Journal of Information Hiding and Multimedia Signal Processing ©2018 ISSN 2073-4212
Ubiquitous International Volume **9**, Number **6**, November **2018**

An Efficient Retrieval Method of Encrypted Speech Based on Frequency Band Variance

Qiu-Yu Zhang, Zi-Xian Ge, Si-Bin Qiao

School of Computer and Communication
Lanzhou University of Technology
Gansu, Lanzhou, 730050, P. R. China
zhangqylz@163.com; zxge727@foxmail.com; qiaosibin@163.com

Received June, 2018; revised August, 2018

Abstract. *In order to improve the retrieval accuracy and retrieval efficiency of the existing content-based encrypted speech retrieval methods, an efficient retrieval method of encrypted speech based on frequency band variance was proposed by using speech perceptual hashing technology. Firstly, the method uses Logistic chaos scrambling to encrypt the user's original speech file and upload it to the cloud encrypted speech library. Secondly, the original speech file is processed by applying pre-processing, framing, and adding window to calculate the frequency band variance of each frame. The obtained frequency band variance is used as a speech feature to construct a hashing, generate the binary perceptual hashing sequence and upload to the cloud's hash feature library. Finally, a hash sequence is generated for query speech and then the normalized Hamming distance algorithm is used to compare it with the cloud hash feature library. Experimental results show that the hash sequence constructed by the proposed method has good discrimination and robustness. Meanwhile, the retrieval performance is effectively improved.*

Keywords: Encrypted speech retrieval, Perceptual hashing, Frequency band variance, Logistic chaotic scrambling, Speech feature extraction.

1. Introduction.

Facing massive multimedia data, how to ensure the safety of the user's data, how to retrieve the required content of massive data accurately and quickly under the condition of data security. These problems have been a hot topic in the field of multimedia retrieval research [1]. At present, the explosive growth of digital audio on the internet has made the high-speed retrieval of audio big data a difficult problem to be solved [2, 3]. However, the content-based encrypted speech retrieval method can retrieve data efficiently and accurately under the premise of ensuring security, and the speech does not need to be downloaded or decrypted during the retrieval process [4], which is a good method to solve this problem.

Currently, the commonly used methods of speech retrieval mainly include original speech retrieval and encrypted speech retrieval. The original speech retrieval technology includes Philips method [2], perceptual hashing method [5, 6], ranking method [7], fingerprint method [8], emotional classification method [9] and deep learning method [10, 11], etc. Encrypted speech retrieval technology includes perceptual hashing method [12], ranking method [13] and fuzzy retrieval [14], etc. However, speech feature extraction is an important step in content-based speech retrieval. Speech feature extraction methods are mainly based on time domain [12], frequency domain [15], time-frequency domain [16], wavelet domain [17], cepstral domain [18] and image domain [19], etc. The original speech retrieval technology has been well developed, but most of them have poor security, and the speech file stored in the cloud can easily be stolen or tamper [8, 10]. Therefore, content-based encrypted speech retrieval technology emerged, it has efficiently improves the security of speech data storage and transmission.

There have been a lot of researches on the content-based encrypted speech retrieval. In 2013, Wang *et al.* [12] proposed a retrieval method of encrypted speech based on perceptual hashing, which use the Chua's chaotic circuit system and the PWL Memristor to encrypt speech, the speech zero-crossing rate was used to extract speech features and retrieve. It has good robustness and the retrieval speed is fast, but it has poor discrimination and high complexity of encryption algorithm. Ibrahim *et al.* [13] proposed a multi-keyword rank-order search method for securely encrypted cloud data. It is excellent in safety and retrieval efficiency, but the retrieval accuracy needs to be strengthened. Wang *et al.* [20] extracted perceptual hashing feature based on time and frequency domain change characteristics, which divided speech into the time domain and frequency domain to extract the perceptual hashing digest. It has better discrimination and robustness as well as high security, but the speed of extracting speech features is low. Zhao *et al.* [21] proposed an encrypted perceptual hashing retrieval method based on multifractal characteristic, which has good robustness and high retrieval accuracy, but it has poor discrimination and the retrieval efficiency is low. He *et al.* [4] proposed a retrieval method of encrypted speech based on syllable-level perceptual hashing, which has better robustness and high security. However, the discrimination of speech hash feature is poor, and retrieval efficiency needs to be improved. Glackin *et al.* [22] present a cloud-based encrypted speech retrieval method, this method has good security and retrieval accuracy. But the retrieval efficiency is not high with the cause of utilizing a high complexity encryption algorithm. It can be seen from the above references that due to the redundancy of speech signals, the speech feature of content-based encrypted speech retrieval method needs good robustness and discrimination. In other words, speech with the same perceptual content can only be mapped to the same hash digest. Moreover, the robustness and discrimination are mutually restricted, and most of the existing methods cannot balance them well, which results in low retrieval accuracy. Speech encryption algorithm will often lose part of the speech features, and the complexity of encryption algorithm will seriously affect the retrieval efficiency. The existing encryption methods may change the speech feature, which affects retrieval performance seriously. And high security and low complexity are two important aspects of encryption method need to reach. In terms of retrieval, the existing methods cannot achieve retrieve results efficiently and accurately.

To solve the above problems, in order to achieve a balance between robustness and discrimination of speech feature, reduce the complexity of encryption algorithm, then improve the accuracy and efficiency of speech retrieval. In this paper, we present an efficient retrieval method of encrypted speech based on frequency band variance. The proposed method uses Logistic chaotic scrambling encryption method to encrypt the speech file, the frequency band variance is used to extract perceptual hashing feature of speech file, which generate the encrypted speech library, and Hamming distance algorithm is used to perform highly efficient matching and retrieval.

The rest of this paper is organized as follows: Section 2 describes the related theories. Section 3 introduces the details of the proposed method. Section 4 gives the experimental results and performance analysis as compared with other related methods. Finally, we conclude our paper in Section 5.

2. **Related Theory.**

2.1. **Frequency Band Variance.** The frequency band variance [23] is actually the variance between the energy of each band of the signal, and the frequency band variance function is represented as follows. The time domain waveform of the original speech signal is $x(n)$, after pre-processing, framing and adding window to obtain the i-th frame speech signal $x_i(m)$. The function is represented as follows:

$$x_i(m) = \omega(m) \times x(iT + m), 1 \leq m \leq N \tag{1}$$

where $\omega(m)$ is the window function, $i = 0, 1, 2, ..., N$ is frame length, T is frame shift length.

The frequency of $x_i(m)$ is obtained through discrete Fourier transform (DFT). The function is represented as follows:

$$X_i(k) = \sum_{m=0}^{N-1} x_i(m) exp(-j\frac{2\pi km}{N}), 0 \leq k \leq N - 1 \tag{2}$$

Let $x_i = \{x_i(1), x_i(2), ..., x_i(N)\}$, and the average value of amplitude E_i is represented as follows:

$$E_i = \frac{1}{N} \sum_{k=0}^{N-1} |X_i(k)| \tag{3}$$

The variance D_i is represented as follows:

$$D_i = \frac{1}{N-1} \sum_{k=0}^{N-1} \left[\left| X_i(k) \right| - E_i \right]^2 \tag{4}$$

where the subscript in E_i and D_i represents the mean and frequency band variance of the i-th frame of speech signal.

In summary, frequency band variance can reflect two information of the speech signal. On one hand, it reflects the degree of fluctuation between the frequency bands of each speech signal. And on the other hand, it reflects the short-time energy of each speech frame, when the short-time energy is more intense and more fluctuating, the value of D_i is greater.

2.2. **Logistic Chaotic Scrambling.** Chaotic sequences [24] are often used for speech scrambling encryption. Logistic chaotic maps are classical model of studying the behavior of dynamic systems, chaos, fractal and other complex systems. Logistic map is also called Logistic iteration, which is a time discrete dynamical system. That is repeated iteration according to Eq. (5):

$$x_{n+1} = \mu x_n (1 - x_n) \tag{5}$$

where $0 < \mu \leq 4$, the value $x_n \in (0, 1)$. When $3.56994 < \mu \leq 4$, the logistic chaotic map is in a chaotic state. The generated sequence $\boldsymbol{X} = \{x_1, x_2, x_3, ..., x_n\}$ is aperiodic, non-convergent, and sensitive to initial values.

In this paper, the proposed method is using the characteristics of Logistic chaotic maps. Through one-to-one mapping relationship between the chaotic sequence and the sequences arranged in ascending order, speech sample points were being encrypted, and the decryption is the reverse process of encryption.

2.3. **Similarity Measure Function.** Similarity measure, a kind of measure that comprehensively assesses the similarity between two things. The closer two things are, the greater their similarity measure will become. Instead, two things have smaller similarity measure will be more alienated. There are variety of methods for similarity measure, such as the normalized Hamming distance [12] and the Euclidean distance [25]. The normalized Hamming distance $\boldsymbol{D}(:,:)$ is used in the proposed method, which is also known as the bit error ratio (BER) to match the speech. The function is represented as follows:

$$BER = D(\boldsymbol{h}_1, \boldsymbol{h}_2) = \frac{1}{m}\sum_{j=1}^{m}\Big|h_1(j) - h_2(j)\Big| \tag{6}$$

where $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are the perceptual hashing features corresponding to the two speech clips, $\boldsymbol{D}$ is the normalized Hamming distance between $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, which represents the ratio of the number of perceptual hashing error bits to the total number of bits.

Comparing the normalized Hamming distance $\boldsymbol{D}$ with the set similarity threshold τ to determine the similarity between speech clips. If $\boldsymbol{D} \leq \tau$, the perceptual content of two speech clips is the same, that they are decided to the same speech. On the contrary, they are decided to different speech.

3. The Proposed Method.

The flow chart of an efficient retrieval method of encrypted speech based on frequency band variance is shown in Fig. 1. The speech signal of the original speech file is pre-processed with framing and adding window, then the frequency band variance of the speech is calculated to construct the hash sequence, which needs to store in hash feature library in the cloud. Meanwhile, the original speech file is encrypted by Logistic chaotic scrambling and stored in the cloud encrypted speech library. The content in the hash feature library establishes one-to-one mapping relationship with the information in encrypted speech library. Then match the speech hash feature extracted as the same method before with the feature in encrypted speech library for retrieval. Finally, the speech file in encrypted speech library corresponding to the successfully matched feature is decrypted and returned as a query result to user.

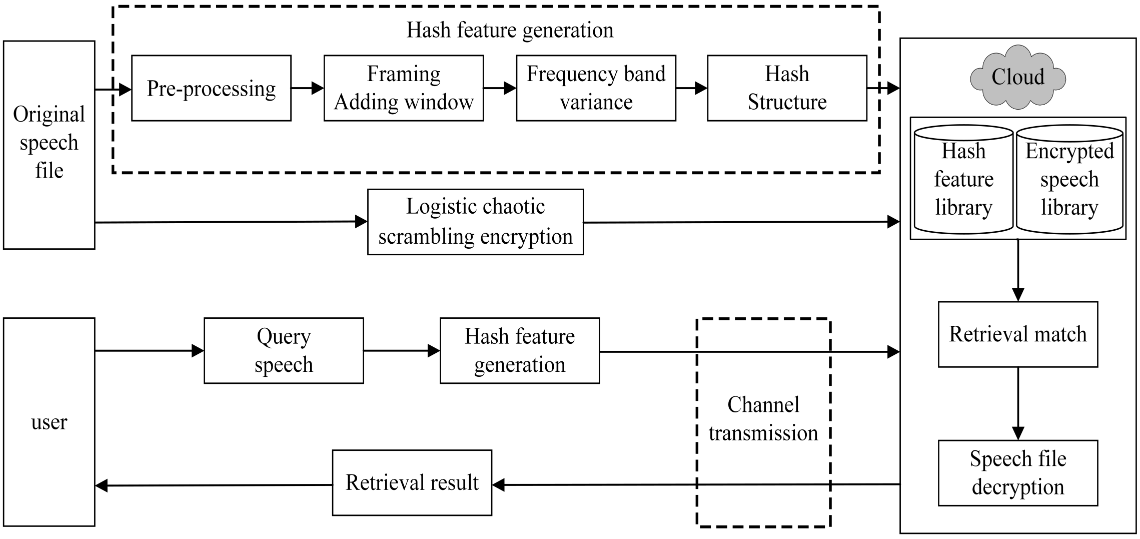


FIGURE 1. Flow chart of the encrypted speech retrieval method.

3.1. Generation of Encrypted Speech Library.

3.1.1. *Processing of speech file encryption.* Considering the security of speech in the transmission process, the original speech file is encrypted by using Logistic chaotic scrambling technology. The speech encryption processing steps are as follows:

Step 1: Original chaotic sequence generation. Select encryption key $[\mu, x_0]$, an original chaotic sequence is generated with Logistic chaotic scrambling according to Eq. (5), which is denoted as $\boldsymbol{X} = \{x_1, x_2, x_3, ..., x_n\}$, $n = 1, 2, 3, ..., M$.

Step 2: Chaotic scrambling sequence generation. The original chaotic sequence $\boldsymbol{X} = \{x_1, x_2, x_3, ..., x_n\}$ obtained in Step 1 is arranged in ascending order to obtain a new sequence $\boldsymbol{K} = \{k_1, k_2, k_3, ..., k_j\}$, $j = 1, 2, 3, ..., M$. Suppose the original speech signal is X and the encrypted speech signal is Y, if the position between original chaotic sequence position and the new sequence satisfies the mapping relationship: $k_j = x_i$, then $Y(j) = X(i)$.

Step 3: Replace the sample points of the original speech signal according to the mapping relationship in Step 2 to obtain the encrypted speech file.

Through the steps of encryption above, the encrypted speech file is uploaded to the cloud encrypted speech library.

3.1.2. *Speech perceptual hashing and hash feature library construction.* Frequency band variance reflects the spectrum distribution of the speech signal. The efficient speech feature extraction method is proposed through frequency band variance.

Step 1: Pre-processing. The speech clip $s(t)$ need a process of pre-emphasis to get the signal $s(t)'$, which makes spectrum of the signal more flat and facilitates subsequent feature extraction.

Step 2: Framing and adding window. Divide the speech clip $s(t)'$ into m frames of non-overlapping frames, which are denoted as $f_i = \{f_i(n) | n = 1, 2, ..., L/m, i = 1, 2, ..., m\}$. Where L is the length of speech clip, m is the total number of frames, $f_i(n)$ is the n-th sample value of the i-th frame. And then these m frames are processed with adding Hamming window.

Step 3: The vector $\boldsymbol{H} = \{D_i | i = 1, 2, ..., M\}$ of frequency band variance of speech signal $f_i(n)$ is calculated according to Eq. (1)-Eq. (4).

Step 4: Hash generation. The vector $\boldsymbol{H}$ is used for hash generation, which generates a hash sequence $\boldsymbol{h} = \{h(i) | i = 1, 2, ..., M\}$, The generation function is represented as follows:

$$\boldsymbol{h}(i) = \begin{cases} 1, & \text{if } \boldsymbol{H}\text{(i)>}\boldsymbol{H}\text{(i-1)} \\ 0, & \text{else} \end{cases} \tag{7}$$

where $i = 1, 2, ..., m$. Through the above steps to complete the feature extraction of speech file, and the generated speech perceptual hashing sequences were stored into the cloud hash feature library. The speech feature in the hash feature library establishes one-to-one mapping relationship with the information in encrypted speech library.

3.2. **Speech Retrieval and Decryption.** The speech retrieval method is based on the binarized hash sequence generated in Section 3.1. Then, the normalized Hamming distance matching in Section 2.3 is used to retrieve by setting a matching threshold. The speech decryption method is the inverse of the speech encryption method.

Step 1: Query speech hash generation. The query speech is treated in the same way as Section 3.1 to generate perceptual hashing features.

Step 2: Matching and retrieval. The normalized Hamming distance mentioned in Eq. (6) is used for matching the perceptual hashing generated in Step 1 with the feature in hash feature library. If the normalized Hamming distance $\boldsymbol{D}$ is less than the set matching threshold τ, it is determined that the hashing feature corresponds to the speech in the encrypted library is the retrieval speech.

Step 3: Speech file decryption. Firstly, the encryption key used in speech encryption process is selected as the decryption key. After that we generate a chaotic sequence the same as Step 1 through Eq. (5), the length of the sequence and the number of sample points do not change, and arrange them from small to large. Then the encrypted speech is inversely replaced according to the position relationship of the elements in the chaotic sequence. Suppose the encrypted speech is Y, the decrypted speech is Z, $Z(i) = Y(j)$ can be set up according to the position relationship, thus obtain the correctly encrypted speech. Finally, the decrypted speech file is returned to the user as a retrieval result.

4. **Experimental Results and Analysis.** In the experiment, we used speech library Texas Instruments and Massachusetts Institute of Technology (TIMIT) and Text to Speech (TTS) as the test speech. The library is composed of 1,280 speech clips that 16 bits signed, 16 kHz, mono, and 4 s long. Experimental platform: Inter Coer i5 @2.50GHz, 2G. Windows 7 SP1, MATLAB R2013a. The main parameters of the experiment are set as follows: non-overlapping framing, length of speech clip $L = 32,000$, number of frame $m = 360$.

4.1. **Performance Analysis of Perceptual Hashing.** The proposed method uses a binary sequence to represent a perceptual hashing digest, which is simple and the data is small. The normalized Hamming distance, also called the bit error ratio (BER), is usually used to calculate its mathematical distance. BER can be calculated to determine whether two perceptual hashing digests represent the same speech. If the BER of two hash digests is less than the pre-set threshold, they are judged having the same speech content. Otherwise, they are judged to represent different speech content. Robustness and discrimination are two main evaluation criteria for perceptual hashing performance. In order to give a clearer description of the perceptual hashing performance, this paper introduces the false acceptance rate (FAR), which refers to the error ratio of different speech contents being judged as the same content.

4.1.1. *Discrimination analysis.* The BER of perceptual hashing values from different speech information mostly follow the normal distribution, and comparing the hash sequences of 1,280 speech data clips can obtain 819,840 BER data. The normal distribution of the obtained BER is shown in Fig. 2.

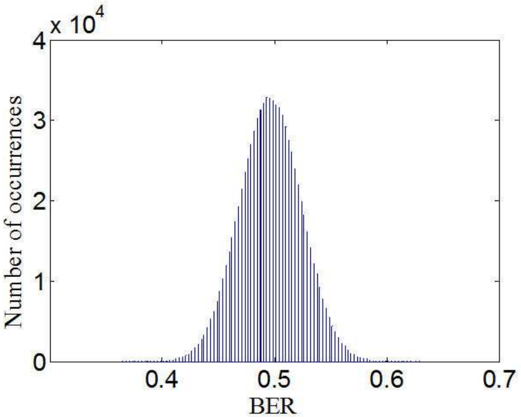


Figure 2. Statistic histogram of 1,280 speech clips matching result.

As shown in Fig. 2, the normalized Hamming distance distributes in the range of 0.3649 to 0.6295, and the result can be approximately fitted as the Gaussian distribution $N(\mu, \sigma)$ with the mathematical expectation $\mu = 0.4963$ and standard deviation $\sigma = 0.0277$, the minimum value is 0.3649. Therefore, the proposed method of this paper achieves better discrimination.

In order to further measure the discrimination of methods under different thresholds, the *FAR* is defined as shown in Eq. (8):

$$FAR(\tau) = \int_{-\infty}^{\tau} f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \tag{8}$$

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

The comparison between the proposed method and the method in Ref. [12, 20] under different thresholds is shown in Table 1.

Table 1. Comparison of *FAR* values

τ	Proposed method	Ref. [12]	Ref. [20]
0.02	1.4486×10^{-66}	1.8849×10^{-29}	3.7864×10^{-70}
0.06	3.3854×10^{-56}	7.5793×10^{-25}	3.2078×10^{-59}
0.10	9.9134×10^{-47}	1.1957×10^{-20}	3.0514×10^{-49}
0.14	3.6436×10^{-38}	7.2567×10^{-17}	3.2653×10^{-40}

As can be seen from Table 1, when $\tau = 0.14$, the *FAR* value of the proposed method is very small. It means that the number of error judgments per 1×10^{38} speech clips is only 3.6, which show that the proposed method has strong anti-collision capability. And when choose different threshold τ, the *FAR* of the proposed method is less than that in Ref. [12]. The performance in Ref. [20] is better than the proposed method with the reason that it chooses a high complexity algorithm for extracting perceptual hashing feature, so it has good discrimination. Therefore, there is a gap between the proposed method and Ref. [20], but the basic requirements of the proposed method have been achieved. Therefore, we can conclude that the proposed method achieves better discrimination.

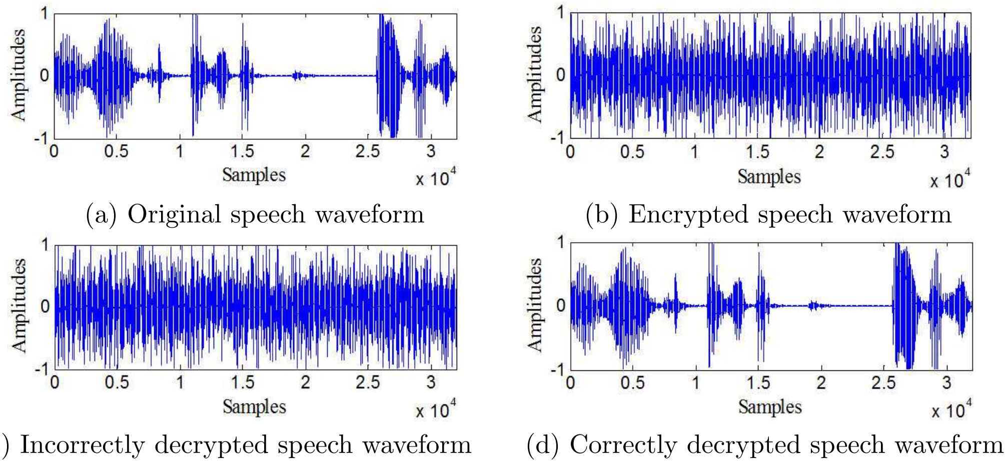
4.1.2. *Robustness analysis.* In order to test the robustness of the proposed method, the content-preserving operation (CPO) shown in Table 2 is firstly applied to the speech file in speech library. The number of 1,280 speech clips in speech library generates the number of 10,240 speech files after the CPO. Then, the *BER* values between original speech library and various CPO are respectively calculated and it is shown in Table 2.

Table 2. The *BER* mean and maximum value of the speech CPO

CPO/Operating means	Proposed method		Ref. [12]		Ref. [20]	
	Mean	Max	Mean	Max	Mean	Max
Volume up /+50%	0.0925	0.2758	0.0203	0.2670	0.0563	0.1236
Volume down /-50%	0.0089	0.1003	4.9668×10^{-4}	0.0630	0.0458	0.0847
Re-quantization /8-16kbps	0.0442	0.3064	0.1354	0.8463	0.0692	0.2458
Re-quantization /16-32kbps	8.7184×10^{-6}	0.0028	0	0	1.0868×10^{-6}	0.0014
Noise addition /50dB	0.0416	0.3036	0.0833	0.3829	0.0663	0.2653
Echo addition /50%	0.2375	0.3148	0.1450	0.3678	0.1242	0.2000
Re-sampling /8-16kbps	0.0304	0.1476	0.0065	0.0982	0.0027	0.0222
Noise reduction /75%	0.1201	0.2563	0.2078	0.8589	0.2069	0.3000

It can be seen from Table 2 that after perceptual robustness test, the maximum *BER* is 0.3148. Under the operation of re-quantization, the *BER* mean of the proposed method is 0.0442, far less than 0.1354 in Ref. [12] and 0.0692 in Ref. [20]. Meanwhile, under the operation of 50 dB noise addition, the *BER* mean of the proposed method is 0.0416, far less than 0.0833 in Ref. [12] and 0.0663 in Ref. [20]. So the proposed method has nice robustness.

4.2. **Performance of Speech Encryption and Decryption.** Logistic mapping based the chaotic scrambling encryption and decryption technology are used to encrypt and decrypt the speech signal during transmission. The encryption and decryption method use the discrete sequences generated by the Logistic chaotic map to sort, and scrambles the original speech sample points according to the arrangement manner. The proposed method can disrupt the correlation of the original speech signals, and then obtain good security. Moreover, the encryption and decryption speed of the proposed method is very fast, which is suitable for the safe transmission process of speech signals. The speech encryption and decryption amplitudes comparisons are shown in Fig. 3.



(a) Original speech waveform

(b) Encrypted speech waveform

(c) Incorrectly decrypted speech waveform

(d) Correctly decrypted speech waveform

Figure 3. Speech encryption and decryption comparisons.

As can be seen from Fig. 3, the speech signal scrambled by Logistic chaos is very vague with respect to the original speech signal, and no feature of the original speech signal can be seen from the encrypted speech signal at all. Therefore, the encrypted speech signal hides the content of the original speech signal very well and ensures the security of the speech.

In order to get the correct decryption signal, the key $[\mu, x_0]$ identical with the encryption is used as the decryption key. The slight changes in x_0 and μ will cause changes in the entire chaotic sequence. Therefore, this encryption method has strong key sensitivity. If the key get slightly change with the decryption key, the error decryption waveform shown in Fig. 3(c) is obtained, and the waveform is still very confusing. The original speech signal can only obtain after using an accurate decryption key.

4.3. **Performance Analysis of Retrieval Method.** The retrieval performance of the proposed method is usually evaluated by the recall ratio (R) and the precision ratio (P). The calculation for R and P are shown in Eq. (9) and Eq. (10).

$$R = \frac{f_T}{f_T + f_L} \times 100\% \tag{9}$$

$$P = \frac{f_T}{f_T + f_F} \times 100\% \tag{10}$$

where f_T is the number of speech clips related to the keyword in retrieval results, f_F is the number that not related to the keyword, and f_L is the number of speech clips related to the keyword but not retrieved.

When the perceptual hashing sequence of the query speech matches the features in hash feature library, a similarity threshold T_2 will be set, $0 < T_2 < 0.5$, and if the normalized Hamming distance $\boldsymbol{D}(\boldsymbol{h}_1, \boldsymbol{h}_2) < T_2$, the match is successful. Therefore, it is crucial to choose the appropriate similarity threshold for the recall ratio and precision ratio of speech retrieval. The discrimination experimental results show that the BER of the perceptual hashing sequence generated by the proposed method for 1,280 speech clips is 0.3649 as a minimum, and the maximum BER in the robustness experiment results is 0.3148, through which the appropriate similarity threshold T_2 can be determined in the range of $0.3148 < T_2 < 0.3649$. And in this range, high recall ratio and precision ratio of retrieving results can be guaranteed. For example, the 600-th speech clip is chosen as the target query speech. A hash sequence is generated for this speech and then retrieve it in the hash feature library. The matching result is shown in Fig. 4.

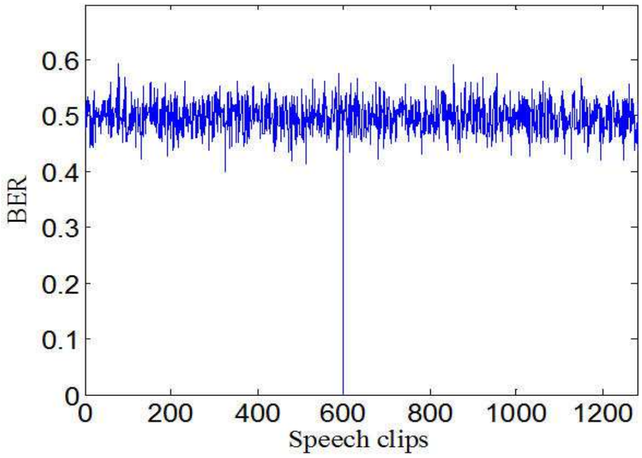


FIGURE 4. Matching result of query digest in hash feature library.

As can be seen from Fig. 4, only when the *BER* is very small, can the matching perceptual hashing retrieve the corresponding speech accurately, and the remaining 1,279 matches failed. The proposed method still guarantees good accuracy in retrieval when the speech modified by CPO, and it is shown in Table 3.

TABLE 3. Comparison of recall ratio after CPO

Operating means	Proposed method	Ref. [12]	Ref. [20]
Volume up/+50%	100%	100%	100%
Volume down/-50%	100%	100%	100%
Re-quantization/8-16kbps	100%	95%	100%
Re-quantization/16-32kbps	100%	100%	100%
Noise addition/50dB	100%	100%	100%
Echo addition/50%	100%	100%	100%
Re-sampling/8-16kbps	100%	99 %	100%
Noise reduction/75%	100%	100%	98%
Invert	100%	100%	100%

As can be seen from Table 3, the retrieval recall ratio of the proposed method after a variety of CPO is stable at 100%, while in the Ref. [12, 20], partial speech cannot be distinguished due to high FAR and poor robustness, eventually lead to some errors in the retrieval results.

The retrieval efficiency is extremely important for speech retrieval. In order to test the complexity and computational efficiency of the proposed method, 1,000 speeches are randomly selected in the TIMIT speech library for testing. Counting running time of retrieval process, which include two parts of feature extraction and feature matching (the matching method is the normalized Hamming distance) and then compare it with Ref. [12, 20], the comparison results is shown in Table 4.

Table 4. Comparison of efficiency of each method

Method	Frequency(GHz)	Speech length(s)	Average running time
Ref. [12]	1.60	4	0.1304
Ref. [20]	2.50	4	0.5218
Proposed method	2.50	4	0.1467

As can be seen from Table 4, the retrieval efficiency of the proposed method is much higher than that of the Ref. [20]. Due to the mutual restriction among the discrimination and robustness of speech perceptual hashing features as well as retrieval efficiency in the process of encrypted speech retrieval. The Ref. [12] chose a simpler feature extraction method which improves the retrieval efficiency at the expense of discrimination. Therefore, the retrieval efficiency of Ref. [12] is faster than the proposed method.

4.4. **The Retrieval Efficiency Comparison of Common Similarity Measure.** There are many kinds of similarity measures, which are effective methods for encrypted speech retrieval. In order to compare the retrieval efficiency of 8 kinds of common similarity measures, the statistical result is calculated by comparing the retrieval efficiency of common similarity measures in different number of speech libraries (50, 200, 500, 800, and 1,280). The retrieval efficiency comparison is shown in Fig. 5.

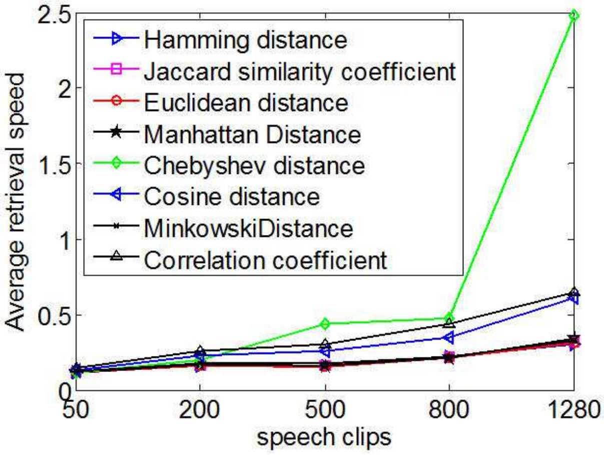


Figure 5. Retrieval efficiency comparison of different method.

As can be seen from Fig. 5, the retrieval speed of common similarity measures will gradually decrease as the increase number of speech clips in the speech library. When the number of speech clips in speech library is less than 200, the retrieval efficiency of the 8 common similarity measure functions are similar, and with the increase number of speech clips, the retrieval efficiency has gradually become different. When the number of speech clips gradually increased to 1,280, the retrieval efficiency was not significantly reduced in the Hamming distance, the Euclidean distance, the Jaccard similarity coefficient, the

Minkowski distance, and the Manhattan distance. The efficiency of the Cosine distance and Correlation coefficient have decreased significantly, and the retrieval efficiency of the Chebyshev distance has decreased most significantly. Therefore, under the experimental environment of this paper, the retrieval efficiency of the Hamming distance and the Euclidean distance are best, which are more suitable for massive content-based encrypted speech retrieval than other six similarity measures.

5. **Conclusions and Future Work.** At present, most content-based encrypted retrieval methods generally have the problems of high complexity of the encryption algorithm, low retrieval efficiency and accuracy, as well as poor discrimination and robustness. In order to solve these problems, this paper proposes an efficient encrypted retrieval method based on frequency band variance. The proposed method uses the speech frequency band variance as the speech feature to construct the hash sequence. At the same time, the Logistic chaotic scrambling is applied to encrypt the original speech file and the normalized Hamming distance algorithm is used for matching and retrieving. The advantages of the proposed method are: 1) The speech frequency band variance is used as a feature to generate perceptual hashing sequences. In the proposed method the extracted hash sequences have good robustness and discrimination, and the efficiency of feature extraction is fast. 2) It can improve security on the basis of efficiency when adopting Logistic chaotic scrambling technology to encrypt and decrypt speech file. 3) Matching hash feature through the method of normalized Hamming distance, the proposed method holds good retrieval efficiency in comparing with common similarity measures when the number of speech clips gradually increases. Moreover, the recall ratio of the proposed method is high. 4) Efficient retrieval can be performed on the encrypted speech data without downloading and decrypting speeches.

In future work, we plan to test long speech clips in the experiment and the protection performance of privacy security need to be further improved.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No. 61862041, 61363078), the Research Project in Universities of Education Department of Gansu Province (2017B-16, 2018A-187). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

[1] Y. V. Murthy and S. G. Koolagudi. Content-Based Music Information Retrieval (CB-MIR) and Its Applications toward the Music Industry: A Review. *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, Article. 45, 2018.

[2] X. Z. Zhang, Y. S. Wang and Z. Zeng, An efficient filtering-and-refining retrieval method for big audio data, *Journal of Computer Research and Development* (in Chinese), vol. 52, no. 9, pp. 2025–2032, 2015.

[3] M. Thangavel, P. Varalakshmi and S. Renganayaki, SMCSRC-Secure multimedia content storage and retrieval in cloud, *Proc. of the 2016 International Conference on Recent Trends in Information Technology (ICRTIT)*, IEEE, Chennai, India, pp. 1–6, 2016.

[4] S. F. 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.

[5] P. Panyapanuwat, S. Kamonsantiroj and L. Pipanmaekaporn. Time-frequency ratio hashing for content-based audio retrieval, *Proc. of the 2017 9th International Conference on Knowledge and Smart Technology (KST)*, IEEE, Chonburi, Thailand, pp. 205–210, 2017.

[6] S. S. Jin, A resilience mask for robust audio hashing, *IEICE TRANSACTIONS on Information and Systems*, vol. 100, no. 1, pp. 57–60, 2017.

[7] J. Qin, X. Liu and H. Lin, Audio retrieval based on manifold ranking and relevance feedback, *Tsinghua Science and Technology*, vol. 20, no. 6, pp. 613–619, 2015.

[8] G. H. Ning, Z. Zhang, X. B. Ren, H. H. Wang and Z. H. He, Rate-coverage analysis and optimization for joint audio-video multimedia retrieval, *Proc. of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, New Orleans, LA, USA, pp. 2911–2915, 2017.

[9] B. Zhang and J. Lin. An Efficient Content Based Music Retrieval Algorithm, *Proc. of the 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, IEEE, Xiamen, China, pp. 617–620, 2018.

[10] M. Dorfer, A. Arzt and G. Widmer. Towards End-to-End Audio-Sheet-Music Retrieval, *Proc. of the NIPS 2016 Workshop on End to End Speech and Audio*, Barcelona, Spain, pp. 1–5, 2016.

[11] H. Y. Lee, P. H. Chung and Y. C. Wu, Interactive Spoken Content Retrieval by Deep Reinforcement Learning, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*(Early Access), pp. 1–13, 2018.

[12] H. X. Wang, L. N. Zhou, W. Zhang and S. Liu, Watermarking-based perceptual hashing search over encrypted speech, *Proc. of the International Workshop on Digital Watermarking*, Springer, Berlin, Heidelberg, pp. 423–434, 2013.

[13] A. Ibrahim, H. Jin and A. A. Yassin, Secure Rank-Ordered Search of Multi-keyword Trapdoor over Encrypted Cloud Data, *Proc. of the 2012 IEEE Asia-Pacific Services Computing Conference (APSCC)*, IEEE, Guilin, China, pp. 263–270, 2012.

[14] J. H. Su, C. Y. Wang and T. W. Chiu, Semantic content-based music retrieval using audio and fuzzy-music-sense features, *Proc. of the 2014 IEEE International Conference on Granular Computing (GrC)*, IEEE, Noboribetsu, Japan, pp. 259–264, 2014.

[15] A. Suksukont and J. Srinonchat. Improving the quality of the speech signal using a FIR band pass filter with Fast Fourier transform, *Proc. of the 2017 International Electrical Engineering Congress (iEECON)*, IEEE, Pattaya, Thailand, pp. 1–4, 2017.

[16] G. Y. Hao, Research of speech perceptual hashing and its application in search over encrypted speech, Master. Thesis, Southwest Jiaotong University, China, 2015.

[17] Q. Y. Zhang, S. B. Qiao, Y. B. Huang and T. Zhang, A high-performance speech perceptual hashing authentication algorithm based on discrete wavelet transform and measurement matrix, *Multimedia Tools and Applications*, vol. 77, no. 16, pp. 21653–21669, 2018.

[18] V. Panagiotou and N. Mitianoudis, PCA summarization for audio song identification using Gaussian mixture models, *Proc. of the 2013 18th International Conference on Digital Signal Processing (DSP)*, IEEE, Fira, Greece, pp. 1–6, 2013.

[19] D. Williams, A. Pooransingh and J. Saitoo, Efficient music identification using ORB descriptors of the spectrogram image, *EURASIP Journal on Audio, Speech, and Music Processing*, vol.2017, no.1, pp. 1–17, 2017.

[20] H. X. Wang and G. Y. Hao, Perceptual Speech Hashing Algorithm Based on Time and Frequency Domain Change Characteristics, *China Patent* No. 2015102405844, 12 Aug. 2015.

[21] H. Zhao and S. F. He, A retrieval algorithm for encrypted speech based on perceptual hashing, *Proc. of the 2016 12th International Conference on Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, IEEE, Changsha, China, pp. 1840–1845, 2016.

[22] C. Glackin, G. Chollet and N. Dugan, Privacy preserving encrypted phonetic search of speech data, *Proc. of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, New Orleans, LA, USA, pp. 6414–6418, 2017.

[23] E. P. B. Reynders and R. S. Langley. Cross-frequency and band-averaged response variance prediction in the hybrid deterministic-statistical energy analysis method. *Journal of Sound and Vibration*, vol. 428, pp. 119–146, 2018.

[24] M. F. A. Elzaher, M. Shalaby and S. H. El. Ramly, An Arnold Cat Map-Based Chaotic Approach for Securing Voice Communication, *Proc. of the 10th International Conference on Informatics and Systems.* ACM, Giza, Egypt, pp. 329–331, 2016.

[25] J. Zhao, Content-based intelligent retrieval and repetitive detection of massive audio, Master. Thesis, Taiyuan University of Technology, China, 2015.