process. The technical process of audio classification and segmentation is designed, and the key technologies involved are discussed.

(3) Support vector machine training algorithm based on incremental learning method, and it is used for the recognition and retrieval of audio examples. Compared with the traditional algorithm, the overall average accuracy is 97.6%, and the error rate is 11.8%, its training time is also significantly reduced compared to the other two. For those audio recognition that require a large sample training library to cover most situations, the algorithm trains the training sub-library in batches, retains the support vector in each training, and removes the non-support vector, which can effectively prevent the error caused by human annotation. The training time was significantly reduced by 47.3%.

The main functions of the audio classification system studied in this paper are feature extraction and experimental analysis of the algorithm which has practical significance for the recognition of future audio, and can be used more widely. However, as audio media is more and more integrated into people’s lives, the amount of audio data is getting larger and larger, the content is more and more complex, and audio data processing, organization, management and retrieval face enormous challenges. The main functions of
the audio classification system implemented in this paper are feature extraction and experimental analysis of the algorithm. The function is still not perfect, and there is a certain distance from the practical requirements. How to improve and reach a practical level is the key work that should be done in the future.

REFERENCES

[1] T. A. Buscheck, Y. Sun, M. Chen, Y. Hao, T. J. Woltery, W. L. Bourcier, B. Court, M. A. Celia, S. Julio Friedmann, and R. D. Aines, “Active CO₂ reservoir management for carbon storage: Analysis of operational strategies to relieve pressure buildup and improve injectivity,” Int. J. Greenhouse Gas Control, vol. 6, pp. 230–245, Jan. 2012.

[2] J. Guo, “An improved incremental training approach for large scaled dataset based on support vector machine,” in Proc. Int. Conf. Big Data Comput. Appl. Technol., Dec. 2017, pp. 149–157.

[3] M. Kumar, Y. Mao, Y. Wang, T. Qiu, Y. Chenggen, and W. Zhang, “Fuzzy-theoretic approach to signals and systems: Static systems,” Inf. Sci., vol. 418–419, pp. 668–702, Dec. 2017.

[4] W. Zhang, J. Yang, Y. Fang, H. Chen, Y. Mao, and M. Kumar, “Analytical fuzzy approach to biological data analysis,” Saudi J. Biol. Sci., vol. 24, no. 3, pp. 563–573, Mar. 2017.

[5] Z. Li, G. Zhou, Z. Zha, and W. Li, “A study on the power generation capacity of piezoelectric energy harvesters with different fixation modes and adjustment methods,” Energies, vol. 9, no. 2, pp. 98–105, 2016.

[6] L.-T. Xie, P. Yan, W.-B. Lu, M. Chen, and G.-H. Wang, “Effects of strain energy adjustment: A case study of rock failure modes during deep tunnel excavation with different methods,” KSCE J. Civil Eng., vol. 22, no. 10, pp. 4143–4154, Oct. 2018.

[7] W. J. Park et al., “System and method for automatic wireless connection between a portable terminal and a digital device,” U.S. Patent 9 722 979, 2017.

[8] C. J. Billings, K. L. Tremblay, P. E. Souza, and M. A. Binns, “Effects of hearing aid amplification and stimulus intensity on cortical auditory evoked potentials,” Audiol. Neurotol., vol. 12, no. 4, pp. 234–246, 2007.

[9] F. Barker, L. Atkins, and S. de Lasigian, “Applying the COM-B behaviour model and behaviour change wheel to develop an intervention to improve hearing-aid use in adult auditory rehabilitation,” Int. J. Audiol., vol. 55, no. sup3, pp. S90–S98, Jul. 2016.

[10] J. Löhler, F. Grabner, B. Wollenberg, P. Schlattmann, and R. Schöneweiter, “Sensitivity and specificity of the abbreviated profile of hearing aid benefit (APHAB),” Eur. Arch. Oto-Rhino-Laryngol., vol. 274, no. 10, pp. 3593–3598, Oct. 2017.

[11] W. Gong, L. Qi, and Y. Xu, “Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment,” Wireless Commun. Mobile Comput., vol. 2018, Dec. 2018, Art. no. 3075849.

[12] A. Álvarez, C. Mendes, M. Raffaelli, T. Luís, S. Paulo, N. Piccinini, H. Arzelus, J. Neto, C. Aliprandi, and A. del Pozo, “Automating live and hearing-aid use in adult auditory rehabilitation,” Audiol. Neurotol., vol. 55, no. 8, pp. 6765–6772, Aug. 2016.

[13] A. AhmedK, S. Aljahdali, and S. Naimatullah Hussain, “Comparative prediction performance with support vector machine and random forest classification techniques,” Int. J. Comput. Appl., vol. 69, no. 11, pp. 12–16, 2014.

[14] L. Liu, T.-J. Pan, and Z.-Y. Zhang, “Incremental support vector machine combined with ultraviolet-visible spectroscopy for rapid discriminant analysis of red wine,” J. Spectrosoc., vol. 2018, Apr. 2018, Art. no. 4230681.

[15] C. Li, P. Liu, C. Zou, F. Sun, J. M. Cioffi, and L. Yang, “Spectral-efficient cellular communications with coexistent one- and two-hop transmissions,” IEEE Trans. Veh. Technol., vol. 65, no. 8, pp. 6765–6772, Aug. 2016.

[16] B. Qi, C. Qin, H. Shang, and L. Xiong, “Simulation of LXP cable insulation on-line monitoring based on TM electromagnetic wave spectrum characteristics,” Acta Microscopica, vol. 28, no. 2, pp. 169–178, 2019.

Linyuan Fan was born in China, in 1984. He received the B.S. degree in mathematics from Nankai University, in 2006, and the Ph.D. degree in mathematics from Peking University, China, in 2012. From 2012 to 2015, he worked as an Assistant Professor at the School of Statistics, Capital University of Economics and Business, where he got an Associate Professor title, in 2016. He is the author of two books and 13 research articles. His research interests include differential geometry, and applications of geometry in data analysis and computer science.