An Encrypted Speech Retrieval Scheme Based on Long Short-Term Memory Neural Network and Deep Hashing

Qiu-yu Zhang[1*], Yu-zhou Li[1] and Ying-jie Hu[1]
[1] School of Computer and Communication, Lanzhou University of Technology
Lanzhou 730050, China
[e-mail: zhangqylz@163.com]
*Corresponding author: Qiu-yu Zhang

Received October 24, 2019; revised April 4, 2020; accepted May 7, 2020; published June 30, 2020

Abstract

Due to the explosive growth of multimedia speech data, how to protect the privacy of speech data and how to efficiently retrieve speech data have become a hot spot for researchers in recent years. In this paper, we proposed an encrypted speech retrieval scheme based on long short-term memory (LSTM) neural network and deep hashing. This scheme not only achieves efficient retrieval of massive speech in cloud environment, but also effectively avoids the risk of sensitive information leakage. Firstly, a novel speech encryption algorithm based on 4D quadratic autonomous hyperchaotic system is proposed to realize the privacy and security of speech data in the cloud. Secondly, the integrated LSTM network model and deep hashing algorithm are used to extract high-level features of speech data. It is used to solve the high dimensional and temporality problems of speech data, and increase the retrieval efficiency and retrieval accuracy of the proposed scheme. Finally, the normalized Hamming distance algorithm is used to achieve matching. Compared with the existing algorithms, the proposed scheme has good discrimination and robustness and it has high recall, precision and retrieval efficiency under various content preserving operations. Meanwhile, the proposed speech encryption algorithm has high key space and can effectively resist exhaustive attacks.

Keywords: Encrypted speech retrieval, long short-term memory neural network, deep hashing, speech feature extraction, 4D hyperchaotic system

1. Introduction

Due to the rapid development of Internet technology and the increasing popularity of communication tools, the collection of multimedia data has become easier. In multimedia data, the semantic function of speech makes speech essentially different from other sounds in nature. It often contains more important information content and is closely related to the privacy and security of individuals and society, such as instructions in the military field, speech evidence in litigation, conference recordings in telecommunications and finance. Therefore, in order to protect its privacy and security, special attention must be paid to the use and storage process. With the continuous development of cloud storage technology, this security measure is particularly important [1]. At present, the front-end encryption of speech is one of the methods to protect speech security in cloud storage environment. However, the encrypted speech often loses most of features, which makes it difficult to retrieve. Therefore, as the increasing scale of speech stored in the cloud, how to efficiently and safely implement encrypted speech retrieval is an urgent problem [2].

Therefore, the content-based encrypted speech retrieval methods was proposed to implement secure search [3-7]. The existing retrieval methods are basically based on speech perceptual hashing technology to extract the perceptual features of speech. Since these perceptual hashing-based encrypted speech retrieval methods utilize the already designed speech features, redesigning the speech feature requires a large number of prior knowledge and experiments, and the retrieval performance of algorithms depend largely on the extracted speech feature. Moreover, deep learning is one of the most important breakthroughs in the artificial intelligence has achieved great success in many fields, such as face retrieval [8,9], cross-modal retrieval [10,11,32], image retrieval [12-15], speech recognition [16,17,22-24], natural language processing [18,19], audio classification [20,21], emotion recognition [25-28] and sound detection [29-31]. As a machine learning method with multiple hidden layers, deep learning can acquire general abstract features by creating neural network models with multiple hidden layers and using a large number of training data. It can also solve the problem of high-dimensional data. Recurrent neural network (RNN) is different from convolutional neural network (CNN), which can process time sequence and model the changes in time sequence. LSTM neural network model as a variant of RNN model has achieved good results in speech recognition [22-24], emotional recognition [25-28] and speech detection [29-31]. In addition, in order to achieve privacy protection for speech in cloud, the speech encryption method is an indispensable technology in the encrypted speech retrieval system. Traditional encryption algorithms such as data encryption standard (DES), advanced encryption standard (AES) and Rivest-Shamir-Adleman (RSA) are no longer suitable for multimedia data encryption, and chaotic systems are widely used for multimedia data encryption with sensitivity, randomness and ergodicity to initial parameters. Especially, hyperchaotic system is more sensitive to parameters [36], which is very suitable for speech encryption.

According to the advantages of the above deep learning technology and LSTM model in different fields, we proposed an encrypted speech retrieval scheme based on LSTM and deep hashing in this paper. Our main innovative work is as follows:

1) An LSTM network model is designed to deep extract the speech feature of Log-Mel Spectrogram/MFCC, which solves the high dimensional and temporality of the speech data.

2) In deep hashing construction, the extracted deep feature is combined with hash function to generate binary deep hashing codes, which has good distinguishability and robustness.

3) A speech encryption algorithm is proposed using 4D quadratic autonomous hyperchaotic system, which can effectively resist exhaustive attacks and improve the security of speech.

4) Introducing the batch normalization algorithm can effectively improve the network fitting speed and reduce the training time. Combining the LSTM network model with deep hashing improves the retrieval accuracy and retrieval efficiency of speech.

The remaining section of this paper is described as follows. Section 2 discusses research related work. Section 3 presents the relevant theories of the research work in this paper. Section 4 details the encrypted speech retrieval scheme and its processing. In Section 5, the encrypted speech retrieval scheme was experimentally verified and compared with the existing methods. Finally, we conclude and look forward in Section 6.

2. Related works

The existing content-based encrypted speech retrieval algorithm [3-7] are all realized by constructing speech perceptual hashing by now. For example, Wang et al. [3] proposed an encrypted speech perceptual hashing retrieval algorithm based on zero-crossing rate and used Chua's chaotic system to encrypt speech. Wang et al. [4] proposed an encrypted speech retrieval scheme, which combines the speech perception hash algorithm based on the time-frequency domain trend transformation and the logistic XOR encryption. Zhao et al. [5] proposed an encrypted speech retrieval algorithm based on the multi-fractal features of speech signals and piecewise aggregation approximation to generate perceptual hashing sequences. He et al. [6] used syllable-level perceptual hashing has better distinguishing and robustness than time domain and frequency domain features. A retrieval method of syllable-level perceptual hashing based on posterior probability feature of syllable segment model is proposed. Zhang et al. [7] proposed an encrypted speech retrieval algorithm that can extract the perceptual hash sequence directly from the encrypted sample speech by using short-term cross-correlation and perceptual hashing. By analyzing the above, the existing content-based encrypted speech retrieval algorithms are based on existing hand-crafted features, and the extracted features are hashed to generate binary codes for retrieval.

The biggest difference between the deep learning methods and the traditional methods is that it can automatically learning features from dataset, rather than using hand-crafted features, which can reduce the data dimension and greatly improve the performance of the system. For example, Dong et al. [8] proposed a deep CNN network which integrates feature extraction and hash learning into a unified optimization framework for face-video retrieval. By using the proposed low rank discriminant binary hash, a better network initialization for hash function learning was realized. Tang et al. [9] proposed a new deep hashing model based on classification and quantization errors for scalable face image retrieval, which can simultaneously learn discriminant image representation and compact binary hash codes. Ma et al. [10] proposed a deep hashing method based on global and local semantics preservation for cross-modal retrieval, which implements large redundancy between similar hash codes and different hash codes from inter-modal views to learn discriminative hash codes. Deng et al. [11] proposed a triple-state-based deep hashing (TDH) network for cross-modal retrieval, which introduces graph regularization to preserve the original semantic similarity between hash codes in Hamming space. Cao et al. [12] proposed a novel deep hashing model, Deep Cauchy Hashing (DCH), which can generate compact binary hashing codes using cauchy quantization loss to achieve efficient Hamming spatial retrieval. Liu et al. [13] proposed a deep self-learning hash algorithm (DSTH), which can generate a set of pseudo-tags by

analyzing the data itself, and then use the discriminant deep model to learn the hash function of the new data. Tang et al. [14] proposed a new discriminant deep quantization hash (DDQH) method, which introduces the batch normalization quantization (BNQ) module to improve retrieval accuracy and simultaneously generate more discriminative hash codes. Cheng et al. [15] proposed an integrated deep hashing algorithm to extract the high-level features of the image, and used kNN, hyperchaos and DNA coding technology to improve the retrieval security.

In addition, the RNN is different from the CNN and other neural networks. It can process the time sequence and model the changes in the time sequence. The basic neural network only establishes weight connections between layers, and the biggest difference of RNN is that weight connections are also established between neurons in layers. For example, Song et al. [16] proposed a new recurrent neural network architecture Skip-RNN, which consists of acoustic model network and skip strategy network. It can dynamically skip less important speech frames to improve network processing speed. Fujimoto et al. [17] proposed a network-based factor modeling framework with various deep convolution recurrent neural networks for noise robust automatic speech recognition (ASR). Morchid et al. [18] proposed a new recurrent unit called "Parsimonious Memory Unit" for natural language processing. It is related based on the assumption that short-term and long-term dependencies, and each hidden neuron must be different in order to better handle terminological dependencies. Korvigo et al. [19] proposed a based on convolutional and stateful recurrent neural network for chemical named entity recognition (NER), which achieves near human level performance on test data sets. Xu et al. [20] proposed a convolutional recurrent neural network (CRNN), which has the non-linearity of a learnable gated linear unit (GLU) that can be applied to Log-Mel Spectrogram. Sang et al. [21] proposed a convolution recurrent neural network that can directly use time domain waveform as input to urban sound classification.

In practical application, in order to solve the long-term dependence of RNN, the improved RNN neural network model is called long short-term memory (LSTM) neural network model. For example, Pradeep et al. [22] proposed a speech recognition method using spectral flatness measurement (SFM) for linear prediction coefficients. Ghorbani et al. [23] proposed a RNN with connection time classification (CTC) cost function for speech recognition on multiple accent English data. Yu et al. [24] proposed a long short-term memory recurrent neural network acoustic model based on attention mechanism and multi-task learning framework, which significantly improved the model's ability to model far-field speech. Xie et al. [25] proposed attention-based dense LSTM for speech emotion recognition, which effectively solved the problem of information loss and degradation in high-level deep neural networks. Tao et al. [26] proposed an advanced LSTM (A-LSTM) for better time background modeling, and used A-LSTM for emotion recognition in weighted aggregate RNN. Ramet et al. [27] proposed a new LSTM-based attention model that can learn localized emotional prominence in a more robust way by considering the order of speech data. Etienne et al. [28] proposed a LSTM framework with data enhancement technology for speech emotion recognition, which can extract high-level features from the original spectrum. Jung et al. [29] proposed a method that combines convolutional bi-directional recurrent neural network (CBRNN) with transfer learning for polyphonic sound event detection, effectively solving the problem of vanishing gradient and over-fitting. Matsuyoshi et al. [30] proposed a weak marker learning method using bi-directional long short-term memory (BLSTM) with the connection time classification (CTC) to reduce the hand-marking cost of learning samples. Liu et al. [31] proposed a method to detect bowel sounds (BS) by LSTM with MFCC. The network is well trained with a large amount of data

and has excellent generalization ability in the same recording environment. Elizalde et al. [32] proposed a cross-pattern search framework based on Siamese neural network, which can retrieve recordings using text or audio queries.

The main function of RNN model is to handle and predict sequence data. In neural network or CNN, layers are fully or partially connected, but the nodes in layers are not connected. The nodes in the hidden layers are connected for RNN model. The input of the hidden layer includes output of the input layer and output of the hidden layer at the last moment. As an excellent variant of RNN model, LSTM network model inherits the characteristics of most RNN models and solves the vanishing gradient problem. Therefore, LSTM network model is very suitable for dealing with problems highly related to time sequence. Based on this characteristics, we proposed an encrypted speech retrieval algorithm based on LSTM and deep hashing in this paper.

3. Relevant theories analysis

3.1 Log-Mel spectrum and MFCC feature

According to the study of human auditory mechanism, it is found that human ear has different auditory sensitivity to sound waves with different frequencies. The human ear can only perceive audio signals in the frequency range of 20 Hz to 20 kHz. According to the auditory perceptual characteristics of the human ear, the Mel filter banks [33] is proposed, which is defined in Eq. (1):

$$f_{mel} = 2595 \times log(1 + f/700) \tag{1}$$

where f is the actual frequency value of speech signal.

Based on the theory of Mel filter banks, the MFCC features are extracted. The processing flow is as follows:

Step 1: Pre-processing, including pre-emphasis, framing, and windowing.

Step 2: Perform the fast Fourier transform (FFT) on the signal to obtain $X_n(k)$, where k=0, 1, 2, ..., N, N represents the number of points of the FFT.

Step 3: Filter with Mel filter $H_m(k)$, where m=0, 1, ..., M-1, M is the number of filters.

Step 4: Calculate the logarithmic energy of each filter bank output, as shown in Eq. (2).

$$s(i) = ln(\sum_{k=0}^{N-1} |X_n(k)|^2 H_m(k)), 0 \le i \le M \tag{2}$$

where $s(i)$ is the logarithmic energy of the i-th Mel filter, $X_n(k)$ is obtained by the FFT of **Step 2**, and $H_m(k)$ is $X_n(k)$ obtained by Mel filtering of **Step 3**.

Step 5: Perform discrete cosine transform (DCT) to obtain MFCC parameters, as shown in Eq. (3).

$$C_{MFCC}(l) = \sum_{m=0}^{M-1} s(i) cos(\frac{\pi l(n-0.5)}{M}), l = 1,2,\cdots,L \tag{3}$$

where L is the dimensions of the MFCC feature, M is the number of filters, and $C_{MFCC}(l)$ represents the MFCC feature of the 1-th dimension.

Compared to the extraction process of the MFCC, the Log-Mel spectrogram [34] only lacks the DCT transform of the **Step 5** that converts the logarithmic Mel spectrum to cepstrum. The extraction process of Log-Mel spectrogram and MFCC is shown in **Fig. 1**.

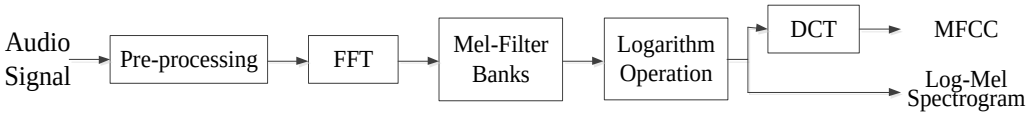


Fig. 1. Log-Mel spectrogram and MFCC extraction process

3.2 Long Short-Term Memory (LSTM) neural network

LSTM [35] is specially designed to solve the long-term dependence problem of common RNN model. It is suitable for processing and predicting important events with relatively long time intervals.

Compared with the basic RNN model, the biggest change of LSTM model is to replace the neural node with a neuron containing input gate, output gate, forget gate and memory unit (Cell). Among them, input gate, output gate and forget gate are all logic units. They do not send output to other neurons. Instead, they are responsible for setting weights at the edges of other parts of the neural network connected with the memory unit, which are used to modify the error function of selective memory feedback as the gradient decreases. The Cell replaces internal storage and maintains data in Cell, which is called Cell state. This cell state runs through the entire LSTM model architecture with only a small amount of linear interaction, which allows information to remain unchanged during transmission. **Fig. 2** shows the structure of LSTM neurons.

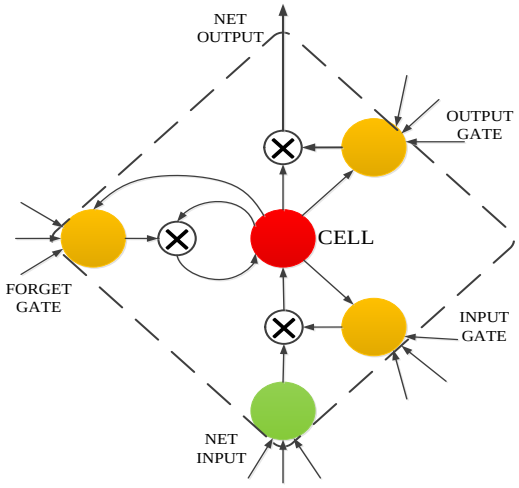


Fig. 2. LSTM neuron structure

As shown in **Fig. 2**, the Cell state of the output gate is preserved by a cyclic connection, and the Cell state information is controlled by multiplication. Since memory capacity is limited, early memories decays exponentially, and Cell state is designed to maintain long-term memory. The input gate, output gate and forget gate in **Fig. 2** have the same control unit structure, which is mainly composed of activation function and multiplication operation.

3.3 4D quadratic autonomous hyperchaotic system

The 4D quadratic autonomous hyperchaotic system [36] is derived from the classical Lorenz system and generates hyperchaotic systems by extending the system dimensions with additional state variables. Hyperchaotic systems have many applications in multimedia data

encryption, and its system equation is defined in Eq. (4):

$$\begin{cases} \dot{x_1} = a(x_2 - x_1) \\ \dot{x_2} = bx_1 - x_2 + ex_4 - x_1x_3 \\ \dot{x_3} = -cx_3 + x_1x_2 + x_1^2 \\ \dot{x_4} = -dx_2 \end{cases} \tag{4}$$

where x_1, x_2, x_3, x_4 are state variables, a, b, c, d, e are positive real parameters of the system. When the initial value $K=(x_0, y_0, z_0, w_0)$ is used as the system key, the four-dimensional chaotic sequence $\boldsymbol{X}=\{x(i), 1\leq i\leq N\}$, $\boldsymbol{Y}=\{y(i), 1\leq i\leq N\}$, $\boldsymbol{Z}=\{z(i), 1\leq i\leq N\}$, $\boldsymbol{W}=\{w(i), 1\leq i\leq N\}$ can be generated for encryption. Where, N is the number of iterations.

The experience shows that when a=10, b=28, c=8/3, d=1, e=16, and initial value K is (1, 1, 1, 1), the Lyapunov index is (0.416, 0.1348, -0.0014, -14.22). Since there are two positive Lyapunov exponents in the system, it is obvious that the system is in hyperchaotic state and has more complex dynamic characteristics.

4. The proposed scheme

4.1 System model

Fig. 3 shows our proposed encrypted speech retrieval system model based on LSTM and deep hashing. The model mainly includes the construction of encrypted speech library, deep hashing construction and generation system hashing index table, and user speech retrieval.

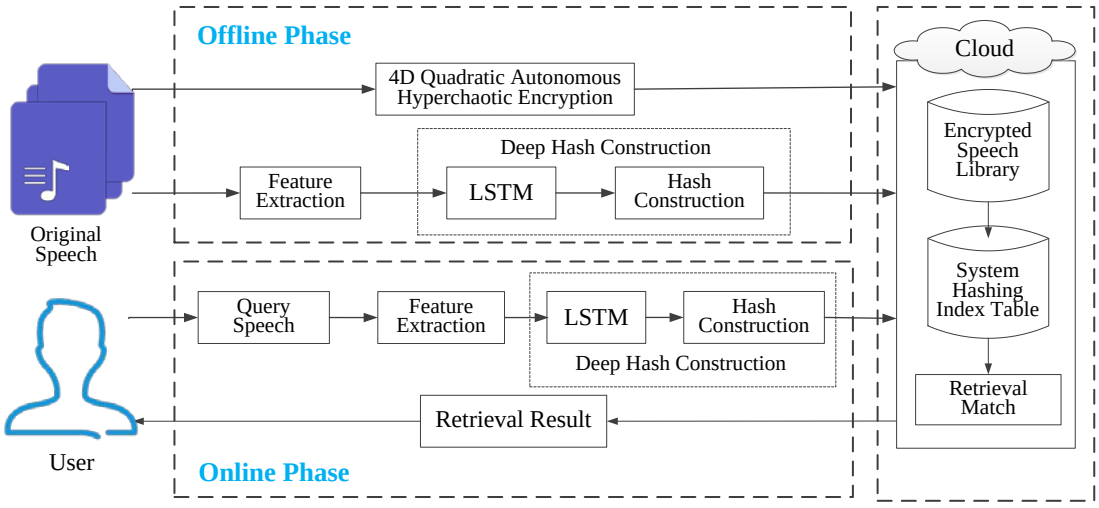


Fig. 3. Encrypted speech retrieval system model based on LSTM and deep hashing

As shown in **Fig. 3**, for the construction of the encrypted speech library, the original speech is encrypted by the 4D quadratic autonomous hyperchaotic encryption algorithm and uploads to the encrypted speech library in the cloud. For the deep hashing construction and generation system hashing index table, the Log-Mel Spectrogram/MFCC of the original speech is first extracted as the training data to pre-train LSTM model. Then deep features extracted from the trained LSTM model is combined with hash function to construct binary deep hash sequence of the speech. A one-to-one mapping relationship with the encrypted

speech is established and uploads to the system hashing index table in the cloud. In the user speech retrieval, the same deep hashing construction method is used to generate the binary deep hashing code of the query speech. Through the normalized Hamming distance algorithm, the generated deep hash sequence is matched in the system hash index table and the retrieval result is returned.

In this model, the encrypted speech library and the system hashing index table are constructed offline, and the user speech retrieval can generate a retrieval index online.

4.2 Construction of encrypted speech library

To ensure the privacy and security of speech data stored in the cloud, the original speech is encrypted by the 4D quadratic autonomous hyperchaotic encryption algorithm described in Section 3.3. **Fig. 4** shows the flow chart of 4D quadratic autonomous hyperchaotic speech encryption.

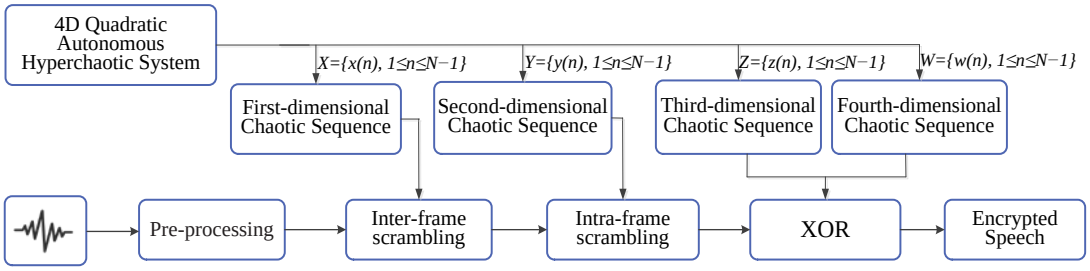


Fig. 4. Flow chart of 4D quadratic autonomous hyperchaotic speech encryption

The speech encryption process is as follows:

Step 1: Pre-processing. The original speech $\boldsymbol{S}=\{s(i), 1\leq i\leq L\}$ is divided into V frames of length N per frame, and each speech frame is represented as $\boldsymbol{S}_x(j)$, where N=256, V=250, L=64,000.

Step 2: Inter-frame scrambling. All speech frames $\boldsymbol{S}_x=\{\boldsymbol{S}_x(j), 1\leq j\leq V\}$ are scrambled using the first-dimensional chaotic sequence $\boldsymbol{X}=\{x(i), 1\leq i\leq V\}$ generated by 4D quadratic autonomous hyperchaotic system. Firstly, the elements of the first-dimensional chaotic sequence $\boldsymbol{X}$ are sorted in ascending order to get $\boldsymbol{X}'$. Then the original position index $\boldsymbol{I}_x$ sequence corresponding to each element of $\boldsymbol{X}'$ is used as the scrambling sequence. Finally, the position scrambling of $\boldsymbol{S}_x$ is performed by $\boldsymbol{I}_x$, and the inter-frame scrambling speech $\boldsymbol{S}_y=\{\boldsymbol{S}_v(j), 1\leq j\leq V\}$ is obtained.

Step 3: Intra-frame scrambling. The scrambling operation is performed on the sample points of each intraframe $\boldsymbol{S}_y(j)$ using the second-dimensional chaotic sequence $\boldsymbol{Y}=\{y(i), 1\leq i\leq N\}$ generated by 4D quadratic autonomous hyperchaotic system. Firstly, the elements of the second-dimensional chaotic sequence $\boldsymbol{Y}$ are sorted in ascending order to get $\boldsymbol{Y}'$. Then the original position index $\boldsymbol{I}_y$ sequence corresponding to each element of $\boldsymbol{Y}'$ is used as the scrambling sequence. Finally, the position scrambling of $\boldsymbol{S}_v(j)$ is performed by $\boldsymbol{I}_y$, and the intra-frame scrambling speech $\boldsymbol{S}_z(j)$ is obtained.

Step 4: XOR diffusion. The third-dimensional chaotic sequence $\boldsymbol{Z}=\{z(i), 1\leq i\leq N\times V\}$ and the fourth-dimensional chaotic sequence $\boldsymbol{W}=\{w(i), 1\leq i\leq N\times V\}$ generated by 4D quadratic autonomous hyperchaotic system are used to diffuse the scrambled one-dimensional speech $\boldsymbol{S}z=\{\boldsymbol{S}z(i), 1\leq i\leq L\}$ forward and backward respectively by Eq. (5) and Eq. (6).

$$S_z'(i)=S_z'(i-1)\oplus z(i)\oplus S_z(i) \tag{5}$$

$$S_z''(i) = S_z''(i+1) \oplus w(i) \oplus S_z'(i) \tag{6}$$

where $\boldsymbol{S}''_z(i)$ is each sample point after XOR diffusion, i=1, 2, ..., L.

Step 5: Restore the speech. Finally, it is reconstructed into time domain speech, and the encrypted speech signal $\boldsymbol{S}'=\{\boldsymbol{S}''_z(i), 1\leq i\leq L\}$ is obtained.

Step 6: Construction of encrypted speech library. The above encryption process is performed on all the original speech in the original speech library, and uploads to the encrypted speech library in the cloud.

4.3 LSTM model construction

Fig. 5 shows the LSTM model of the proposed scheme. This scheme uses the LSTM model to learn speech representation, which can break through the limitations of hand-crafted features.

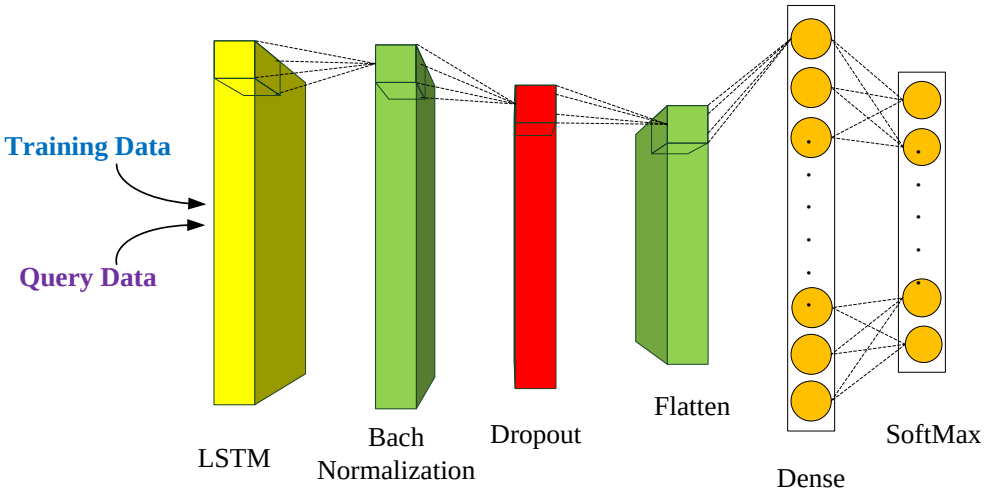


Fig. 5. The LSTM model proposed in this paper

As shown in **Fig. 5**, the framework consists mainly of LSTM and 2 fully connected layers. The number of filters in the LSTM is 128, and the Batch Normalization algorithm is introduced to improve the network fitting speed and reduce the training time. In addition, the Dropout layer is introduced to prevent Over-fitting during the training process of the deep learning network, and the experimentally verified setting parameter is 0.75. The Flatten layer is designed to turn data into one-dimensional data for input to the fully connected layer. The fully connected layer Dense contains 384 nodes as the feature extraction layer, and the rectification linear unit (ReLU) is selected as the nonlinear activation function. Finally, the speeches with the same content are divided into one class, which is divided into 10 class. And the SoftMax layer is used as the network output layer. In the experimental process, we experimented by adding or deleting layers and modifying parameters, and verified the optimal results of the network model through loss, test accuracy, recall and precision.

The LSTM model shown in **Fig. 5**, and it is implemented by Python's Keras library. The loss function for training is binary cross entropy. The optimization algorithm is stochastic gradient descent (SGD).

4.4 Deep hashing construction

The research has shown that the features extracted in the fully connected layer have been widely used in image and audio fields. However, due to the extracted features are high-dimensional vectors, the retrieval efficiency is very low in large corpora. In this paper, we convert the extracted feature vector into binary sequence to facilitate effective speech retrieval and reduce computational cost. The Hamming distance can be used to achieve fast retrieval and matching. The construction process of the binary deep hashing sequence is as follows:

Step 1: Speech feature extraction. The Log-Mel spectrogram and MFCC are extracted from the speech using the feature extraction method described in Section 3.1. In the feature extraction stage, the Librosa library [37] extracts audio features at a sampling rate of 16 kHz with frame length and frame shift set to 25 ms and 10 ms, respectively. And the Hamming window function is used. The Log-Mel spectrogram/MFCC feature coefficients is 20 dimensions. Since the speech in the experimental dataset has different durations, the input speech duration is fixed to 4s.

Step 2: Deep feature extraction. The extracted features as input to trained LSTM model in Fig. 5, and the deep features $\boldsymbol{H}=\{H(i)|i=1, 2, \ldots, M\}$ are extracted from the fully connected layer Dense.

Step 3: Deep hashing construction. The extracted deep feature $\boldsymbol{H}=\{H(i)|i=1, 2, \ldots, M\}$ generates hashing sequence $\boldsymbol{h}=\{h(i)|i=1, 2, \ldots, M\}$ by Eq. (7), where M is the number of extracted features (the number of Dense nodes).

$$h(i)=\begin{cases}1 & H(i)>H_{meadian}\\ 0 & H(i)\le H_{meadian}\end{cases},i=1, 2,\cdots,M \tag{7}$$

where $H_{meadian}$ is the median of the feature vector $\boldsymbol{H}$, and M is the length of the deep binary hashing sequence (the number of Dense nodes is 384).

Step 4: Construct a system hashing index table. According to the above steps, the deep hashing sequences ($\boldsymbol{h}_1$, $\boldsymbol{h}_2$, …, $\boldsymbol{h}_x$) of all the original speeches ($\boldsymbol{S}_1$, $\boldsymbol{S}_2$, …, $\boldsymbol{S}_N$) are obtained. The deep hashing sequence generated establish a one-to-one mapping relationship of *Key-Value* with the encrypted speech, and upload to the system hashing index table in the cloud.

4.5 User speech retrieval

When querying the speech, the querying user submits the speech to be queried online, and the encrypted speech retrieval can be performed by means of “no downloading, no decryption”. The speech retrieval process is as follows:

Step 1: Submit the query speech. Given the speech q to be queried, the Dense layer output $\boldsymbol{H}_q$ of the trained LSTM model is first extracted as a deep feature, and then the deep binary code $\boldsymbol{h}_\mathbf{q}$ is obtained by the Eq. (7).

Step 2: Retrieve the matching. The deep hashing sequence $\boldsymbol{h}_\mathbf{q}$ generated in S**tep 1** and the hash sequence $\boldsymbol{h}_\mathbf{x}$ in the system hashing index table are matched by the normalized Hamming distance (also known as the bit error rate, BER) algorithm $D(\boldsymbol{h}_\mathbf{x}, \boldsymbol{h}_\mathbf{q})$. The BER is calculated as shown in Eq. (8):

$$D(h_x,h_q)=\frac{1}{M}\sum_{i=1}^{M}\left(\left|h_x(i)-h_q(i)\right|\right)=\frac{1}{M}\sum_{i=1}^{M}h_x(i)\otimes h_q(i),i=1, 2, \cdots,M \tag{8}$$

where M is the length of the deep binary hashing sequence.

In retrieval, the threshold $T(0<T<0.5)$ is set as the similarity measurement threshold. If $D(\boldsymbol{h}_x, \boldsymbol{h}_q)<T$, the retrieval is successful and the decrypted speech is returned to user; otherwise, the retrieval is fail. And the retrieval accuracy is very relevant to the setting of the threshold T.

4.6 Speech decryption

The retrieved speech is decrypted, and the decryption process is the reverse process of encryption. The speech decryption process is as follows:

Step 1: Import the encrypted speech $\boldsymbol{S}z=\{\boldsymbol{S}z(i), 1\le i\le L\}$ where L=64,000, and use the same key as the encryption to generate the chaotic sequence using the 4D quadratic autonomous hyperchaotic system.

Step 2: XOR diffusion. The fourth-dimensional chaotic sequence $\boldsymbol{W}=\{w(i), 1 \le i \le N\times V\}$ and the third-dimensional chaotic sequence $\boldsymbol{Z}=\{z(i), 1 \le i \le N\times V\}$ generated by 4D quadratic autonomous hyperchaotic system, and use chaotic sequences to inverse diffuse the encrypted speech $\boldsymbol{S}z=\{\boldsymbol{S}z(i), 1\le i\le L\}$ by Eq. (9) and Eq. (10).

$$S'_z(i) = S'_z(i+1) \oplus w(i) \oplus S_z(i) \tag{9}$$

$$S''_z(i) = S''_z(i-1) \oplus z(i) \oplus S'_z(i) \tag{10}$$

where $\boldsymbol{S}''_z(i)$ is each sample point after XOR diffusion, i=1, 2, ..., L.

Step 3: Intra-frame scrambling. For the speech $\boldsymbol{S}''_z=\{\boldsymbol{S}''_z(i), 1\le i\le L\}$ in **Step 2** is divided into V frames of length N per frame, then the speech of each frame is represented as $\boldsymbol{S}''_z(j)$, where N=256, V=250, L=64,000. Then, the intraframe sampling points are scrambled using the second-dimensional chaotic sequence $\boldsymbol{Y}=\{y(i), 1 \le i \le N\}$ generated by the 4D quadratic autonomous hyperchaotic system. Firstly, the elements of the second-dimensional chaotic sequence $\boldsymbol{Y}$ are sorted in ascending order to get $\boldsymbol{Y}'$. Then the original position index $\boldsymbol{I}_y$ sequence corresponding to each element of $\boldsymbol{Y}'$ is used as the scrambling sequence. Finally, the position scrambling of $\boldsymbol{S}''_z(j)$ is performed by $\boldsymbol{I}_y$, and the intra-frame scrambling speech $\boldsymbol{S}_y(j)$ is obtained.

Step 4: Inter-frame scrambling. All speech frames $\boldsymbol{S}_y=\{\boldsymbol{S}_y(j), 1 \le j\le V\}$ are scrambled using the first-dimensional chaotic sequence $\boldsymbol{X}=\{x(i), 1 \le i \le V\}$ generated by 4D quadratic autonomous hyperchaotic system. Firstly, the elements of the first-dimensional chaotic sequence $\boldsymbol{X}$ are sorted in ascending order to get $\boldsymbol{X}'$. Then the original position index $\boldsymbol{I}_x$ sequence corresponding to each element of $\boldsymbol{X}'$ is used as the scrambling sequence. Finally, the position scrambling of $\boldsymbol{S}_y$ is performed by $\boldsymbol{I}_x$, and the inter-frame scrambling speech $\boldsymbol{S}_x=\{\boldsymbol{S}_x(j), 1 \le j\le V\}$ is obtained.

Step 5: Restore the speech. Finally, the speech $\boldsymbol{S}_x=\{\boldsymbol{S}_x(j), 1 \le j\le V\}$ obtained in **Step 4** is reconstructed into time domain speech, and the decrypted speech signal $\boldsymbol{S}=\{s(i), 1\le i\le L\}$ is obtained.

5. Experimental results and performance analysis

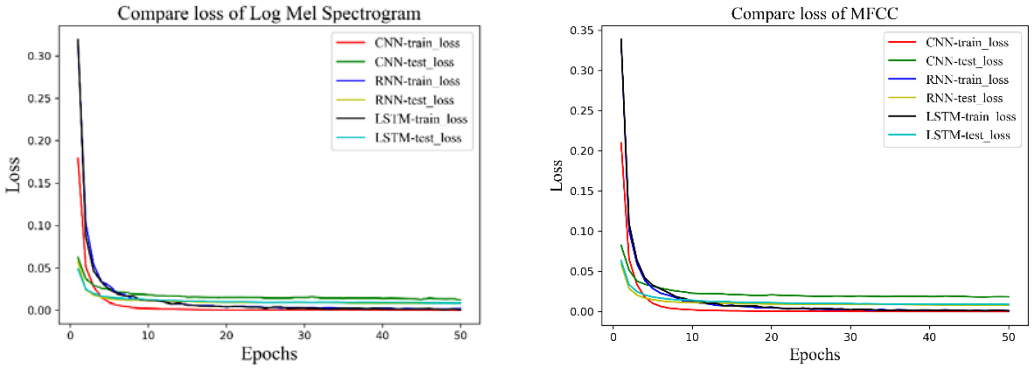
We use the speech in the THCHS-30 [38] as the experimental data. It is an open Chinese speech database published by the center for speech and language technology (CSLT) of Tsinghua University. A single-channel wav format speech segment with a frequency of 16 kHz and a sampling accuracy of 16 bits. In the LSTM model training stage, 10 segments of

speech with different contents are selected from 17 different speakers. We perform17 speech content preserving operations (CPOs) including amplitude adjustment, noise addition, re-quantization, resampling, and MP3. A total of 3,060 speeches were obtained. The robustness of the system is improved while increasing the amount of data. In the performance analysis, in order to better verify the feasibility, 1,000 speeches with 4s are randomly selected in the speech library for evaluation. In order to better test the retrieval efficiency, 10,000 speeches with 4s are randomly selected for evaluation.

The experimental hardware platform is: Intel(R) Core (TM) i7-8750H CPU, 2.20GHz, 8GB of memory. The software environment is: Windows 10, MATLAB R2017b, JetBrains PyCharm Community Edition 2019.1.3 x64.

5.1 Performance analysis of the proposed LSTM model

LSTM model is the most important part of encrypted speech retrieval. Our main goal is to propose a deep learning network to extract deep semantic feature of speech, and test the performance of the model through pre-training model. **Fig. 6** shows the train/test loss curves of Log-Mel Spectrogram/MFCC features in CNN, RNN and LSTM models.



(a) Log-Mel Spectrogram (b) MFCC

Fig. 6. Log-Mel spectrogram and MFCC train/test loss curves

As shown in **Fig. 6**, the LSTM model converges faster and loses less in the test phase. This is because although CNN model still has performance advantages in classification, it is still relatively scarce in dealing with some sequential tasks. As an improved version of RNN model, LSTM network model inherits the characteristics of most RNN models, which makes LSTM more suitable for handling problems highly related to time sequence.

To better compare the reliability of CNN, RNN and LSTM models, the train accuracy of the models for the two features can be calculated. **Table 1** shows the train accuracy of CNN, RNN and LSTM models with Log-Mel Spectrogram/MFCC is used as input.

Table 1. Performance comparison of test accuracy

Networks	Methods	Test Accuracy (%)
CNN	Log-Mel Spectrogram	99.59
	MFCC	99.48
RNN	Log-Mel Spectrogram	99.66
	MFCC	99.61
LSTM	Log-Mel Spectrogram	99.69
	MFCC	99.64

It can be seen from **Table 1** that the Log-Mel Spectrogram and MFCC have different performances in the CNN, RNN and LSTM network models. The accuracy of using the LSTM model is significantly higher than other network models.

To further test the feasibility of CNN, RNN and LSTM network models, the Eq. (11) is used to calculate the AP (Average Precision) for the speech after different speech CPOs. Then the mAP (mean Average Precision) of **Table 2** is obtained by the Eq. (12).

$$AP(q) = \frac{\sum_{k}^{n}(\frac{1}{k} \times rel(k))}{m}, rel(k) \in \{0,1\} \tag{11}$$

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{12}$$

where Q represents the number of query, $AP(q)$ represents the query accuracy of the q-th, and n represents the number of speeches, $rel(k)$ represents whether the retrieved k-th speech is related to the query speech (correlation 1, irrelevant 0).

Table 2. Performance comparison of *mAP* accuracy

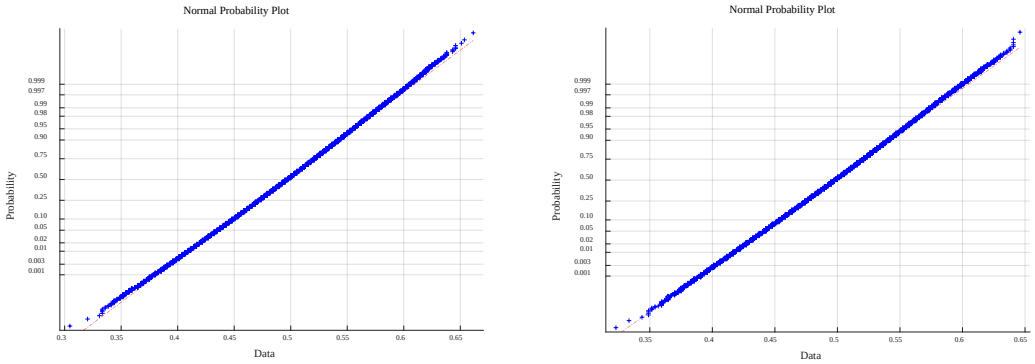
Networks	Methods	mAP (%)
CNN	Log-Mel Spectrogram	91.58
	MFCC	85.39
RNN	Log-Mel Spectrogram	92.01
	MFCC	87.18
LSTM	Log-Mel Spectrogram	92.45
	MFCC	89.32
Ref. [32]	MFCC	56
	Walnet	72

The larger the mAP value, the better the retrieval algorithm. **Table 2** shows that the LSTM method in this paper is superior to CNN, RNN and Ref. [32]. It can be concluded that the LSTM model can more truly represent or simulate the cognitive process of human behavior, logic and nerve organization, so it is more suitable for learning time sequence.

5.2 Performance comparison with existing perceptual hashing methods

In the proposed scheme, LSTM model can achieve better retrieval results. And the normalized Hamming distance (also known as BER) is used to analyze the discrimination and robustness of the proposed deep hashing algorithm. The degree of similarity between speeches can be determined by calculating the BER between deep hash sequences. The BER obtained from hash sequences of different speech content is basically obeys normal distribution. To better verify the discrimination at different thresholds, the false accept rate (FAR) as shown in Eq. (13) is introduced. Meanwhile, to better test the reliability of the algorithm, 1,000 speeches were randomly selected for evaluation in the THCHS-30. Through pairwise matching of hashing codes on 1,000 speeches, and 1,000×999/2 = 499,500 BER data are obtained.

Fig. 7 shows the BER normal probability distribution of the log-Mel Spectrogram and MFCC in 1,000 speeches.



(a) Log-Mel Spectrogram (b) MFCC

Fig. 7. The BER normal probability distribution

It can be seen from **Fig. 7**(a) and **Fig. 7**(b) that the probability distribution of the BER values almost overlaps with the probability curve of the normal distribution, the binary deep hashing sequence obtained by the proposed algorithm basically obeys the normal distribution.

According to the De Moivre-Laplace central limit theorem, the BER is used as the similarity measure and the BER approximates the normal distribution $\mu = p, \delta = \sqrt{p(1-p)/N}$, where N is the hashing sequence length, μ is the BER mean, δ is the BER standard deviation, and p is the probability of the hash sequence 0, 1 occurring. The more the distribution curves of BER coincide, the better the performance of our algorithm. In this paper, the length of the deep hash sequence is n = 384, which can calculate the mean value μ=0.5 and standard deviation δ=0.0255 of the parameters of the theoretical normal distribution. In the experiment, the BER mean of Log-Mel spectrogram is μ_0=0.4963, and the standard deviation is δ_0=0.0364. The BER mean of MFCC is μ_1=0.4977, and the standard deviation δ_1=0.0339. The FAR with different thresholds τ can be calculated according to Eq. (13).

$$FAR(\tau) = \int_{-\infty}^{\tau} f(x \mid \mu, \delta)\, dx = \int_{-\infty}^{\tau} \frac{1}{\delta\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\delta^2}} dx \tag{13}$$

where τ is the threshold of hash matching, μ is the BER mean, δ is the BER standard deviation, and the x is BER.

Table 3 shows the comparison results between the proposed scheme and the existing perceptual hashing based encrypted speech retrieval methods Ref. [3, 5, 6, 7] at different thresholds. The lower FAR of perceptual hashing algorithm, the higher the anti-collision performance and the better the discrimination of the algorithm.

Table 3. Comparison of FAR with existing methods at different thresholds

τ	Log-Mel	MFCC	Ref. [3]	Ref. [5]	Ref. [6]	Ref. [7]
0.02	2.00×10^{-39}	2.14×10^{-45}	1.88×10^{-29}	4.27×10^{-26}	8.44×10^{-42}	4.78×10^{-38}
0.04	2.38×10^{-36}	7.66×10^{-42}	4.22×10^{-27}	4.25×10^{-24}	1.60×10^{-38}	4.22×10^{-35}
0.06	2.10×10^{-33}	1.94×10^{-38}	7.58×10^{-25}	3.48×10^{-22}	2.20×10^{-35}	2.81×10^{-32}
0.08	1.37×10^{-30}	3.47×10^{-35}	1.07×10^{-22}	2.35×10^{-20}	2.20×10^{-32}	1.40×10^{-29}
0.10	6.62×10^{-28}	4.39×10^{-32}	1.20×10^{-20}	1.30×10^{-18}	1.59×10^{-29}	5.25×10^{-27}
0.12	2.37×10^{-25}	3.93×10^{-29}	1.05×10^{-18}	5.96×10^{-17}	8.40×10^{-27}	1.48×10^{-24}
0.14	6.31×10^{-23}	2.50×10^{-26}	7.26×10^{-17}	2.25×10^{-15}	3.21×10^{-24}	3.15×10^{-22}
0.16	1.24×10^{-20}	1.12×10^{-23}	3.95×10^{-15}	6.97×10^{-14}	8.90×10^{-22}	5.02×10^{-20}

As shown in **Table 3**, the FAR of MFCC are lower than the Ref. [3, 5, 6, 7] under different thresholds. The FAR of the Log-Mel Spectrogram is lower than the Ref. [3, 5, 7] at different thresholds, which is very close to Ref. [6]. Therefore, the method has strong anti-collision and discrimination, which can meet the retrieval needs.

Robustness means that the speech has the same hash code as the speech processed by the CPOs. For the robustness of the experimental algorithm, the experiment used software Gold Wave 6.38 and MATLAB R2017b to perform content preserving operation with 1,000 test speeches. **Table 4** shows the average BER after 5 operations such as MP3 compression (128 kbps, MP3), re-quantization (16→8→16bit, R.Q), amplitude increase or decrease of 3 dB (+3dB, -3dB), and 30 dB narrowband Gaussian noise (G.N).

It can be seen from **Table 4**, the Log-Mel spectrogram is more robust than the MFCC. The BER of Log-Mel spectrogram is 9.609×10^{-4} after a 3 dB decrease in amplitude, which is less than Ref. [3, 5, 6]. The BER is 0.0221after the 3 dB increase in amplitude, less than Ref. [5]. And the BER is 0.1291 and 0.2310 after the R.Q and G.N respectively, less than Ref. [3]. For the MFCC, the BER is 0.0019after the 3 dB decrease in amplitude which is less than Ref. [5, 6]. And the BER is 0.0316 after the 3 dB increase in amplitude, less than the literature Ref. [5]. The BER is 0.2698 after G.N, less than Ref. [3]. The robustness is lower than Ref. [7] mainly because Log-Mel Spectrogram and MFCC are less robust, and the LSTM network model has less data during pre-training.

Table 4. Comparison of Robustness with existing methods at different CPOs

CPOs	Log-Mel	MFCC	Ref. [3]	Ref. [5]	Ref. [6]	Ref. [7]
MP3	0.0567	0.1082	0.0177	0.0038	0.0016	0.0090
R.Q	0.1291	0.1621	0.1354	0.0693	0.0026	-
-3dB	9.609×10^{-4}	0.0019	0.0018	0.0139	0.0042	0.0038
+3dB	0.0221	0.0316	0.0183	0.0476	0.0039	0.0160
G.N	0.2310	0.2698	0.2719	-	-	0.0248

5.3 Analysis of retrieval performance

Recall and precision are important indexes to measure the performance of retrieval algorithm. The calculation methods of the recall rate R and the precision rate P are shown in Eq. (14) and Eq. (15), respectively.

$$R = \frac{f_T}{f_T + f_L} \times 100\% \tag{14}$$

$$P = \frac{f_T}{f_T + f_F} \times 100\% \tag{15}$$

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 retrieval, the similarity threshold T ($0<T<0.5$) is set. And the retrieval is successful if the normalized Hamming distance $D(\boldsymbol{h}_x, \boldsymbol{h}_q)<T$. The setting of the threshold directly determines the recall R and the precision P of the retrieval algorithm. For the discrimination experimental, the minimum BER values of Log-Mel spectrogram/MFCC in 1,000 speeches were 0.3047 and 0.3229, respectively. For the robustness experimental, the maximum BER values are 0. 2370 and 0.2943, respectively. To avoid miss retrieval and achieve high performance, the Log-Mel spectrogram/MFCC similarity thresholds were set to T_0=0.30 and

T_1=0.32, respectively. **Table 5** shows the recall rate R and the precision rate P calculated by the Eq. (14) and Eq. (15).

Table 5. Comparison of the recall rate R and the precision rate P with existing methods under different CPOs

	Methods	MP3	R.Q	-3dB	+3dB	G.N
Recall rate R	Log-Mel	100%	100%	100%	100%	99.0%
	MFCC	100%	100%	100%	100%	93.0%
	Ref. [3]	92%	95%	97%	96%	-
	Ref. [5]	100%	100%	100%	100%	-
	Ref. [6]	96%	98%	-	-	98%
	Ref. [7]	100%	-	100%	100%	100%
Precision rate P	Log-Mel	100%	100%	100%	100%	100%
	MFCC	99.9%	100%	100%	100%	100%
	Ref. [3]	92%	92%	96%	93%	-
	Ref. [5]	100%	100%	100%	100%	-
	Ref. [6]	96%	97%	-	-	97%
	Ref. [7]	100%	-	100%	100%	100%

As shown in **Table 5**, the Log-Mel Spectrogram/MFCC methods proposed in this paper can still retain a high recall rate R and a precision rate P after several content preserving operations. And compared with the Ref. [3, 5, 6, 7] under several content retention operations, the recall rate R and precision rate P are similar or even better than those. The retrieval recall rate R and precision rate P of this method after CPOs are superior to Ref. [3, 6] except for the G.N operation, and the performance is comparable to Ref. [5, 7]. Therefore, our algorithm has good retrieval performance under several content retention operations.

The recall and precision are mutually influential, and ideally both are high. But in general, the higher the recall, the lower the precision. Drawing Precision-Recall curve can visually observe the interaction between recall and precision. **Fig. 8** shows a comparison of the Precision-Recall curve between the proposed scheme and the existing algorithm [3, 5, 6, 7].

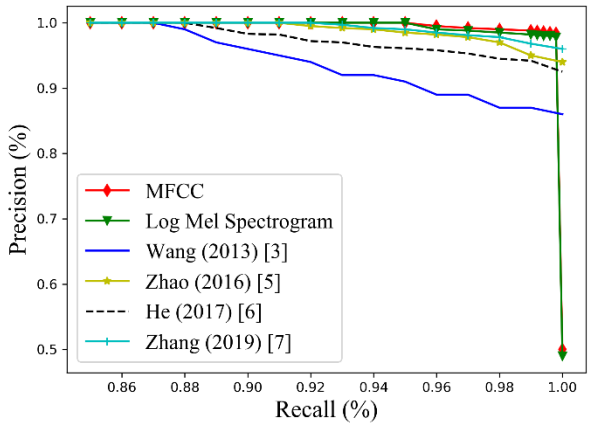
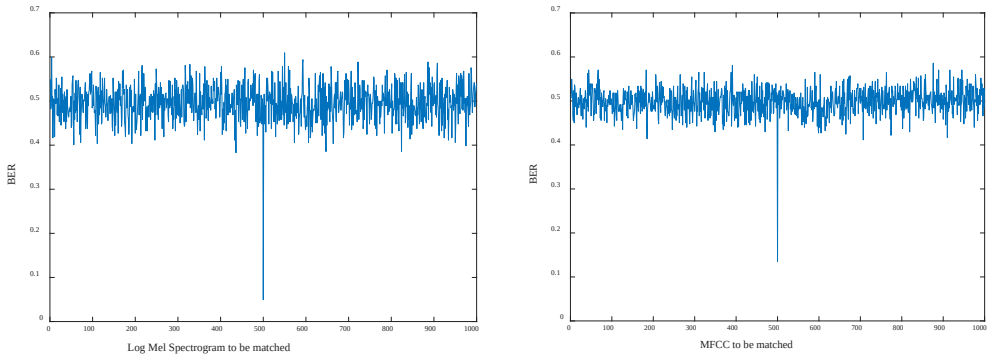


Fig. 8. Precision-Recall curve of Log-Mel Spectrogram/MFCC

As shown in **Fig. 8**, the larger the area enclosed by Precision-Recall curve and X-Y coordinate axis, the better the retrieval performance of the retrieval algorithm. Since recall and precision are mutually influential, the method in this paper has the greatest impact on

precision rate when recall is 1. The area of the Precision-Recall curve shown in **Fig. 8** is larger than Ref. [3, 5, 6, 7]. Therefore, our algorithm has good retrieval performance.

For the speech retrieval experiment, all query speeches are processed through 5 kind of CPOs, and then matched in the system hashing index table. **Fig. 9** is an example in which the 500-th speech as query speech. The hash value is obtained after MP3 operation, and the BER is calculated to get the retrieval result.



(a) Log-Mel Spectrogram (b) MFCC

Fig. 9. Log-Mel spectrogram/MFCC matching results in system hashing index table

As shown in **Fig. 9**, except for the BER between the query speech and the 500-th speech in the system hashing index table, the other BERs are greater than the thresholds T_0=0.30 and T_1=0.32, the retrieval is successful.

For the retrieval efficiency of the algorithm, 10,000 speeches were randomly selected in the THRHS-30 for evaluation. The average retrieval time of the algorithm is calculated and compared with Ref. [3, 5, 6, 7]. **Table 6** shows the experimental results.

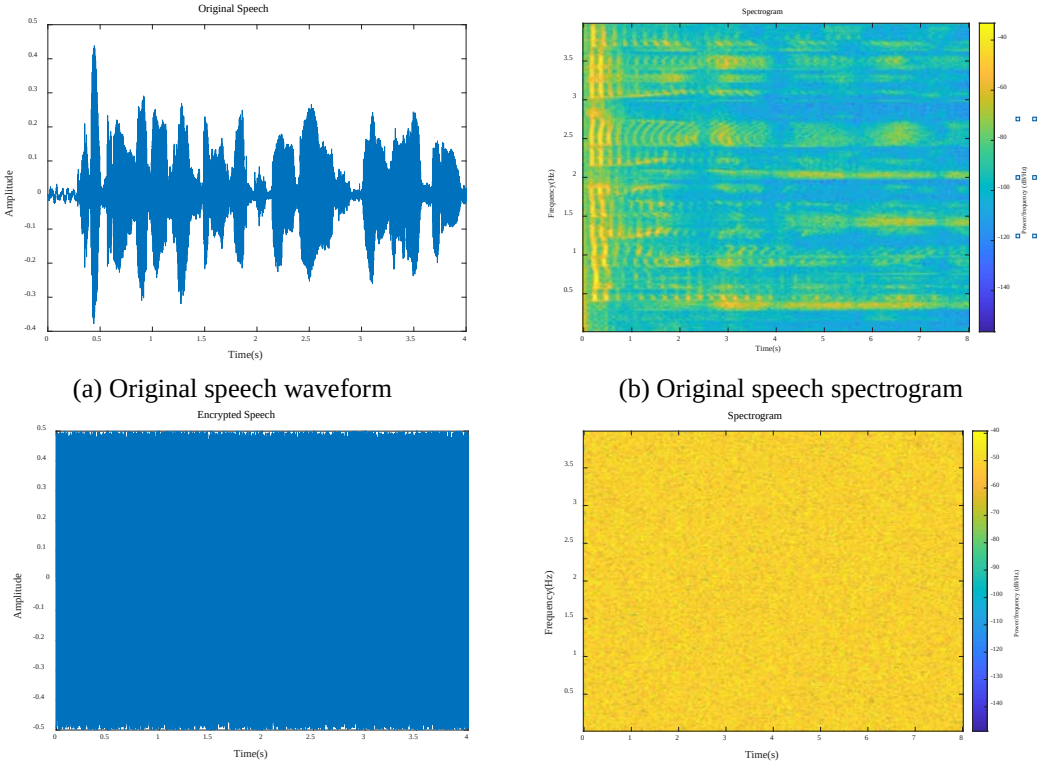
Table 6. Comparison of retrieval efficiency with existing methods

Methods	Frequency (GHz)	Speech length (*s*)	Average running time (*s*)
Ref. [3]	1.60	4	0.2613
Ref. [5]	3.20	4	3.7937
Ref. [6]	3.20	4	4.2032
Ref. [7]	2.90	4	0.4932
Log-Mel Spectrogram	2.20	4	0.5548
MFCC	2.20	4	0.5198

Average retrieval time is an important index to evaluate the retrieval system. It can be seen from **Table 6**, the Log-Mel Spectrogram/MFCC methods proposed in this paper have similar average retrieval time, and the retrieval efficiency is higher than that Ref. [5, 6]. The retrieval efficiency is about 7 times of Ref. [5], 8 times of Ref. [6] and the retrieval efficiency is similar to Ref. [7]. This is because the method in this paper is based on Log-Mel Spectrogram/MFCC to further extract the deep features, which reduce the dimension of the original features, and can use a shorter hash code to achieve efficient retrieval. Ref. [5, 6] only extract speech features and then construct perceptual hash for retrieval. The retrieval efficiency is lower than Ref. [3,7] because Ref. [3,7] uses relatively simple feature extraction methods with low feature dimension and the LSTM network model is computationally time consuming. Experimental results show that our scheme has better retrieval efficiency.

5.4 Security analysis

A good encryption system must have a large enough key space to defend against exhaustive attacks. For the encryption, when the key space of the algorithm is larger than $2^{100} \approx 10^{30}$, it can resist exhaustive attacks. We encrypt the speech data using the speech encryption algorithm described in Section 4.2 and the selected key is $\boldsymbol{K}$= (1,1,1,1). **Fig. 10** is the waveform and spectrogram of original speech and encrypted speech. Where **Fig. 10**(a) is the original speech waveform, **Fig. 10**(b) is the original speech spectrogram, **Fig. 10**(c) is the encrypted speech waveform and **Fig. 10**(d) is the encrypted speech spectrogram.



(a) Original speech waveform (b) Original speech spectrogram

(c) Encrypted speech waveform (d) Encrypted speech spectrogram

Fig. 10. Waveform and spectrogram of original speech and encrypted speech

As can be seen from **Fig. 10**(c), the encrypted speech waveforms are evenly distributed, similar to noise waveforms, and there are few features that can be utilized. Since the pixel points of the encrypted speech spectrum **Fig. 10**(d) are randomly distributed, and there are no spectral peaks of any speech. The results show that the chaotic effect of the algorithm is good and the security is high. In this paper, the speech encryption algorithm uses the initial value $\boldsymbol{K}$= (x_0, y_0, z_0, w_0) as the key and double-precision floating-point data accurate to 12 decimal places is used. The key space can reach $2\times10^{12}\times2\times10^{12}\times2\times10^{12}\times2\times10^{12}=16\times10^{48}\approx2^{164}$ and the key space is huge. If the system parameters a, b, c, d, e and the number of iterations are taken into account, the key space will be larger enough to resist exhaustive attack.

To evaluate the encrypted speech algorithm, we analyzes the speech quality perceptual evaluation (PESQ) of encrypted speech and decrypted speech. PESQ [39] is the mean opinion score (MOS) recommended by the Telecommunication Standardization Sector (ITU-T) P.862 from 1.0 (worst) to 4.5 (best) PESQ-MOS. It is generally expected that the encrypted speech PESQ-MOS can be decreased to 1.0 or lower, and the decrypted speech PESQ-MOS can increase to 2.5 or even higher. We randomly selected 20 speeches for experimentation. The average PESQ-MOS values obtained are shown in **Table 7**.

Table 7. PESQ-MOS for encrypted and decrypted speech

Type	PESQ-MOS	
	Our method	Ref. [7]
Encrypted speech	0.8888	1.0305
Decrypted speech	4.4997	4.5000

It can be seen from **Table 7**, the average encrypted speech PESQ-MOS is only 0.8888. From the experimental results, it can be concluded that the encrypted speech quality is extremely poor and the encryption algorithm is highly secure. The decrypted speech PESQ-MOS is 4.4997, indicating that the recovered speech has a higher auditory quality. Compared with Ref. [7], the encryption and decryption effect are better. Therefore, the speech encryption method can meet the security requirements of the system.

6 Conclusions

In this paper, we propose an encrypted speech retrieval scheme based on LSTM neural network and deep hashing. The proposed scheme not only realizes the efficient retrieval of massive speech in cloud environment, but also effectively avoids the risk of sensitive information leakage. Main contributions of this paper include two aspects: firstly, LSTM network is used to extract deep semantic features of speech and combining with hash function to generate deep hashing codes. The normalized Hamming distance is used to achieve retrieval matching, which improves the efficiency and accuracy of massive speech retrieval. Secondly, inspired by the existing privacy protection technology, speech encryption algorithm is implemented by 4D quadratic autonomous hyperchaotic system with good encryption performance, which can effectively enhance the security and privacy of speech data in cloud environment. Compared with the existing content-based encrypted speech retrieval methods, the proposed scheme has good discrimination and robustness. Meanwhile, it has a high recall, precision and retrieval efficiency, and the proposed speech encryption method has better security.

In addition, it is still insufficient that the gradient problem of RNN model has been solved in LSTM network model. Moreover, due to the time-consuming problem of LSTM network model, it is difficult to achieve efficient retrieval of irregular long speech. In the future, we will try to solve these problems.

Acknowledgments

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.

References

[1] F. Boucenna, O. Nouali, S. Kechid and M. T. Kechadi, "Secure Inverted Index Based Search over Encrypted Cloud Data with User Access Rights Management," *Journal of Computer Science and Technology*, vol. 34, no. 1, pp. 133-154, Jan. 2019. Article (CrossRef Link)

[2] C. Glackin, G. Chollet, N. Dugan, N. Cannings, J. Wall, S. Tahir and M. Rajarajan, "Privacy preserving encrypted phonetic search of speech data," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6414-6418, March 5-7, 2017. Article (CrossRef Link)

[3] H. Wang, L. Zhou, W. Zhang and H. Liu, "Watermarking-based perceptual hashing search over encrypted speech," in *Proc. of Int. Workshop on Digital Watermarking. Springer, Berlin, Heidelberg*, pp. 423-434, July 9, 2014. Article (CrossRef Link)

[4] H. X. Wang and G. Y. Hao, "Encryption speech perceptual hashing algorithm and retrieval scheme based on time and frequency domain change characteristics," *China Patent CN104835499A*, Aug 12, 2015. Article (CrossRef Link)

[5] H. Zhao and S. He, "A retrieval algorithm for encrypted speech based on perceptual hashing," in *Proc. of 12th Int. Conf. on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1840-1845, Aug 13-15, 2016. Article (CrossRef Link)

[6] S. He and H. Zhao, "A retrieval algorithm of encrypted speech based on syllable-level perceptual hashing," *Computer Science & Information Systems*, vol. 14, no. 3, pp. 703-718, 2017. Article (CrossRef Link)

[7] Q. Zhang, L. Zhou, T. Zhang and D. H. Zhang, "A retrieval algorithm of encrypted speech based on short-term cross-correlation and perceptual hashing," *Multimedia Tools and Applications*, vol. 78, no. 13, pp. 17825-17846, July, 2019. Article (CrossRef Link)

[8] Z. Dong, C. Jing, M. Pei and Y. Jia, "Deep CNN based binary hash video representations for face retrieval," *Pattern Recognition*, vol. 81, pp. 357-369, September, 2018. Article (CrossRef Link)

[9] J. Tang, Z. Li and X. Zhu, "Supervised deep hashing for scalable face image retrieval," *Pattern Recognition*, vol. 75, pp. 25-32, March, 2018. Article (CrossRef Link)

[10] L. Ma, H. Li, F. Meng, Q. Wu and K. Ngan, "Global and local semantics-preserving based deep hashing for cross-modal retrieval," *Neurocomputing*, vol. 312, pp. 49-62, October 27, 2018. Article (CrossRef Link)

[11] C. Deng, Z. Chen, X. Liu, X. Gao and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27 , no. 8, pp. 3893-3903, Aug. 2018. Article (CrossRef Link)

[12] Y. Cao, M. Long, B. Liu and J. Wang, "Deep cauchy hashing for hamming space retrieval," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1229-1237, 2018. Article (CrossRef Link)

[13] Y. Liu, J. Song, K. Zhou, L. Yan, L. Liu, F. Zou and L. Shao, "Deep self-taught hashing for image retrieval," *IEEE transactions on cybernetics*, vol. 49, no. 6, pp. 2229-2241, June, 2019. Article (CrossRef Link)

[14] J. Tang, J. Lin, Z. Li and J. Yang, "Discriminative deep quantization hashing for face image retrieval," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12, pp. 6154-6162, Dec. 2018. Article (CrossRef Link)

[15] S. L. Cheng, L. J. Wang, G. Huang and A. Y. Du, "A privacy-preserving image retrieval scheme based secure kNN, DNA coding and deep hashing," *Multimedia Tools and Applications*, pp.1-23, May, 2019. Article (CrossRef Link)

[16] I. Song, J. Chung, T. Kim and Y. Bengio, "Dynamic Frame Skipping for Fast Speech Recognition in Recurrent Neural Network Based Acoustic Models," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4984-4988, April 15-20, 2018. Article (CrossRef Link)

[17] M. Fujimoto and H. Kawai, "Comparative Evaluations of Various Factored Deep Convolutional Rnn Architectures for Noise Robust Speech Recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4829-4833, April 15-20, 2018. Article (CrossRef Link)
[18] M. Morchid, "Parsimonious memory unit for recurrent neural networks with application to natural language processing," *Neurocomputing*, vol. 314, pp. 48-64, November, 2018. Article (CrossRef Link)
[19] I. Korvigo, M. Holmatov, A. Zaikovskii and M. Skoblov, "Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules," *Journal of cheminformatics*, vol. 10, no. 1, pp. 28, May, 2018. Article (CrossRef Link)
[20] Y. Xu, Q. Kong, W. Wang and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 121-125, April 15-20, 2018. Article (CrossRef Link)
[21] J. Sang, S. Park and J. Lee, "Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms," in *Proc. of 26th European Signal Processing Conference (EUSIPCO)*, pp. 2444-2448, Sept. 3-7, 2018. Article (CrossRef Link)
[22] R. Pradeep and K. S. Rao, "Incorporation of Manner of Articulation Constraint in LSTM for Speech Recognition," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3482-3500, August, 2019. Article (CrossRef Link)
[23] S. Ghorbani, A. E. Bulut and J. H. L. Hansen, "Advancing Multi-Accented Lstm-CTC Speech Recognition Using a Domain Specific Student-Teacher Learning Paradigm," in *Proc. of IEEE Spoken Language Technology Workshop (SLT)*, pp. 29-35, Dec. 18-21, 2018. Article (CrossRef Link)
[24] Z. Yu, Z. Pengyuan and Y. A. N. Yonghong, "Long short-term memory with attention and multitask learning for distant speech recognition," *Journal of Tsinghua University (Science and Technology)*, vol. 58, no. 3, pp. 249-253, 2018. Article (CrossRef Link)
[25] Y. Xie, R. Liang, Z. Liang and L. Zhao, "Attention-Based Dense LSTM for Speech Emotion Recognition," *IEICE TRANSACTIONS on Information and Systems*, vol. 102, no. 7, pp. 1426-1429, July, 2019. Article (CrossRef Link)
[26] F. Tao and G. Liu, "Advanced LSTM: A study about better time dependency modeling in emotion recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2906-2910, April 15-20, 2018. Article (CrossRef Link)
[27] G. Ramet, P. N. Garner, M. Baeriswyl and A. Lazaridis, "Context-Aware Attention Mechanism for Speech Emotion Recognition." in *Proc. of IEEE Spoken Language Technology Workshop (SLT)*, pp. 126-131, Dec. 18-21, 2018. Article (CrossRef Link)
[28] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers and B. Schmauch, "Cnn+ lstm architecture for speech emotion recognition with data augmentation," in *Proc. of Workshop on Speech, Music and Mind 2018*, 21-25, 2018. Article (CrossRef Link)
[29] S. Jung, J. Park and S. Lee, "Polyphonic Sound Event Detection Using Convolutional Bidirectional Lstm and Synthetic Data-based Transfer Learning," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 885-889, May 12-17, 2019. Article (CrossRef Link)
[30] T. Matsuyoshi, T. Komatsu, R. Kondo, T. Yamada and S. Makino, "Weakly labeled learning using BLSTM-CTC for sound event detection," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1918-1923, Nov. 12-15, 2018. Article (CrossRef Link)
[31] J. Liu, Y. Yin, H. Jiang, H. Kan, Z. Zhang, P. Chen, B. Zhu and Z. Wang, "Bowel Sound Detection Based on MFCC Feature and LSTM," in *Proc. of IEEE Biomedical Circuits and Systems (BioCAS)*, pp. 1-4, Oct. 17-19, 2018. Article (CrossRef Link)
[32] B. Elizalde, S. Zarar and B. Raj, "Cross Modal Audio Search and Retrieval with Joint Embeddings Based on Text and Audio," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4095-4099, May 12-17, 2019. Article (CrossRef Link)

[33] S. B. Davis and P. Mermelstein, "Evaluation of acoustic parameters for monosyllabic word recognition in continuously spoken sentences," *The Journal of the Acoustical Society of America*, vol.64, pp. s180-s181, 1978. Article (CrossRef Link)
[34] Y. Xu, Q. Kong, W. Wang and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 121-125, April 15-20, 2018. Article (CrossRef Link)
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997. Article (CrossRef Link)
[36] H. Wang and G. Dong, "New dynamics coined in a 4-D quadratic autonomous hyper-chaotic system," *Applied Mathematics and Computation*, vol. 346, pp. 272-286, April 1, 2019. Article (CrossRef Link)
[37] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. of the 14th python in science conference*, pp. 18-24, 2015. Article (CrossRef Link)
[38] D. Wang and X. Zhang, "Thchs-30: A free Chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015. Article (CrossRef Link)
[39] ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs, ITU-T, Jan. 2002. Article (CrossRef Link)

Qiu-yu Zhang, Researcher/PhD supervisor, graduated from Gansu University of Technology in 1986, and then worked at school of computer and communication in Lanzhou University of Technology. He is vice dean of Gansu manufacturing information engineering research center, 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.

Yu-zhou Li, Master degree candidate, he received the BS degrees in communication engineering from Lanzhou University of Technology, Gansu, China, in 2016. His research interests include audio signal processing and application, multimedia authentication and retrieval techniques.

Ying-jie Hu, Doctoral candidate, she received the MS degree in computer software and theory from Lanzhou University, Lanzhou, China, in 2011, and now working as a lecturer in the school of computer and communication in Lanzhou University of Technology. Her research interests include multimedia information processing and application, information security, multimedia authentication and retrieval techniques.