Audio Fingerprint Retrieval Method Based on Feature Dimension Reduction and Feature Combination

Qiu-yu Zhang*, Fu-jiu Xu, and Jian Bai
School of computer and communication, Lanzhou University of Technology
Lanzhou 730050, China
[e-mail: {zhangqylz, xufujiu1, baijian0213}@163.com]
*Corresponding author: Qiu-yu Zhang

Received July 26, 2020; revised December 6, 2020; accepted February 11, 2021; published February 28, 2021

Abstract

In order to solve the problems of the existing audio fingerprint method when extracting audio fingerprints from long speech segments, such as too large fingerprint dimension, poor robustness, and low retrieval accuracy and efficiency, a robust audio fingerprint retrieval method based on feature dimension reduction and feature combination is proposed. Firstly, the Mel-frequency cepstral coefficient (MFCC) and linear prediction cepstrum coefficient (LPCC) of the original speech are extracted respectively, and the MFCC feature matrix and LPCC feature matrix are combined. Secondly, the feature dimension reduction method based on information entropy is used for column dimension reduction, and the feature matrix after dimension reduction is used for row dimension reduction based on energy feature dimension reduction method. Finally, the audio fingerprint is constructed by using the feature combination matrix after dimension reduction. When speech's user retrieval, the normalized Hamming distance algorithm is used for matching retrieval. Experiment results show that the proposed method has smaller audio fingerprint dimension and better robustness for long speech segments, and has higher retrieval efficiency while maintaining a higher recall rate and precision rate.

Keywords: Audio Fingerprint Retrieval, MFCC, LPCC, Feature Dimension Reduction, Feature Combination

This work is supported by the National Natural Science Foundation of China (No. 61862041, 61363078). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

http://doi.org/10.3837/tiis.2021.02.008 **ISSN: 1976-7277**

1. Introduction

With the explosive growth of the number of digital audio on the Internet, high-speed retrieval in audio big data has become an urgent problem to be solved. Audio fingerprint retrieval technology uses short audio fingerprint data instead of audio itself for retrieval, which can effectively improve the efficiency of audio retrieval, but the amount of fingerprint data corresponding to audio big data is also quite large, and traditional audio fingerprint retrieval methods have been difficult to meet the requirement of fast and accurate retrieval in the audio big data environment [1]. Therefore, the audio retrieval technology has been widely concerned by many researchers.

At present, the existing audio fingerprint methods mainly include the Philips fingerprint and the Shazam fingerprint. The audio retrieval mainly has three types of methods: keyword detection, key speaker detection and key audio detection, although these technologies are relatively mature, there are still many defects. With the increasing of data, the amount of corresponding fingerprint data also increasing, and the problem of dimensionality disaster also emerging, Which result in the exponential growth of calculation and data when searching for similar fingerprints in high dimensions due to the dimensionality disaster [2]. In order to solve the above problems, it is necessary to select a suitable dimensionality reduction method to ensure most of the original audio information can be retained while reducing the audio feature dimension as much as possible.

The selection of speech features will directly affect the performance of audio fingerprints. Currently, the main features are commonly used including the Mel-frequency cepstral coefficient (MFCC), linear predictive coefficient (LPC), linear prediction cepstrum coefficient (LPCC), formant feature, Spectral entropy feature, energy feature, etc. [3], various features have prominent advantages in retrieval efficiency, distinguishability, robustness and so on, respectively. However, the appropriate processing of features can achieve optimization of audio fingerprint performance, such as weighting processing, feature combination processing, etc., which can compensate for the drawback of the original features or amplify the advantages.

In recent years, scholars have proposed many methods in audio fingerprinting, feature extraction, and feature dimension reduction. In terms of audio fingerprint extraction, the traditional Philips fingerprint and Shazam fingerprint are usually selected for improvement. For examples, Chu et al. [4] proposed an energy band calculation method based on the peak, which improves the robustness of audio fingerprints and has good robustness to volume changes and amplitude changes. Plapous et al. [5] used the IIR filter instead of the Fourier transform to extract audio fingerprints, and conducted down-sampling on the extracted fingerprints, which effectively improved the retrieval speed, but with poor robustness to noise. Chen et al. [6] proposed an improved Philips fingerprint based on wavelet transform, which can realize the accurate retrieval of 1s speech segments and have high robustness to the retrieval of short speech segments. Sun et al. [7] optimized the Shazam fingerprint, proposed a method based on the dynamic region feature peak pair selection, which improved the accuracy of audio fingerprint retrieval. Kamesh et al. [8] selected peaks in the alternate time periods of the spectrogram to extract audio fingerprints, and sorted the amplitudes of the time periods in descending order to increase the frequency matching depth, effectively reducing the computational complexity and improving the retrieval efficiency and retrieval accuracy. Sun et al. [9] proposed an efficient audio fingerprint retrieval method based on subband spectral centroids, set seed segments to select subbands that need to extract features, extracts audio fingerprints based on subband spectral centroids, and set a hit count threshold

during the retrieval phase, improved the recall rate and precision rate, but the query index time is longer. Terchi et al. [10] proposed an audio fingerprint based on discrete wavelet transform, which extracts the audio fingerprint through multi-resolution decomposition of the discrete wavelet transform, which effectively improves the robustness of the audio fingerprint. Lin et al. [11] proposed a feature extraction method based on spectral baseband phase, the spectral phase reconstruction is used to reduce the impact of the noise, which has better robustness under low SNR and poorer performance under noise. Jiang et al. [12] proposed an audio fingerprint based on the optimal solution selection algorithm for lifting wavelet packets, adopting minimum entropy to decompose optimal wavelet packet and construct the audio fingerprint, which has good robustness. Chu et al. [13] proposed a speech endpoint detection method in a noisy environment, filter out noise through a speech enhancement algorithm based on the best improved logarithmic spectrum amplitude estimation, but the robustness of high frequency noise is poor.

Feature fusion and feature combination methods are widely used in speech recognition and classification because they can reflect more information of the original speech. Wang et al. [14] adopted combined features to improve the performance of audio fingerprint, and constructed audio fingerprint by combining chromatograms, sonic graphs, and energy spectral density. The audio fingerprint constructed by this combined feature performance is better than the audio fingerprint constructed by a single feature. Borjian et al. [15] proposed an adaptive feature extraction method based on type recognition, used fusion feature to train the recognizer, and proposed an audio retrieval algorithm based on Kullback-Leibler distance, which increased the type recognition rate. Dash et al. [16] proposed feature fusion based on biological features and speech features, using wavelet transform and artificial neural network for feature extraction and training to improve speech retrieval accuracy. Archana et al. [17] combined MFCC and PLP as the feature of audio recognition, this feature has a low false accept rate (FAR) and false reject rate (FRR), but has better robustness in noisy environments.

In terms of dimensionality reduction, Vavrek et al. [18] proposed a dynamic time warping (DTW) algorithm based on weighted sequence to reduce the dimensionality of audio feature matrix, which improves the performance of precise retrieval and fuzzy retrieval. Cha *et al*. [19] proposed the multiple sub-fingerprint matching principle, offset matching principle and termination strategy, which effectively solved the problem of audio fingerprint dimensionality disaster, and achieved fast and efficient retrieval through high-dimensional audio fingerprint.

In order to further improve the efficiency of audio fingerprint retrieval, some scholars optimize the audio fingerprint retrieval method. Fernando et al. [20] proposed a speech retrieval method based on entropy feature, which uses an approximate Balltree proximity retrieval scheme to reduce retrieval time at the cost of retrieval accuracy. Wang et al. [21] proposed an entropy-based audio fingerprint retrieval technology, which uses the maximum common string, edit distance, and DTW algorithm to achieve retrieval in the retrieval phase. Yao et al. [22] significantly increased the retrieval speed of audio fingerprints through sampling and counting methods during the retrieval phase, using the inverted index reduced the storage space of audio fingerprints, which has higher recall rate and precision rate. Sun et al. [23] optimized the hash table of audio fingerprints through the Fibonacci sequence and adjusted the length of the hash table through the right shift operation, which effectively reduced memory consumption and improved system retrieval efficiency. Yao et al. [24] proposed an audio fingerprint retrieval method based on the turning point alignment fingerprint matching algorithm, this scheme can effectively enhance the anti-time scaling

ability and improve the retrieval accuracy. Zhang et al. [25] divided the audio fingerprint into multiple sub-fingerprints by distinguishing between silent and voiced segments, and improved it in the audio fingerprint extraction stage and matching stage, which is highly robust to noisy environments. Liang et al. [26] proposed an audio fingerprint algorithm based on double fingerprint recognition of short speech segments, and constructed two groups of audio fingerprints for retrieval and matching through Wallish transform and threshold determination of spectrum map, which improved robustness to a certain extent.

By analyzing the above research, most of the existing audio fingerprint methods are for short speech segments, and there are relatively few researches on the retrieval of long speech segments, the existing audio fingerprints have low efficiency in retrieving long speech segments and is less robust than short speech segments. To solve these problems, this paper adopts long speech segments of 20s as the object of study and a robust audio fingerprint retrieval method based on feature dimension reduction and feature combination is proposed. The proposed method uses feature dimension reduction method based on energy and information entropy to reduce the dimension of the feature matrix, and combines the MFCC and LPCC to extract the audio fingerprint of speech, which improves the robustness of the audio fingerprint. The main innovation work of this paper are:

1) A feature combination method based on MFCC and LPCC is proposed. This combined feature has a high recall rate and precision rate for various speech content preserving operations (CPO) of long speech segments.

2) The feature dimension reduction method based on information entropy is used to perform column reduction on the high-dimensional feature matrix, which can retain most of the original audio information; the feature dimension reduction method based on energy is used to perform row reduction on the high-dimensional feature matrix, which can achieve efficient feature extraction and fingerprint construction while ensuring the robustness of long speech features.

3) In the audio fingerprint retrieval stage, the normalized Hamming distance is used to retrieve the audio fingerprint, which effectively improves the retrieval efficiency.

The remaining part of this paper is organized as follows. Section 2 describes the relevant theories in detail, including speech feature extraction algorithms, feature dimension reduction methods and feature combination methods. The detailed audio fingerprint retrieval method is described in Section 3. Section 4 gives the experimental results and the performance analysis compared with other related methods. Finally, we conclude our paper in Section 5.

2. Related Theories

2.1 MFCC Feature

MFCC [27] is based on the characteristics of the human auditory system (HAS) in which the speech is analyzed through the conclusion of the human hearing experiment. Since human subjective perception of the spectrum is not linear, transforming the frequency spectrum of the speech signal into the perceived frequency domain can obtain more comprehensive and representative audio features. MFCC reflects the features of the short-term amplitude spectrum of speech from the perspective of human ear perception, and the meaning of the expression is more accurate, and the anti-noise performance is also better.

At present, the dynamic differential MFCC is mostly used for MFCC, because the dynamic differential MFCC can reflect more information of the original signal, but the long

speech segments can extract more features than the short speech segments, and the dimension of feature matrix extracted from long speech is large, the feature dimension reduction method is needed to reduce the dimension when using the feature matrix. Therefore, this paper doesn't use dynamic differential MFCC.

2.2 LPCC Feature

LPCC [28] is based on the characteristics of the human vocalization mechanism. When calculating LPCC, LPC must be calculated first. LPC passes through a linear combination of the sampled values at a certain moment before a certain sampling time to estimate and forecast. In this paper, the traditional full-pole model is used to calculate the linear prediction coefficient (LPC), then the fast Fourier transform (FFT) is used for LPC, the logarithm operation is performed on the results, and finally the LPCC feature is obtained by inverse Fourier transform.

However, compared with MFCC, LPCC is less robust to noise, but LPCC can reflect the features of different speakers, and can make up for some of the features that cannot be reflected by MFCC.

2.3 Feature Dimension Reduction

Due to the high feature dimension extracted from long speech segments, if the feature matrix of long speech segments is directly used to classify or retrieve the speech, the efficiency is often very low and time-consuming. In order to achieve audio fingerprint construction and retrieval efficiently, it is necessary to reduce dimensionality of feature matrix to reduce the amount of feature data [17]. In this paper, the feature dimension reduction process adopted is as follows:

Step 1: Feature extraction. Extracting the features of the original speech to get the feature matrix $\boldsymbol{N}$ of $p \times q$.

Step 2: Matrix dimension reduction. Energy-based feature selection algorithm and information entropy-based feature selection algorithm are used to dimensionality reduction algorithm, the feature matrix is mapped or processed by feature selection.

Step 3: Reconstruction feature matrix. The feature matrix after dimensionality reduction is reconstructed to obtain the feature matrix $\boldsymbol{N'}$ of $p' \times q'$, where $p'<p$ and $q'<q$.

The dimensionality reduction method is mainly divided into two methods: secondary feature extraction and feature selection. Both methods can reduce the dimension of the feature matrix. At present, secondary feature extraction through mapping is widely used, and high-dimensional data is converted into low-dimensional data through mapping, but the mapping will have a certain impact on the original feature matrix of the signal, which is not very obvious in the processing of short sound segments, but will have a more serious effect on the robustness when processing long speech segments' feature matrices. Therefore, in order to improve the robustness of audio fingerprints, this paper selects a feature selection method that has little effect on the robustness of long speech features for dimensionality reduction. By setting a certain standard, the feature matrix is selected according to this standard, and the features that don't meet the conditions are discarded, so that the dimension can be effectively reduced, and the features that meet the conditions form the new feature matrix, the features after dimensionality reduction can still reflect the information of the original audio well while reducing the amount of data. The influence of feature selection on the original feature matrix mainly depends on the standard adopted to filter the features, an excellent standard can effectively reduce the matrix dimension while retaining most of the original speech features.

2.4 Feature Combination

The feature combination method [14] is widely used in speech emotion recognition and speech classification because it can combine the advantages of different features to more fully reflect the characteristics of the original speech. The feature fusion method is as follows:

Step 1: Extract the x kinds of features of the original speech to obtain x kinds of different feature matrixes $\boldsymbol{A}$ and $\boldsymbol{B}$ etc.

Step 2: Reduce the dimension of different feature matrices to the same dimension through the dimensionality reduction method of Section 2.3 to obtain the x kinds of feature matrices $\boldsymbol{A'}$, $\boldsymbol{B'}$, etc. of $p'\times q'$ after dimension reduction. The dimensionality-reduced feature matrix is spliced left and right to obtain the combined feature matrix $\boldsymbol{Z}$=($\boldsymbol{A}$, $\boldsymbol{B}$...) of $p'\times xq'$.

Step 3: The combined feature matrix $\boldsymbol{Z}$ is processed through machine learning methods such as GMM, KNN, CNN according to the recognition or classification requirements, and the required feature matrix $\boldsymbol{Z'}$ is output.

At present, feature fusion methods are widely used in image big data processing recognition and classification in intelligent transportation management [29], audio content authentication, recognition, retrieval and other fields, but they are most suitable for speech emotion recognition and classification, and need to use machine learning methods for processing. This paper needs to retrieve long speech segments, and even if the long speech feature matrix has been dimension-reduced, the amount of data is still large, so that training through machine learning is time-consuming, the features after dimensionality reduction cannot reflect all the information of the original audio, it will have a greater impact on the results of machine learning training. Therefore, MFCC and LPCC are selected for feature combination in this paper, MFCC reflects the non-linear characteristics of human ear hearing, LPCC reflects the difference of human vocal tracts, reducing the dimensionality of the combined matrix of MFCC and LPCC to combine the characteristics of the two features and acquire reflect more information of the original speech quickly and efficiently.

3. The Proposed Method

This paper combines the existing audio fingerprint retrieval methods, takes long speech segments as the object of study, uses feature selection to reduce the dimension of the high-dimensional feature matrix, and uses the method of feature combination to combine the characteristics of different features, and proposes a robust audio fingerprint retrieval method based on feature dimension reduction and feature combination.

3.1 The Model of Audio Fingerprint Retrieval System

Fig. 1 shows the model of the audio fingerprint retrieval system based on feature dimension reduction and feature combination. The model is mainly composed of three parts: audio fingerprint construction, generation of audio fingerprint library and the retrieval of query speech.

The processing procedure of the audio fingerprint retrieval system is as follows:

Step 1: Audio fingerprint construction. Extracting the original speech feature matrix and construct the audio fingerprint.

Step 2: Generation of audio fingerprint library. The audio fingerprint is used to establish the audio fingerprint index, and the audio fingerprint library is generated after establishing a one-to-one mapping relationship with the corresponding original speech.

Step 3: Retrieval of query speech. When the query user submits the query speech, the same audio fingerprint construction method of **Step 1** is used to extract the audio fingerprint of the speech to be queried. Then match the extracted audio fingerprint with the audio fingerprint sequence in the index table of the audio fingerprint library by calculating the normalized Hamming distance, and return the search result to the query user.

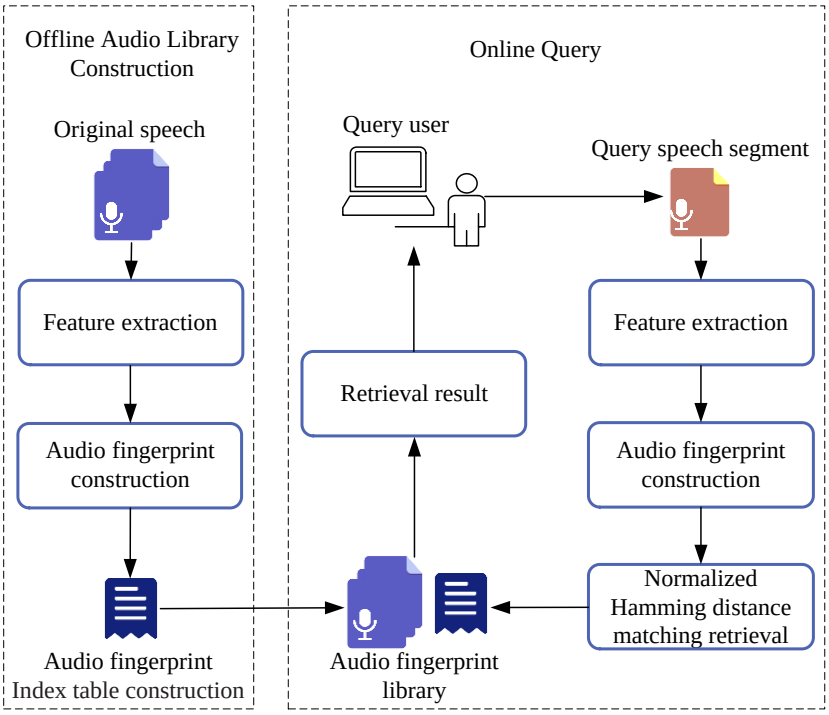

Fig. 1. The model of audio fingerprint retrieval system

3.2 Feature Extraction and Audio Fingerprint Construction

Fig. 2 shows the flow chart of audio fingerprint construction process. The process consists of four parts: feature extraction, feature combination, feature dimension reduction and audio fingerprint construction.

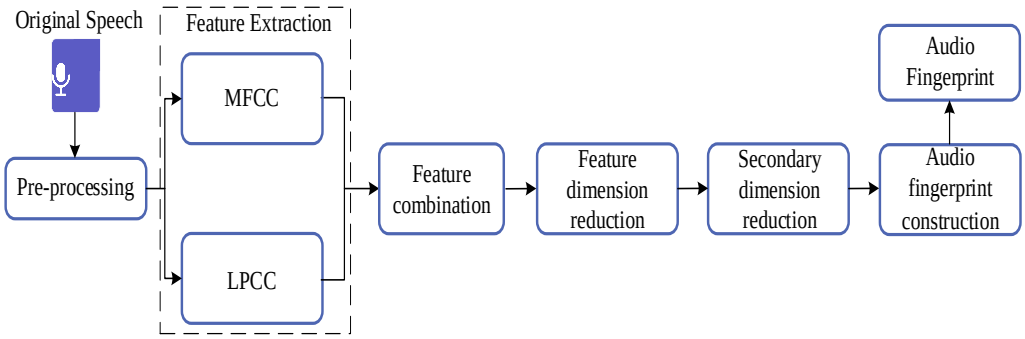

Fig. 2. The flow chart of audio fingerprint construction process

The processing procedure of the audio fingerprint construction is as follows:

Step 1: Pre-processing and feature extraction. The original speech is preprocessed by windowing and framing, and then MFCC and LPCC are extracted as the features of the audio fingerprint.

Step 2: Feature combination. MFCC feature matrix and LPCC feature matrix are combined to obtain a combined feature.

Step 3: Feature dimension reduction. In order to reduce the data of the extracted features while ensuring less loss of information, a feature selection method based on information entropy is used to achieve column dimension reduction of the combined feature matrix.

Step 4: Secondary feature dimension reduction. The combined features are subjected to energy-based feature selection, and energy is used as a selection standard to select audio frames to achieve row dimension reduction of the feature matrix, thereby further reducing the feature dimension.

Step 5: Audio fingerprint construction. The dimension-reduced feature matrix is used to construct the audio fingerprint according to the audio fingerprint construction method and output the audio fingerprint.

3.2.1 Feature Combination

Aiming at the problem that the fingerprint robustness of long speech is poor compared with that of short speech, this paper uses feature combination method, combining with the characteristics of MFCC and LPCC to improve the recall rate and precision rate of audio fingerprint in various audio processing. **Fig. 3** shows the flow chart of feature combination process. The process consists of two parts: feature combination, feature dimension reduction.

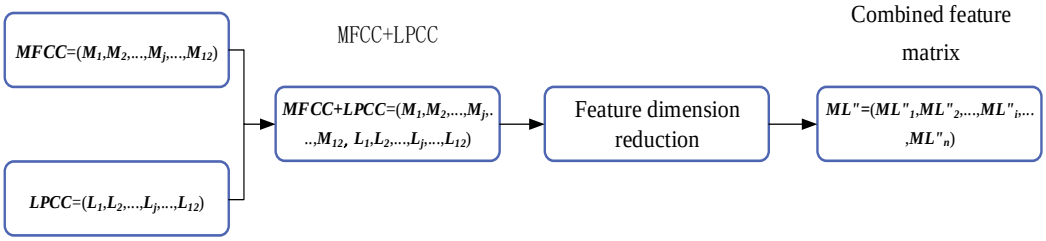

Fig. 3. The flow chart of audio feature combination process

Step 1: Feature extraction. In the feature extraction stage, the frame length and frame shift are set to 32ms and 10ms respectively, the Hamming window function is used, and the MFCC and LPCC features of the speech are respectively extracted according to the feature extraction methods of Section 2.1 and Section 2.2. Among them, 24 Mel filters are set during MFCC feature extraction to obtain a 12-dimensional MFCC feature matrix $\boldsymbol{MFCC}=(\boldsymbol{M_1}, \boldsymbol{M_2}, ..., \boldsymbol{M_j}, ..., \boldsymbol{M_{12}})$, where $\boldsymbol{M_j}=(mel_{(1)j}, mel_{(2)j}, ..., mel_{(i)j}, ..., mel_{(n)j})$ is the j-th feature vector of the MFCC feature matrix, and n is the number of frames. The coefficient of linear prediction is set to 12 during LPCC extractio to obtain a 12-dimensional LPCC feature matrix $\boldsymbol{LPCC}=(\boldsymbol{L_1}, \boldsymbol{L_2}, ..., \boldsymbol{L_j}, ..., \boldsymbol{L_{12}})$, where $\boldsymbol{L_j}=(lpcc_{(1)j}, lpcc_{(2)j}, ..., lpcc_{(i)j}, ..., lpcc_{(n)j})$ is the j-th feature vector of the LPCC feature matrix.

Step 2: Combination of feature matrixes. The extracted feature matrix is constructed according to the feature combination method in Section 2.4, the combined feature matrix is $\boldsymbol{ML} = (\boldsymbol{MFCC, LPCC}) = (\boldsymbol{M_1}, \boldsymbol{M_2}, ..., \boldsymbol{M_j}, ..., \boldsymbol{M_{12}}, \boldsymbol{L_1}, \boldsymbol{L_2}, ..., \boldsymbol{L_j}, ..., \boldsymbol{L_{12}})$.

3.2.2 Information Entropy-based Feature Dimension Reduction Method

At present, feature matrices extracted from long speech segments have larger dimensions. For example, the 12-dimensional MFCC feature extracted by 24 Mel filters has the best performance in the retrieval of short speech segments, but it will generate a huge amount of data for long speech segment retrieval, and the retrieval time will increase exponentially, so it's necessary to reduce the dimensionality of the audio feature matrix. Each column vector of the high-dimensional feature matrix contains different amounts of audio information, some column vectors contain most of the audio information, while some column vectors contain less information. For example, in the 12-dimensional MFCC matrix, the 5 column vectors with the highest amount of information contains more than 67% of the information. Due to the feature combination method used in this paper further increases the dimensionality of feature matrix, this paper proposes a feature dimension reduction method based on information entropy. The method selects features by calculating the information entropy of each column vector of the feature matrix, according to the information entropy of each column vector, several column vectors with more information are selected to form the new feature matrix. This method can maintain more original speech features and ensure the robustness of the fingerprint, while reducing the amount of data as much as possible and improving the retrieval efficiency.

The process of feature dimension reduction based on information entropy is as follows:

Step 1: Normalization of the matrix. Normalize the combined feature matrix $\boldsymbol{ML}$ constructed in Section 3.2.1, the normalized feature matrix is $\boldsymbol{ML'}=(\boldsymbol{ML'_1}, \boldsymbol{ML'_2}, ..., \boldsymbol{ML'_i}, ..., \boldsymbol{ML'_{12}}, \boldsymbol{ML'_{13}}, \boldsymbol{ML'_{14}}, ..., \boldsymbol{ML'_{2i}}, ..., \boldsymbol{ML'_{24}})$, where $\boldsymbol{ML'_j}=(ml'_{(1)j}, ml'_{(2)j}, ..., ml'_{(i)j}, ..., ml'_{(n)j})$ is the j-th dimensional normalized feature vector of the combined feature matrix.

Step 2: Calculation of information entropy. Calculate the information entropy of the normalized feature matrix, the information entropy is calculated as shown in (1) and (2).

$$m_{ij} = \frac{ml'_{ij}}{\sum_{i=1}^{n} ml'_{ij}} \tag{1}$$

$$e_i = -\sum_{i=1}^{n} m_{ij} * \ln m_{ij} \tag{2}$$

where e_i is the information entropy of each dimension of the MFCC feature matrix, and the information entropy matrix $\boldsymbol{E}=(e_1,e_2,...,e_{12})$.

Step 3: Feature selection. According to the information entropy matrix $\boldsymbol{E}$, the column vectors of feature matrix are sorted from large to small, and 10 of them with the largest information content form a new 10 dimensional combined feature matrix $\boldsymbol{ML''}=(\boldsymbol{ML''_1}, \boldsymbol{ML''_2}, ..., \boldsymbol{ML''_i}, ..., \boldsymbol{ML''_n})$, where $\boldsymbol{ML''_j}=(ml_{(1)j}, ml_{(2)j}, ..., ml_{(i)j}, ..., ml_{(n)j})$ is the j-th dimensional feature vector of the feature matrix after feature dimension reduction, and the feature matrix $\boldsymbol{ML''}$ is the feature matrix after feature dimension reduction method based on information entropy.

3.2.3 Energy-based Feature Dimension Reduction Method

Due to the large amount of long speech data and the number of frames in the feature extraction stage, after the feature dimension reduction of the column feature matrix based on

the information entropy, it is necessary to continue the row dimension reduction of the feature matrix. Therefore, we use energy as the feature dimension reduction parameter for row dimension reduction (That is secondary feature dimension reduction), to further reduce the amount of feature matrix data while ensuring the robustness of audio fingerprints.

The process of feature dimension reduction based on energy is as follows:

Step 1: Divide the long speech segment into n frames according to the framing method of **Step 1** in Section 3.2.1, and then divide the divided signal into five segments on average, and the number of frames in each segment is z.

Step 2: With in the range of $[f_1, f_{(z-30)}]$ of each segment, perform FFT on each frame signal to obtain frequency domain signal $X_i(k)$, and then performed the logarithm operation of each frame, the calculating formula of the logarithmic energy feature is shown in (3).

$$E(i) = \ln(1 + \frac{\sum_{k=0}^{l-1} X_i^2(k)}{c}), 1 \le i \le (z-30) \tag{3}$$

where k=0, 1, ..., l-1, l is the frame length of each frame, and c is a constant.

Step 3: By comparing the logarithmic energy of each frame to determine a frame $f_{\max}$ with the highest energy of each segment, taking the frame $f_{\max}$ as the starting point, 30 frames are taken backward, and $[f_{\max}, f_{(\max+30)}]$ is used as the frame for extracting features of each segment.

Step 4: Construct an n×1 empty matrix $\boldsymbol{T}$, set the position where the feature needs to be extracted to 1, and set the remaining positions to 0 to obtain the feature selection matrix $\boldsymbol{T'}$.

Step 5: Multiply the feature matrix $\boldsymbol{ML''}$ by the feature selection matrix $\boldsymbol{T'}$ to construct a new matrix, and the feature matrix $\boldsymbol{ML'''}$ is constructed with the dimensions of new matrix whose data is not 0. The number of rows of the feature matrix is reduced from n to 155, the matrix $\boldsymbol{ML'''}$ is new feature matrix after dimension reduction.

3.2.4 Construction of Audio Fingerprint Library

This paper improves the retrieval performance by improving the traditional Philips fingerprint retrieval algorithm during the audio fingerprint construction stage. The traditional Philips fingerprint retrieval algorithm uses the Euclidean distance for audio fingerprint retrieval through the sliding window. In this paper, the high-dimensional audio fingerprint matrix is reconstructed into a one-dimensional audio fingerprint matrix, and the one-dimensional audio fingerprint matrix is retrieved using the Hamming distance. The proposed method can effectively reduce the retrieval time of audio fingerprints.

The process of audio fingerprint library construction is as follows:

Step 1: The feature matrix $\boldsymbol{ML'''}$ is used to construct the audio fingerprint $\boldsymbol{h}$=(h_1, h_2, ..., h_{10}) by threshold judgment, the threshold judgment formula is shown in (4).

$$h_{(i)j} = \begin{cases} 1 & ML'''_{(i)j} \ge ML'''_{(i-1)j} \\ 0 & ML'''_{(i)j} < ML'''_{(i-1)j} \end{cases}, \begin{matrix} i = 1,2,...,155 \\ j = 1,2,...,10 \end{matrix} \tag{4}$$

Step 2: Each column of the 155×10 audio fingerprint $\boldsymbol{h}$=(h_1, h_2, ..., h_{10}) is transposed to reconstruct the 1550×1 audio fingerprint $\boldsymbol{h'}$=(h^T_1, h^T_2, ..., h^T_{10}).

Step 3: Generation of audio fingerprint library. According to **Steps** 1 and 2, the feature

matrix of all the speeches in the corpus is processed to obtain the audio fingerprint $\boldsymbol{h_x}$. The obtained audio fingerprint is used to build a linear index table, and the one-to-one mapping relationship between each audio fingerprint and the corresponding original speech is established to generate the audio fingerprint library.

3.3 Audio Fingerprint Retrieval

When querying the speech, the first 20s of the query speech $\boldsymbol{Q}$ is used to extract the audio fingerprint $\boldsymbol{h_Q}$ by the audio fingerprint method based on feature dimension reduction and feature combination, calculating the bit error rate (BER) in the linear index table by the normalized Hamming distance between the audio fingerprint and $\boldsymbol{h_x}$ in the audio fingerprint library. The normalized Hamming distance formula is shown in (5).

$$D(\boldsymbol{h}_x, \boldsymbol{h}_Q) = \frac{1}{m}\sum_{i=1}^{m}(\left|\boldsymbol{h}_x(\boldsymbol{i}) - \boldsymbol{h}_Q(\boldsymbol{i})\right|), \quad i = 1,2,...,m \tag{5}$$

where m is the length of the audio fingerprint.

When the user retrieves, setting the similarity threshold to T ($0.35<T<0.5$). If the normalized Hamming distance $D(\boldsymbol{h_x}, \boldsymbol{h_Q}) < T$, the retrieval is successful and the system will return to the query speech; otherwise the retrieval fails. The similarity threshold directly affects the robustness of audio fingerprint retrieval, in order to avoid missing detection and improve the robustness as much as possible, the similarity threshold is set to T=0.4.

4. Experimental Results and Performance Analysis

In the experiment, the speech library is constructed from the speech in the THCHS-30 speech library [30]. This paper uses a single-channel WAV format speech segment with the frequency of 16 kHz, and the sampling accuracy of 16bit, each speech is 20s, and the number of speech is 1000. In the audio fingerprint construction stage, We perform several kinds speech CPOs including resample, MP3 compression, narrowband Gaussian noise addition (30dB, 20dB, 10dB, 5dB, 0dB), use background noise and factory noise in NoiseX-92 to add noise in the original speech, a total of 10,000 kinds of speeches are obtained as a database. Experimental hardware environment: Intel(R) Core(TM) i5-7300HQ CPU, 2.50GHz, 8GB of memory. The software environment is: Windows 10, MATLAB R2017a.

4.1 Robustness and Retrieval Performance Analysis

The evaluation of the robustness of audio fingerprints is mainly through recall rate and precision rate. The calculation methods of the recall rate R and the precision rate P are shown in (6) and (7).

$$R = \frac{f_T}{f_T + f_L} \times 100\% \tag{6}$$

$$P = \frac{f_T}{f_T + f_F} \times 100\% \tag{7}$$

where f_T is the retrieved relevant speech, f_L is the relevant speech that is not retrieved, and f_F is the retrieved irrelevant speech.

In order to test the recall rate and the precision rate of the proposed method under different speech CPOs, the experiment used Gold Wave 6.38 and MATLAB R2017a to perform the 6 kinds of CPO shown in **Table 1** on 1,000 speeches.

The table lists the recall rate and the precision rate after 6 kinds of CPO: MP3 compression (128kbps, MP3), resample (16b→32b→16b, R.S), 30dB background noise addition (B.N), 30dB factory noise addition (F.N), and 30dB narrowband Gaussian noise addition (G.N).

In terms of robustness analysis, this experiment uses the feature dimension reduction algorithms of Section 3.2.2 and Section 3.2.3 to extract MFCC feature and LPCC feature of the original speech to construct audio fingerprint, this paper compares the audio fingerprint based on the combined features with the audio fingerprints of two features and the existing audio fingerprint methods of [6, 7, 22, 23, 25, 27], where [6, 7] is an improved method based on Shazam fingerprint, [22, 23, 25, 27] is an improved method based on Philips fingerprint.

Table 1. Comparison of the recall rate and precision rate with existing methods under different CPOs

Query index	Methods	Speech length (*s*)	MP3	R.S	B.N	F.N	G.N
Recall rate	Proposed	20	100%	100%	100%	100%	100%
	MFCC	20	100%	99.8%	99.8%	100%	100%
	LPCC	20	99.7%	99.9%	27%	91%	86%
	[6]	5	100%	-	-	-	100%
	[7]	3.4	100%	100%	99.8%	99.8%	99.6%
	[22]	6	99.71%	99.75%	-	-	99.74%
	[23]	3	99.87%	99.87%	-	-	99.45%
	[25]	5	97.1%	97.1%	-	-	-
	[27]	15	-	-	82.5%	-	86.5%
Precision rate	Proposed	20	100%	100%	100%	100%	100%
	MFCC	20	100%	100%	99.9%	100%	100%
	LPCC	20	97.2%	97.5%	45%	52.6%	49%
	[6]	5	100%	-	-	-	99.5%
	[7]	3.4	100%	100%	100%	100%	100%
	[22]	6	99.98%	100%	-	-	100%
	[23]	3	100%	100%	-	-	100%
	[25]	5	91.2%	91.2%	-	-	-
	[27]	15	-	-	94.2%	-	96.6%

As shown in **Table 1**, the combined feature is more robust than MFCC and LPCC that under the same feature dimension reduction process. The proposed method has good retrieval performance under different CPOs, compared with the audio fingerprint algorithm based on short speech segments in [6, 7, 22, 23, 25] under several CPOs, the proposed method can get similar or even better recall rate and precision rate, and compared with the audio fingerprint algorithm based on long speech segments that is robust to noise proposed in [27], recall rate and precision rate of the proposed method in terms of background noise and narrowband Gaussian noise are higher than [27].

In order to further test the retrieval performance of audio fingerprints, the proposed method uses F-score index to evaluate the audio fingerprints and obtain the comparison data shown in **Table 2**.

Table 2. Comparison of the F-score of with existing algorithm under different CPOs

Query index	Methods	Speech length (s)	MP3	R.S	B.N	F.N	G.N
F-score	Proposed	20	100%	100%	100%	100%	100%
	MFCC	20	100%	99.9%	99.9%	100%	100%
	LPCC	20	98.4%	98.7%	33.8%	66.7%	62.4%
	[6]	5	100%	-	-	-	99.7%
	[7]	3.4	100%	100%	99.9%	99.9%	99.8%
	[22]	6	99.8%	99.9%	-	-	99.9%
	[23]	3	99.9%	99.9%	-	-	99.7%
	[25]	5	94.1%	94.1%	-	-	-
	[27]	15	-	-	88%	-	91.3%

The calculation method of F-score index is shown in (8).

$$F = \frac{2PR}{P+R} \times 100\% \tag{8}$$

It can be seen from **Table 2** that after feature dimension reduction based on information entropy and energy, combined features have better retrieval performance than MFCC and LPCC that under the same feature dimension reduction processing. In addition, the proposed method can ensure high retrieval performance after MP3 compression, resample, background noise, factory noise and narrowband Gaussian noise, and the retrieval performance is higher than [6, 7, 22, 23, 25, 27].

Existing studies have shown that drawing P-R (Precision-Recall) curves can intuitively and comprehensively reflect the performance of the proposed audio fingerprint algorithms. **Fig. 4** shows the comparison results of the P-R curves of the proposed method with the other method in [6, 22, 25, 27], due to the retrieval performance in [6] is similar to [7, 23], this paper chooses [6, 22, 25, 27] to compare.

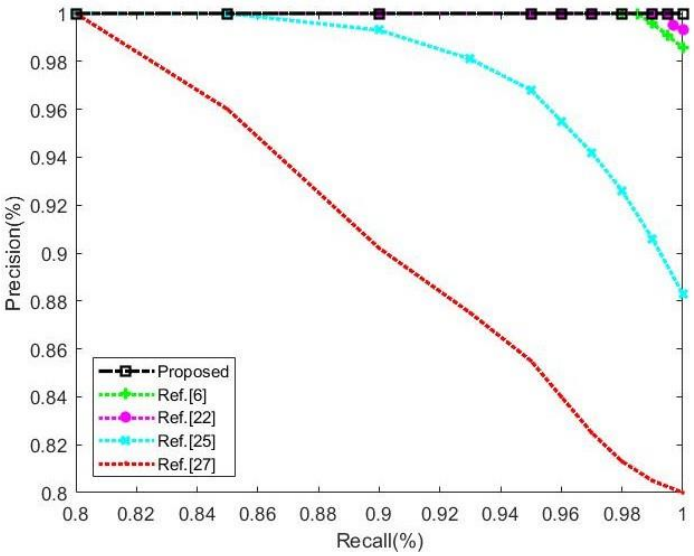

Fig. 4. Comparison of the robustness of audio fingerprint method under different method

As shown in **Fig. 4**, the area bounded by the P-R curve and the x-y coordinate axis of the proposed method is the larger than [6, 22, 25, 27]. This indicates that the retrieval

performance of the proposed method is better than [6, 22, 25, 27]. In addition, since recall rate and precision rate are mutually influenced, the proposed method has the greatest influence on recall when recall is 1.

4.2 Robustness Analysis of Low SNR Noise

This paper uses MATLAB R2017a to add 5 kinds of noise operations in 1,000 speeches, the noise operation includes adding 30dB narrowband Gaussian noise (30dB), 20dB narrowband Gaussian noise (20dB), 10dB narrowband Gaussian noise (10dB), 5dB narrowband Gaussian noise (5dB) and 0dB narrowband Gaussian noise (0dB) in the speech. In order to further test the robustness of the audio fingerprint extracted by the proposed method to noise, the noise robustness of the audio fingerprint is evaluated using *mAP* (mean Average Precision). *AP* refers to the average precision of different recall rates, and *mAP* is the average of each category of *AP* and the area under the P-R curve. The calculation method of *mAP* is shown in (9).

$$mAP = \frac{1}{K}\sum_{y=1}^{K}\sum_{r=1}^{f_T+f_L}\frac{P_y^r}{f_T+f_L}\times 100\% \tag{9}$$

where K is the number of speeches in the speech library, and f_T+f_L is the relevant speech that is queried.

Table 3 shows the compares the robustness of the proposed method with the other method in [6, 7, 22, 23, 25, 27] under different SNR.

Table 3. Comparison of robustness with existing methods under different SNR

Query index	Methods	30dB	20dB	10dB	5dB	0dB
Recall rate	Proposed	100%	100%	98.9%	95.4%	85%
	[6]	100%	100%	-	-	-
	[7]	99.6%	-	-	-	-
	[22]	-	100%	-	-	-
	[23]	-	100%	-	-	-
	[25]	-	96.17%	91.43%	86.03%	72.07%
	[27]	-	-	89.5%	88.45%	59.4%
mAP	Proposed	100%	99.85%	98.75%	93.57%	62.79%
	[6]	99.75%	96.5%	-	-	-
	[7]	99.6%	-	-	-	-
	[22]	-	99.69%	-	-	-
	[23]	-	99.9%	-	-	-
	[25]	-	73.29%	72.61%	69.73%	62.59%
	[27]	-	-	81.4%	76.1%	36.25%

As shown in **Table 3**, the combined features of the proposed method have good robustness to noise when the SNR is high. When the SNR is higher than 20dB, the query speech can be accurately retrieved. When the SNR is 20dB, the robustness is higher than [6, 22], which is similar to [23]. As the SNR is reduced from 20dB to 5dB, the recall rate and *mAP* of combined features are changing slower than [25, 27]. When the SNR is reduced from 5dB to 0dB, the *mAP* of combined features decreases significantly, but still higher than [25, 27].

Drawing P-R (Precision-Recall) curves can intuitively and comprehensively reflect the retrieval performance of audio fingerprint retrieval algorithms under noise processing. **Fig. 5**

shows the comparison results of the P-R curves of the proposed method under different SNR, which can more intuitively reflect the impact of noise interference on the robustness of the proposed method.

As shown in **Fig. 5**, the area bounded by the P-R curve and the x-y coordinate axis of the proposed method is the largest and has the best retrieval performance when the SNR is 20dB, with the decrease of SNR, the area gradually decreases. When the SNR drops from 20dB to 5dB, the decrease in area is smaller. When the SNR drops from 5dB to 0dB, the decrease in area is greater. This indicates that the retrieval performance of the proposed method is better at high SNR, while the robustness of the proposed method at low SNR has a great influence.

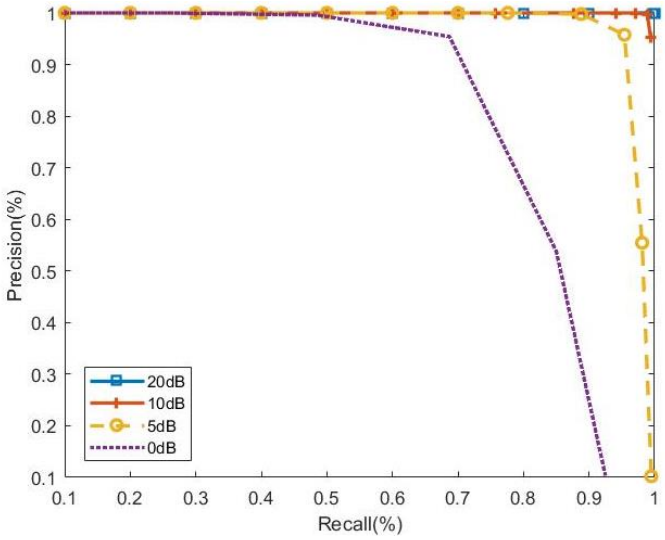

Fig. 5. Comparison of the robustness of the proposed method under different SNR

4.3 Search Efficiency Analysis

Retrieval efficiency is important index to evaluate the speech retrieval algorithm. In order to test the audio fingerprint retrieval efficiency of the proposed method, 10,000 speeches with a length of 20s were selected in the speech library for retrieval performance evaluation, and calculate the average retrieval time of the proposed method (including audio fingerprint construction time and retrieval matching time).

Table 4 shows the compares the retrieval efficiency of the proposed method and [6, 7, 22, 23, 25, 27].

Table 4. Comparison of retrieval efficiency with existing methods

Methods	Speech length (*s*)	Average running time (*s*)
[6]	5	0.720
[7]	3.4	0.430
[22]	6	0.712
[23]	3	0.970
[25]	5	0.8539
[27]	15	2.780
The proposed method	20	0.5328

As shown in **Table 4**, the retrieval efficiency of the proposed method is higher than [6, 22, 23, 25, 27], and slightly lower than [7]. However, the speech length of the audio fingerprint

constructed by the proposed method is 6 times that of Ref. [7, 23], 4 times that of [6, 25] and 3 times that of [22]. As MFCC, LPCC, logarithmic energy feature and information entropy are extracted respectively in the construction of audio fingerprint in this paper, the construction time of audio fingerprint is relatively high. However, the audio fingerprint retrieval is optimized, and the feature dimension reduction method is used to reduce the dimension of audio fingerprint, so the proposed method still ensures the robustness of audio fingerprint and realizes the fast retrieval of audio fingerprint. [22, 23] use the traditional Philips fingerprint method, use the Fibonacci sequence and sampling counting method to optimize the retrieval method, so the retrieval of short speech segments takes less time, while [27] takes a long time to extract features, resulting in low retrieval efficiency.

5. Conclusions

In order to solve the problems of low efficiency and poor robustness of the existing audio fingerprint method when using long speech segments for speech retrieval, this paper base on the advantages of feature combination in speech emotion recognition and the advantages of feature dimension reduction methods in processing high-dimensional data, an audio fingerprint retrieval method based on feature dimension reduction and feature combination is proposed. The proposed method constructs a combined feature matrix by combining MFCC and LPCC, and the combined feature matrix can reflect more information of the original audio. The feature dimension reduction method based on information entropy and energy are used to reduce the dimension of the feature matrix, which can effectively reduce the dimension of the feature matrix while retaining most of the features. The audio fingerprint is constructed from the combined features after feature dimension reduction, and the traditional Philips audio fingerprint algorithm is improved in the retrieval stage, and the audio fingerprint is retrieved through the normalized Hamming distance algorithm. Experimental results show that the proposed method can effectively combine the characteristics of MFCC feature and LPCC feature, and can effectively reduce the dimension of the feature matrix under the premise of ensuring robustness, and the constructed audio fingerprint has good robustness. The retrieval stage can achieve high recall rate and precision rate for long speech segments. Meanwhile, the retrieval accuracy and retrieval efficiency is effectively improved.

References

[1] N. M. Patil and M. U. Nemade, “Content-based audio classification and retrieval using segmentation, feature extraction and neural network approach,” *Advances in Intelligent Systems Computing*, vol. 924, no. 1, pp. 263-281, May 2019. Article (CrossRef Link)

[2] M. Mahdi, A. R. Ahmad, and R. Ismai, “Similarity search techniques in exploratory search: a review,” in *Proc. of TENCON 2018 - 2018 IEEE Region 10 Conference*, pp. 2193-2198, Oct. 28-31, 2018. Aritcle (CrossRef Link)

[3] G. Sharma, K. Umapathy, and S. Krishnan, “Trends in audio signal feature extraction methods,” *Applied Acoustics*, vol. 158, pp. 1-21, Jan. 2020. Article (CrossRef Link)

[4] R. Chu, B. Niu, S. Yao, and J. Liu, "Peak-based Philips fingerprint robust to pitch-shift for massive audio retrieval," in *Proc. of IEEE 5th International Conference on Multimedia Big Data (BigMM)*, pp. 314-320, Sep. 2019. Article (CrossRef Link)

[5] C. Plapous, S. A. Berrani, and B. Besset, “A low-complexity audio fingerprinting technique for embedded applications,” *Multimedia Tools Applications*, vol. 77, no. 5, pp. 5929-5948, Mar. 2018. Article (CrossRef Link)

[6] D. Chen, W. Zhang, Z. Zhang, W. Huang, and J. Ao, "Audio retrieval based on wavelet transform," in *Proc. of 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 531-534, May 2017. Article (CrossRef Link)

[7] X. Sun, W. Zhang, and D. Chen, "Movie retrieval based on Shazam algorithm," in *Proc. of 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pp. 1129-1133, Dec. 2018. Article (CrossRef Link)

[8] V. Kamesh, N. Pampana, M. Sinha, and S. Bandopadhaya, "Audio fingerprinting with higher matching depth at reduced computational complexity," in *Proc. of 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)*, pp. 1162-1166, July 2018. Article (CrossRef Link)

[9] J. Sun, J. Zhang, and Y. Yang, "Effective audio fingerprint retrieval based on the spectral sub-band centroid feature," *Journal of Tsinghua University (Science and Technology)*, vol. 57, no. 4, pp. 382-387, Apr. 2017. Article (CrossRef Link)

[10] Y. Terchi and S. Bouguezel, "Key-dependent audio fingerprinting technique based on a quantisation minimum-distance hash extractor in the DWT domain," *Electronics Letters*, vol. 54, no.11, pp. 720-722, May 2018. Article (CrossRef Link)

[11] X. Lin and X. Kang, "Exposing speech tampering via spectral phase analysis," *Digital Signal Processing*, vol. 60, pp. 63-74, Jan. 2017. Article (CrossRef Link)

[12] Y. Jiang, C. Wu, and K. Deng, "An audio fingerprinting extraction algorithm based on lifting wavelet packet and improved optimal-basis selection," *Multimedia Tools Applications*, vol. 78, no. 21, pp. 30011-30025, Nov. 2019. Article (CrossRef Link)

[13] X. Zhu, D. Huang, Y. Lu, and S. Fu, "Pilot speech endpoint detection in aircraft cockpit noisy environment," *Computer Engineering*, vol. 44, no. 1, pp. 317-321, Jan. 15, 2018. Article (CrossRef Link)

[14] M. Wang, K. Li, L. Luo, X. Song, Z. Zhou, and H. Qin, "An subarea localization algorithm based on combination features using representative audio fingerprint," in *Proc. of IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 374-380, Mar. 29-31, 2019. Article (CrossRef Link)

[15] N. Borjian, E. Kabir, S. Seyedin, and E. Masehian, "A query-by-example music retrieval system using feature and decision fusion," *Multimedia Tools*, vol. 77, no. 5, pp. 6165-6189, Mar. 2018. Article (CrossRef Link)

[16] D. Dash, P. Ferrari, S. Malik, and J. Wang, "Overt speech retrieval from neuromagnetic signals using wavelets and artificial neural networks," in *Proc. of IEEE Global Conference on Signal and Information Processing (GLOBALSIP 2018)*, pp. 489-493, Nov. 2018. Article (CrossRef Link)

[17] A. Dinesh and K. E. Bijoy, "Privacy preserving speech, face and fingerprint based biometrie authentication system using secure signal processing," in *Proc. of 2017 2nd International Conference on Communication Systems, Computing and IT Applications(CSCITA)*, pp. 164-168, Apr. 2017. Article (CrossRef Link)

[18] J. Vavrek, P. Viszlay, and M. Lojka, "Weighted fast sequential DTW for multilingual audio Query-by-Example retrieval," *Journal of Intellgent Information Systems*, vol. 51, no. 2, pp. 439-455, Oct. 2018. Article (CrossRef Link)

[19] G. H. Cha, "An efficient search algorithm for fingerprint databases," *Journal of Information Science and Engineering*, vol. 35, no. 2, pp. 471-484, Mar. 2019. Article (CrossRef Link)

[20] F. Luquesuarez, A. Camarenaibarrola, and E. Chavez., "Efficient speaker identification using spectral entropy," *Multimedia Tools Applications*, vol. 78, no. 12, pp. 16803-16815, June 30, 2019. Article (CrossRef Link)

[21] W. Wang, Z. Chen, X. Meng, and W. Li, "Research and implementation of identifying music through performances using entropy based audio-fingerprint," *Computer Science*, vol. 44, no. Z6, pp. 551-556, Dec. 2017. Article (CrossRef Link)

[22] S. Yao, B. Niu, and J. Liu, "Audio identification by sampling sub-fingerprints and counting matches," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 1984-1995, Sep. 2017. Article (CrossRef Link)

[23] N. Sun, W. P. Zhao, M. Chen, and C. Li, "An improved algorithm of Philips audio fingerprint retrieval," *Computer Engineering*, vol. 44, no. 1, pp. 280-284, Jan. 15, 2018. Article (CrossRef Link)
[24] S. Yao, B. Niu, and J. Liu, "Enhancing sampling and counting method for audio retrieval with time-stretch resistance," in *Proc. of 2018 IEEE 4th International Conference on Multimedia Big Data (BigMM)*, pp. 1-5, Sep. 2018. Article (CrossRef Link)
[25] X. Zhang, G. Zhan, W. Wang, P. Zhang, and Y. Yan, "Robust audio retrieval method based on anti-noise fingerprinting and segmental matching," *Electronics Letters*, vol. 56, no. 5, pp. 245-247, Mar. 2020. Article (CrossRef Link)
[26] T. Liang, X. Chen, C. Xu, and L. He, "Parallel double audio fingerprinting," in *Proc. of 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 344-348, Nov. 2018. Article (CrossRef Link)
[27] J. Lin, J. C. Yang, X. Y. Zhang, and X. C. Li, "Robust audio retrieval method based on fingerprint factors," *Journal of Data Acquisition and Processing*, vol. 31, no. 5, pp. 1020-1027, Sep. 2016. Article (CrossRef Link)
[28] G. Aggarwal and L. Singh, "Classification of intellectual disability using LPC, LPCC, and WLPCC parameterization techniques," *International Journal of Computers and Applications*, vol. 41, no. 6, pp. 470-479, Nov. 2019. Article (CrossRef Link)
[29] Y. Liu, C. Yang, and Q. Sun, "Thresholds based image extraction schemes in big data environment in intelligent traffic management," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-9, June 2020. Article (CrossRef Link)
[30] D. Wang and X. Zhang, "Thchs-30: A free Chinese speech corpus," *arXiv preprint arXiv*:1512.01882, 2015. Article (CrossRef Link)

Qiu-yu Zhang is a researcher and PhD supervisor. He graduated from Gansu University of Technology in 1986, and then worked at school of computer and communication in Lanzhou University of Technology. He is a CCF senior member, a member of IEEE and ACM. His research interests include network and information security, information hiding and steganalysis, image understanding and recognition, multimedia communication technology.

Fu-jiu Xu is a Master degree candidate. He received the BS degrees in communication engineering from University of Jinan, Shandong, China, in 2018. His research interests include audio signal processing and application, digital watermark, and multimedia authentication.

Jian Bai is a Master degree candidate. He received the BS degree in electronics and communication engineering from Shanxi University, Shanxi, China, in 2017. His research interests include audio signal processing and application, multimedia authentication and retrieval techniques.