Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 2467 – 2482
© TÜBİTAK
doi:10.3906/elk-1907-94

An encrypted speech authentication method based on uniform subband spectrum variance and perceptual hashing

Qiuyu ZHANG[*], **Denghai ZHANG**, **Liang ZHOU**
School of Computers and Communication, Lanzhou University of Technology, Lanzhou, P.R. China

Received: 13.07.2019 • **Accepted/Published Online:** 02.01.2020 • **Final Version:** 25.09.2020

Abstract: In a real-world cloud server, a speech signal is prone to suffer various attacks, such as malicious muting and tampering. In such a context, the privacy security of the speech owner will not be guaranteed. In order to achieve content authentication of encrypted speech in the cloud server, an efficient encrypted speech authentication method based on uniform subband spectrum variance and perceptual hashing is proposed. Firstly, the original speech is scrambled by Henon mapping to construct an encrypted speech library in the cloud, through extracting uniform subband spectrum variance of the encrypted speech and constructing a hashing sequence to generate a hashing template of the cloud. In this way, a one-to-one correspondence between the encrypted speech and the hashing sequence is built. Secondly, the authentication digest of encrypted speech is extracted according to the inquiry result. Finally, the authentication digest and the hashing sequence in the cloud are matched by the Hamming distance algorithm. The experimental results demonstrate that the proposed method has great security and efficiency, and it can directly extract the authentication digest from encrypted speech. The authentication digest not only has good discrimination and robustness, but it accurately locates the tampered area for malicious substitution and mute attacks.

Key words: Encrypted speech authentication, perceptual hashing, uniform subband spectrum variance, feature extraction of encrypted speech, tamper location

1. Introduction

With the rapid development of cloud technology and speech signal processing technology, the number of digital speeches has increased dramatically. Cloud technology makes it easy and convenient for people to access multimedia information. However, an open and multitenancy cloud storage supplier is not a completely trusted third-party organization, and it brings convenience but significant security and privacy risks [1]. For instance, when multimedia information is transmitted or stored in the cloud, it may be attacked or tampered with by hackers, making speech lose its original meaning [2]. In fact, the biggest disadvantage of cloud storage is that the confidentiality and integrity of speech is hard to guarantee [3].

As an important aspect of multimedia information application scenarios, many speech recordings contain some sensitive information, such as trade secrets, court evidence, and other important information [4]. For instance, some speech recordings as legal evidence are stored in the cloud. If these evidence files are attacked or tampered with by hackers, there will be unfair and unpredictable legal consequences. Therefore, it is necessary to improve the security factor and privacy of speech during storage and transmission in the cloud. Meanwhile, these speech recordings, after being transmitted through the cloud, need to undergo content authentication and

[*] Correspondence: zhangqylz@163.com

are judged in terms of whether they were tampered with by hackers. For the speech query user, verifying the confidentiality and integrity of a speech is also an urgent problem to be solved. Hence, it is important to study privacy protection for speech and content authentication of encrypted speech.

AES encryption [5] and chaotic encryption [6] are used for speech privacy protection. These methods usually use a chaotic system to generate random sequences to perform scrambling and replacement operations, making the speech waveform whiten. Furthermore, they have perfect encryption performance and high security. However, such these encryption methods lose many perceptual features while exchanging good encryption performance, making it difficult to directly extract perceptual features from encrypted speech.

The technologies used in current speech authentication research fall into two general categories: digital watermarking and perceptual hashing. Wang et al. [7] proposed a speech tampering detection scheme for based on digital watermarking, which used formant enhancement to improve watermark robustness. However, an authentication error would occur if the attackers used desynchronized attacks. In [8], a speech content authentication algorithm based on digital watermarking was proposed. By scrambling the segmented samples and calculating the scrambled autocorrelation, the quantized correlation was used to embed the watermark bits. The scheme had a strong ability to resist desynchronized attacks. Liu et al. [9] proposed a new speech content authentication algorithm based on Bessel–-Fourier moments, thus defending against the desynchronized attacks of insertion and deletion. Liu also proposed a speech authentication algorithm based on digital watermarking [10]. The cross-correlation coefficients and frame numbers of speech signals were embedded into the speech signal as digital watermarks; this algorithm had wonderful robustness and it could locate the tampered position. Qian et al. [4] proposed a dual watermark scheme for speech authentication. It transformed speech into a matrix, and it used the matrix and the hash function to generate dual watermarks for speech authentication. The results showed that the algorithm had great security and it could locate the tampered position. Qian et al. [11] also proposed a high-efficiency encrypted speech content authentication algorithm based on the integer wave transform; it embedded the approximate coefficients of the encrypted speech as a digital watermark into the detailed coefficients. This method had good robustness and strong ability to resist desynchronized attacks.

By analyzing the above research, we can see that most speech authentication methods based on digital watermarking do not consider the problem of speech information streaking. However, embedding the watermark into the speech signal reduced the recovery quality of speech and complicated the authentication system. Therefore, the technology of perceptual hashing was proposed.

Perceptual hashing is constructed by extracting perceptual features directly from a speech. Chen et al. [12] proposed a nonnegative matrix factorization based on the linear prediction coefficient (LPC) as an authentication digest; however, the robustness and discrimination of the algorithm were bad. Li et al. [13] proposed a perceptual hashing method based on the modified discrete cosine transform (MDCT) combined with a nonnegative matrix (NMF). The algorithm extracted the MDCT coefficients of speech with the NMF dimension reduction method to construct the perceptual hashing sequence; this method had nice robustness. Zhang et al. [14] proposed a time–frequency domain perceptual hashing authentication algorithm that used the spectral features of the speech with the discrete wavelet transform to generate a ternary perceptual hashing sequence. The experimental results showed that it had good robustness and discrimination. In [15], an efficient perceptual hashing algorithm based on improved spectral entropy for speech authentication was proposed. The speech signal was framed and the minimum mean square error coefficient matrix was obtained. Then the spectral entropy parameter matrix of each frame was calculated by the improved spectral entropy method.

Experiments showed that it had better discrimination and robustness. In order to solve the problem of speech privacy protection in the speech retrieval model, Wang et al. [2] proposed an encrypted speech retrieval method. The speech was encrypted first by a chaotic system. Then the perceptual hashing extracted from the original domain as a digital watermark was embedded into the encrypted speech, which increased the complexity of the retrieval system. In [16], a speech retrieval method based on perceptual hashing was proposed; it extracted the perceptual hashing from the original domain as a digital watermark to embed it into the encrypted speech. This could ensure the privacy of the speech. He et al. [17] proposed an encrypted speech retrieval scheme based on perceptual hashing. This method extracted the perceptual hashing from the original domain as a digital watermark and embedded it into the encrypted speech, but this increased the complexity of the retrieval system. Zhang et al. [18] proposed an encrypted speech retrieval method based on short-term cross-correlation and perceptual hashing. The retrieval digest was extracted from the encrypted speech, and the method decreased the complexity of the retrieval system.

Through the above research analysis, we can see that the authentication digest or retrieval digest was extracted from the original speech. These schemes not only greatly reduce confidentiality but also increase uncertainty. In addition, they also increase the complexity of authentication and the retrieval system.

From what has been discussed above, in order to solve the confidentiality and ensure integrity of the speech in the cloud, an encrypted speech authentication method based on uniform subband spectrum variance and perceptual hashing is proposed in this paper. The main contributions of our approach can be summarized as follows:

1) A time–frequency domain grouping scrambling speech encryption method based on Henon mapping is designed in this paper. It ensures the privacy of speech stored in the cloud, and perceptual features such as the authentication digest can be directly extracted from the encrypted speech.

2) The uniform subband spectrum variance of the encrypted speech is used to construct the perceptual hashing sequence for authentication, and the perceptual hashing has great discrimination, robustness, and high authentication efficiency.

3) When the encrypted speech is tampered with by substitution and mute attacks, the tampered area can be accurately detected and located.

Compared with the existing authentication methods based on perceptual hashing, the authentication digest can be extracted directly from the encrypted speech in our method. The proposed method is simple and more efficient against authentication while ensuring the privacy of speeches transmitted in the cloud. The whole authentication process in our method is operated on the encrypted speech, which has perfect security.

The rest of this paper is organized as follows. Section 2 describes the related theories. Section 3 gives a model of the encrypted speech authentication, encrypted speech algorithm, perceptual hashing extraction, and speech authentication. Section 4 gives the experimental results, performance analysis, and results obtained from comparison with other related methods. Finally, we conclude our paper in Section 5.

2. Related theory

2.1. Henon mapping

Henon mapping is often used in image encryption [19, 20], because it generates a random sequence with perfect security. Since the proposed encryption algorithm in this paper focuses on the security of the key, we introduce the Henon chaotic system [21] into the encryption method of this paper to generate a key. The Henon mapping is defined as follows:

$$\boldsymbol{T}(i) = \begin{cases} x_{i+1} = y_i + 1 - ax_i^2 \\ y_{i+1} = bx_i \end{cases} . \tag{1}$$

It can be seen from (1) that the Henon mapping iterates the whole chaotic equation from two variables x and y, and this is more complicated than one-dimensional or multiple one-dimensional chaotic systems. For the parameters $1.54 < a < 2, 0 < |b| < 1$, the system is in a hyperchaotic state. At this time, the chaotic sequence generated by the system does not overflow when under a large number of iteration rounds, which accords with the requirements of various statistical characteristics and is suitable for the generation of an encryption key. In summary, Henon mapping is easy to implement, relatively complex, and highly secure. Therefore, Henon mapping is a suitable key generator for speech encryption.

2.2. Uniform subband spectrum variance

The existing band variance is calculated by the variance of each spectral line, which have large fluctuations and low stability. Therefore, in our method, the uniform subband separation band variance [22] is directly extracted from the encrypted speech.

First, a fast Fourier transform is performed on each frame of data N, and there are $N/2+1$ lines in the positive frequency domain. The $N/2+1$ discrete Fourier transform postamplitude spectrum $X_i = X_i(1), X_i(2), ..., X_i(N/2+1)$ is divided into q subbands (i denotes the ith frame), and each subband contains $p = fix[(N/2+1)/q]$ lines (where $fix[.]$ indicates its integer part). Then the subband $XX_i(m)$ is:

$$XX_i(m) = \sum_{k=1+(m-1)p}^{1+(m-1)p+(p-1)} |X_i(K)|, \tag{2}$$

where m represents the number of spectral lines in the positive frequency domain within a frame of speech.

Letting $XX_i = XX_i(1), XX_i(2), \cdots, XX_i(q)$, the subband mean $E_{i,1}$ is:

$$E_{i,1} = \frac{1}{q}\sum_{k=1}^{q} XX_i(k). \tag{3}$$

Subband variance $D_{i,1}$ is:

$$D_{i,1} = \frac{1}{q-1}\sum_{k=1}^{q}[XX_i(k) - E_{i,1}]^2. \tag{4}$$

The band variance can account for the undulation of the band and the involved energy, but it cannot better differentiate the noise and speech segment in the lower signal to noise ratio (SNR) speech. However, the subband variance method can distinguish the noise and the speech segment well. Consequently, the uniform subband band variance separation as the perceptual feature is extracted from the encrypted speech to construct a perceptual hashing sequence in our method.

3. The proposed authentication algorithm of encrypted speech

3.1. The authentication model of encrypted speech

Figure 1 shows an encrypted speech authentication model for encrypted speech retrieval; the model can extract the authentication digest under the premise of ensuring speech privacy.

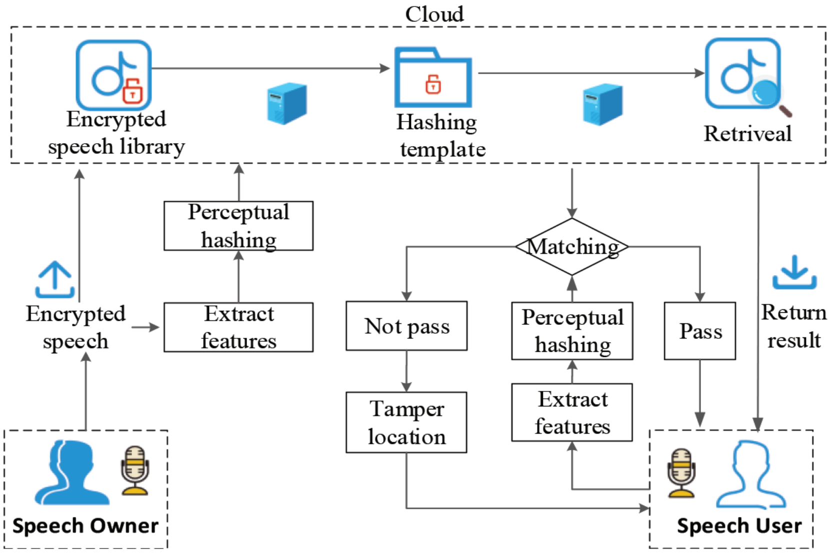


Figure 1. The authentication model of encrypted speech based on perceptual hashing.

As shown in Figure 1, the model can achieve speech content authentication while ensuring speech privacy. The model consists of the following three parts. (1) The speech owner encrypts the speech according to the key, ensuring the security of the speech signal. Then the speech holder uploads it to the cloud server to form an encrypted speech library. (2) The key holder extracts the perceptual features of the sample encrypted speech, constructing a perceptual hashing sequence and generating a hashing template. (3) The legally authorized speech user first extracts the hashing sequence of the queried speech, constructs a search digest, and obtains the query result through matching with the hashing template stored in the cloud by the Hamming distance algorithm. In order to verify the authenticity and integrity of the queried speech, perceptual features are first extracted from the queried encrypted speech; these are then matched with the hashing index in the hashing template stored in the cloud. If the matching is successful, the speech is a legal speech; otherwise, it is a speech that has been tampered with tampering attacks.

The main innovation of the proposed model is that it solves the problem of speech information streaking and enhances the privacy of speech content in the cloud. The authentication digest is constructed by directly extracting the uniform subband variance from the encrypted speech, which simplifies the authentication model. Moreover, the digest has better discrimination, abstractness, and robustness for various content-preserving operations (CPOs). Compared with the digital watermark system, the proposed model has lower system complexity. In the whole process, the model has high security and reliability because the speech always exists in an encrypted form in the cloud.

3.2. Speech encryption

In order to protect the security and privacy of speech stored in the cloud, it is necessary to encrypt the speech before uploading it to the cloud. The encryption algorithm should ensure the privacy of the speech while ensuring that the encrypted speech still has perceptual features and while constructing the hashing sequence. The scrambling algorithm based on the time–frequency domain only scrambles speech samples and does not do XOR and replacement samples operations. This not only guarantees the speech privacy, but also ensures partially perceptual features in the encrypted speech. Meanwhile, the Henon mapping in the chaotic state generates the random sequence, which has strong random and high security. Hence, Henon mapping is used to generate the key. The interframe grouping in the time domain and intraframe grouping in the frequency domain are combined to scramble each other between groups.

Figure 2 shows the processing flow of the encryption algorithm.

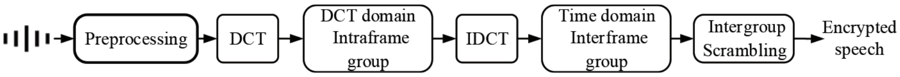


Figure 2. The processing flow of the encryption algorithm.

As can be seen from Figure 2, the encryption algorithm first performs frame preprocessing on the speech signal by using a frame length of 256 and frame shift of 256. A speech has a total of 64,000 sample points, which are divided into 250 frames, and each frame has 256 sample points. In order to have better encryption performance without losing most of the perceptual features, the authentication digest can be extracted directly from the encrypted speech. Therefore, the speech is divided into 64, 32, or 16 groups by 4, 8, and 16 sampling points in the discrete cosine transform (DCT) domain. In the time domain, the speech is divided into a group per 2 frames, in total 125 groups, or it is divided into a group per 5 frames, in total 50 groups. We select 20 speeches in the speech library for testing. These encrypted speeches are tested by the objective mean opinion score (MOS) value, which evaluates the performance of the encryption algorithm. The MOS value comparison results are shown in Table 1.

Table 1. The MOS value of the encrypted speech under different groups.

Intraframe (group)	Interframe (group)	MOS value	Interframe (group)	MOS value
64	125	0.4116	50	0.5591
32	125	0.4995	50	0.6376
16	125	0.5240	50	0.7519

As can be seen from Table 1, when the number of interframe groups is constant, the MOS value becomes lower as the number of intraframe groups increases. This leads to worse speech quality, which indicates that the encryption performance becomes better. However, the loss of perceptual features also increases. In order to balance key security, encryption performance, and perceptual feature performance, the proposed encryption method adopts four sampling points per interframe as a group in the frequency domain and five frames as a group in the time domain. This not only ensures better encryption performance, but it also guarantees key security and that general key attacks can be resisted. Finally, all of the original speeches are encrypted and uploaded to form an encrypted speech library in the cloud.

3.3. The construction of encryption speech perceptual hashing sequence

Speech and noise have different characteristics in the frequency domain, but in lower SNR speech, the uniform subband variance can better distinguish the speech segment and the noise than the band variance in the spectral domain. Therefore, the proposed method extracts the subband separation variance as a perceptual feature from the encrypted speech of lower SNR to construct the perceptual hashing sequence.

Step 1: Speech preprocessing. Let the speech signal be $x(n)$. Using (5), we can eliminate the direct current (DC) component because the existence of the DC component will affect the calculation of the uniform sub-band spectrum variance:

$$x(n) = x(n) - mean(x(n)), \tag{5}$$

where $mean(.)$ is the mean of the speech signal.

Step 2: Framing the windowing. For this, the spectrum is leaked by truncating the speech directly. Accordingly, in order to improve the occurrence of this situation, the speech signal $x(n)$ is processed by a Hamming windowing. The speech signal $x(n)$ has no overlapping framing, the frame length $wlen$ is set to 256, and the frame shift inc is set to 256. After windowing processing, $x_i(m)$ represents the ith frame speech signal.

Step 3: Uniform subband separation. According to (2), one subband is composed of $p = 4$ points, and $x_i(m)$ is divided into $q = 32$ subbands. The subband spectrum is $XX_n(k)$, where $1 \leq k \leq 32$.

Step 4: Feature extraction. Define the subband spectrum as $XX_i = XX_i(1), XX_i(2), \cdots, XX_i(q)$ of all of the frames, and the subband mean value $E_{i,1}$ of the amplitude is calculated according to (3). Then, using (4), we can calculate the variance of the band after each subband separation, thus generating the feature matrix parameter $\boldsymbol{D}_w(1, Z_m)$, where Z_m denotes the number of feature vectors.

Step 5: Hashing construction. A binary hashing construction is performed on the characteristic parameter matrix $\boldsymbol{D}_w(1, Z_m)$ to generate a hashing sequence $\boldsymbol{H}(1, 2, ..., q_m)$ of 0 and 1. The specific construction method is as follows:

$$\boldsymbol{H}_i(x) = \begin{cases} 1, & \boldsymbol{D}_w(j+1) > \boldsymbol{D}_w(j), \\ 0, & \boldsymbol{D}_w(j+1) \leq \boldsymbol{D}_w(j), \end{cases} \qquad j = 1, 2 \cdots Z_m, \tag{6}$$

where j is the jth feature vector in the feature parameter matrix $\boldsymbol{D}_w$.

3.4. Speech authentication

When the unauthenticated speech extracts the perceptual hashing sequence, which is matched with the hashing template by the normalized Hamming distance (as shown in (7)), where S_1 and S_2 represent two different speech segments, H_1 and H_2 are hashing sequences generated by two speech segments and M represents a hashing sequence length. The normalized Hamming distance $W(:,:)$ means the bit error rate (BER), and that is the unauthenticated speech perceptual hashing matching error rate with the perceptual hashing sequence of the hashing template:

$$W(S_1, S_2) = D(H_1, H_2) = \frac{1}{M} \sum_{k=1}^{M} |H_1(k) - H_2(k)|. \tag{7}$$

To better describe the entire authentication model, we make two assumptions as follows, where τ is the hashing sequence matching threshold:

A. If $W(S_1, S_2) \leq \tau$ is established, the authentication is passed.

B. If $W(S_1, S_2) > \tau$ is established, the authentication is not passed.

According to the above hypothesis, when the Hamming distance is less than or equal to the matching threshold of the authentication model, we consider that the two speeches have the same content and the authentication is passed. Conversely, if the Hamming distance is greater than the matching threshold, we consider that the two speeches do not have exactly the same content, and the authentication is not passed.

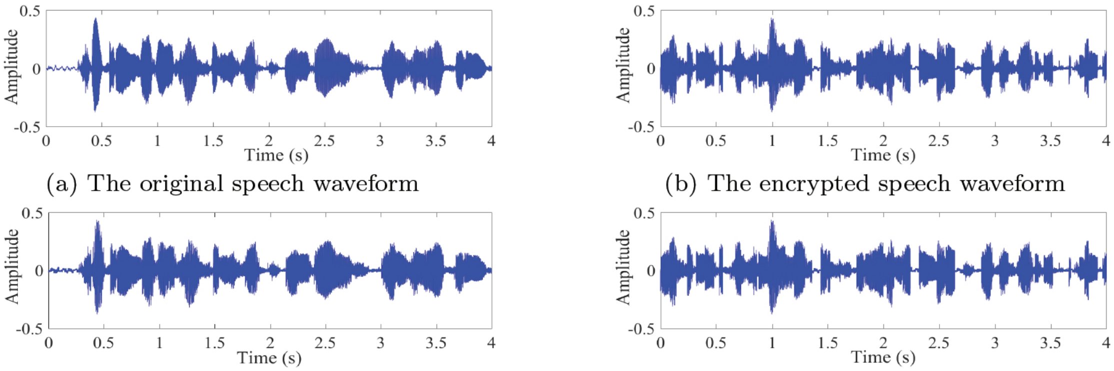
4. Experimental results and analysis

The experimental hardware platform is an Intel Core i5-5200U CPU @2.20GHz, RAM 4 GB, and the experimental software platform is MATLAB R2016a under the Windows 10 operating system. The experimental data used are the speech samples in the Texas Instruments and Massachusetts Institute of Technology (TIMIT) and Text to Speech (TTS) speech libraries; the speeches are encrypted to generate an encrypted speech library in the cloud. The speech library contains 640 segments of audio recorded by men and women, yielding a total of 1280 segments. Each speech segment is 4 s in length and in WAV format. The speeches adopt a 16-kHz sampling frequency with 16-bit sampling accuracy.

4.1. Encryption performance analysis

In order to solve the privacy and security of speech storage or transmission in the cloud, an encryption algorithm based on Henon mapping combined with time–frequency domain grouping scrambling is designed. It has good encryption performance and great security. We select a 4-s speech segment from the original speech library with a sampling frequency of 16 kHz. The specific encryption steps are as follows. (1) Two pseudorandom sequences are generated by Henon mapping and the speech signal is framed. (2) Each frame in the DCT domain is divided into a group per 4 sampling points, in total 64 groups. (3) The intraframe groups are scrambled by one of the two pseudorandom sequences. (4) After the speech inverse DCT, the speech in the time domain is divided into a group per 5 frames, in total 50 groups. (5) Another random sequence is used for scrambling intergroups in the time domain, and the encrypted speech is output.

Figure 3 shows the speech waveforms before and after encryption and decryption. Figure 3a shows the original speech waveform, Figure 3b shows the encrypted speech waveform, Figure 3c shows the speech waveform following successful decryption, and Figure 3d shows the speech waveform following failed decryption.

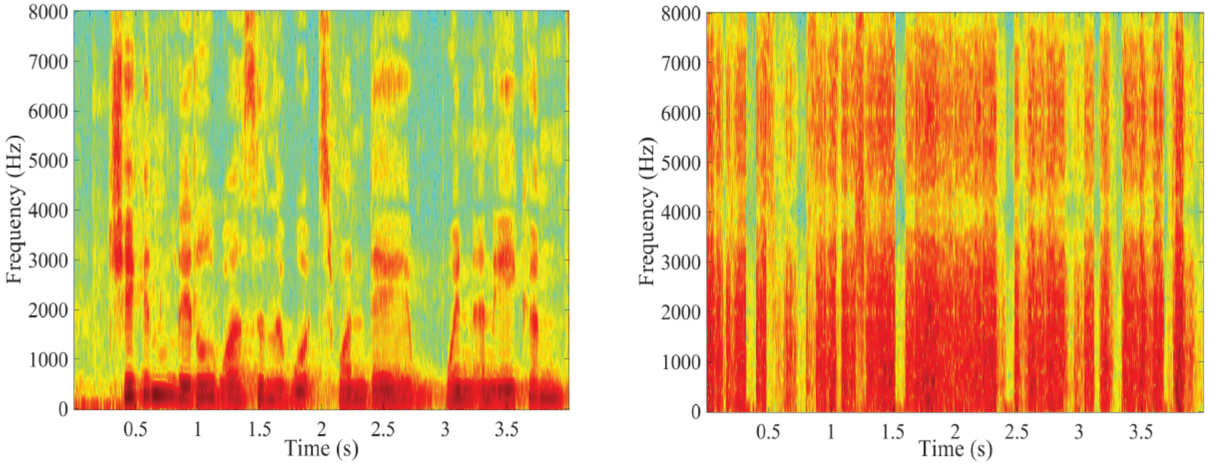


(a) The original speech waveform (b) The encrypted speech waveform

(c) The speech waveform following successful decryption (d) The speech waveform following failed decryption.

Figure 3. The speech encryption and decryption waveform comparison.

The waveform of Figure 3b is clearly disorganized compared with that of Figure 3a. This shows that the encryption algorithm has good encryption performance and can guarantee speech privacy in the cloud. According to the number of intergroups and intragroups of the encryption algorithm, the length of the key is $64! \times 50!$, which can resist general key attacks. In this way, the privacy protection of the speech can be ensured during transmission. Meanwhile, the authentication digest is directly extracted from the encrypted speech for authentication. To illustrate the performance of the encryption algorithm from the perspective of peaks and energy, the speech spectra of the speech before and after encryption are shown in Figure 4.



(a) The original speech spectrogram (b) The encrypted speech spectrogram

Figure 4. Comparison of speech spectrograms before and after encryption.

By comparing Figure 4a with Figure 4b, we can see that the energy in Figure 4a is mainly concentrated in the low-frequency band of 0–1 kHz, while the energy in Figure 4b is diffused in the whole frequency band. The peak of the speech signal in Figure 4a can be clearly seen, but the peak of the speech signal in Figure 4b cannot be seen clearly. This indicates that the encryption algorithm performs well. Meanwhile, it can be seen from Figure 4b that the encrypted speech spectrum does not completely turn into noise and whiten, which makes it possible to extract the perceptual features in the encrypted speech.

As discussed above, the energy and peak of the speech signal reveal the encryption performance from an intuitive perspective. From an objective evaluation perspective, the perceptual evaluation of speech quality (PESQ) is used to analyze the encrypted and decrypted speech. The MOS value is used to evaluate the speech quality. The score of PESQ-MOS ranges from 1.0 (worst) to 4.5 (best). It is hoped that the MOS value of the encrypted speech will be close to 1.0 or smaller than 1.0, which means that the encryption performance is good. The quality of the recovered speech after decryption is expected to be close to 4.5, which would mean that the quality of speech recovery is excellent. In the experiment, we select 20 representative speeches from the encrypted speech library to test PESQ-MOS values and take the average MOS value. Table 2 compares encryption and decryption PESQ-MOS values from Ref. [18] and the proposed method.

Table 2. The PESQ-MOS value of encrypted and decrypted sample speeches.

Type	The proposed method	Ref. [18]
Encrypted sample speech	0.5591	1.0350
Decrypted sample speech	4.5000	4.5000

As shown in Table 2, the PESQ-MOS value after speech encryption is less than 1.0. It is nearly half that of Ref. [18], which illustrates that our method has better encryption performance. Furthermore, the PESQ-MOS value of the decrypted speech is 4.5, which shows that the recovery quality is excellent. Therefore, the proposed encryption method can be effectively applied to privacy protection, and it can extract the perceptual features directly from the encrypted speech for authentication, which greatly simplifies the authentication model.

4.2. Discrimination performance analysis

Discrimination is an important evaluation indicator of perceptual hashing performance. It is measured by the false accept rate (FAR), which is a BER calculated from two different speech segments. The discrimination in this paper is calculated by 1280 encrypted speech segments. These speech segments obtain a total of 816,004 BER values. Figure 5 shows the normal distribution diagram of the BER for different speech contents.

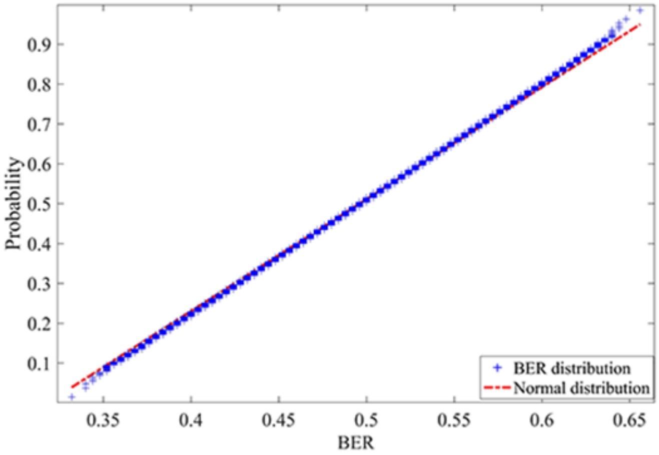


Figure 5. The normal distribution diagram of different speech content BERs.

According to the De Moiver–Laplace center limit theorem, the Hamming distance approximately obeys the normal distribution of $\mu = p, \sigma = \sqrt{\mu(1-\mu)/N}$, where N is the length of the perceptual hashing sequence, μ is the BER mean, σ is the BER standard deviation, and p is the probability that a perceptual hashing sequence occurs. In this paper, N=250. We calculate the theoretical normal value $\mu = 0.5$ and standard deviation $\sigma = 0.0316$ according to the De Moiver–Laplace center limit theorem. The experimental values are $\sigma = 0.0347$ and $\mu = 0.4966$, which are very close to the theoretical values. As shown in Figure 5, the BER distribution map almost coincides with the standard normal distribution line, which illustrates that the BER distribution diagram obeys the approximate normal distribution. Hence, the proposed method has better discrimination.

Considering the above analysis, the discrimination is discussed from a theoretical view. We intend to conduct further analysis of the perceptual hashing discrimination, and the FAR is calculated according to (8):

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

where μ is mean BER, τ is an authentication threshold, σ is the standard deviation, and x is BER.

Table 3 compares the FAR values under different thresholds from the proposed method and those in Ref. [12, 13, 17, 18].

The smaller the FAR value, the better the discrimination. After applying the perceptual hashing algorithm, the FAR value of the encrypted speech is significantly larger than that of the original speech because the encryption algorithm loses the speech perceptual feature while exchanging good encryption performance. Therefore, if we want to extract good perceptual features, we have to balance between encryption performance and perceptual hashing performance. The FAR value of the proposed method applied to the encrypted speech

Table 3. Comparison of FAR values under different thresholds.

τ	The proposed method		Ref. [12]	Ref. [13]	Ref. [17]	Ref. [18]
	Encrypted speech	Original speech				
0.1	1.49×10^{-30}	5.20×10^{-35}	2.95×10^{-22}	2.94×10^{-21}	1.59×10^{-29}	5.25×10^{-27}
0.15	8.56×10^{-24}	3.44×10^{-27}	3.42×10^{-17}	1.14×10^{-16}	5.56×10^{-23}	4.11×10^{-21}
0.20	6.28×10^{-18}	2.09×10^{-20}	1.38×10^{-13}	1.11×10^{-12}	2.63×10^{-17}	5.45×10^{-16}
0.25	5.94×10^{-14}	1.16×10^{-14}	1.94×10^{-10}	2.75×10^{-09}	1.69×10^{-12}	1.23×10^{-11}
0.30	7.32×10^{-9}	6.09×10^{-10}	9.69×10^{-07}	1.68×10^{-06}	1.51×10^{-08}	4.77×10^{-08}

or the original speech is significantly smaller than those of Ref. [12, 13, 17, 18]. Ref. [18] uses short-term cross-correlation to construct a perceptual hashing, which causes the correlation after speech encryption to decrease, thus leading to a decrease in the FAR value. In [12], because the linear prediction analysis is approximated by the minimum mean square error, there is a certain error that causes the FAR value to decrease. Ref. [13] adopts a 16-kHz sampling frequency for this paper, resulting in the number of frames decreasing. However, it is sensitive to the number of frames, thus leading to a decrease in the FAR value. When the matching threshold is τ =0.3, only 7.32 segments are misjudged for 10^9 speeches in the proposed method. From Table 3, we see that the FAR value of this study is lower than the FAR of the comparison algorithm. In summary, the discrimination of the proposed method is better than those of Ref. [12, 13, 17, 18].

4.3. Robustness performance analysis

Another evaluation indicator of perceptual hashing is robustness. We usually do some CPOs on the speech signals while using them, such as volume increase or decrease etc, and these operations usually affect the speech perceptual features. Therefore, the designed hashing algorithm must have robustness for conventional operations. For testing the robustness of the perceptual hashing algorithm, nine kinds of CPOs are selected for speech processing. The specific CPOs are shown in Table 4.

Table 4. CPO types.

Type	Parameters	Abbreviation
Volume adjustment 1	−50%	$V \uparrow$
Volume adjustment 2	+50%	$V \downarrow$
Resampling 1	$16 \rightarrow 8 \rightarrow 16$ kHz	$R8 \rightarrow 16$
Resampling 2	$32 \rightarrow 8 \rightarrow 16$ kHz	$R32 \rightarrow 16$
MP3 compression 1	32 kbps	M.32
Filter 1	12 step Butterworth filtering 3.4 kHz	B.W
Filter 2	12 step FIR filtering 3.4 kHz	F.I.R
Echo	Delay 100ms, −50%	E

In order to better declare the robustness of the proposed method, Refs. [12, 13, 17, 18] are used for the same CPOs. Table 5 shows the comparison results of average bit error rate.

As shown in Table 5, the robustness of the proposed method is better than that of Ref. [12] because Ref. [12] uses LPCs. In linear prediction analysis, sampling points of speech can be approximated from several

Table 5. Comparison of average BER.

τ	The proposed method		Ref. [12]	Ref. [13]	Ref. [17]	Ref. [18]
	Encrypted speech	Original speech				
$V\downarrow$	0.0002	0.0002	0.0016	0.0040	0.0042	0.0038
$V\uparrow$	0.0173	0.0188	0.0415	0.0256	0.0039	0.0160
$R.8\rightarrow 16$	0.0026	0.0021	0.0260	0.0012	0.0026	0.0033
$R.32\rightarrow 16$	0.0221	0.0200	0.1219	0.0098	-	0.0223
M.32	0.0322	0.0376	0.1147	0.0218	0.0016	0.0090
M.128	0.0059	0.0049	0.0727	0.0035	-	0.0086
B.W	0.1412	0.0445	0.4098	0.1422	-	0.1339
F.I.R	0.1520	0.0496	0.4303	0.1615	-	0.1394
E	0.1137	0.0946	0.2015	0.0923	-	-

past sample values, which is greatly affected by the amplitude of the speech. Hence, the robustness of Ref. [12] is worse for volume increase and filtering. Compared with Ref. [13], the volume increase, volume decrease, MP3 compression, and filtering have better robustness; however, the resampling has less robustness because Ref. [13] adopted 44.1-kHz sampling in the original paper but has 16-kHz sampling in this paper. This leads to a frame length reduction and a decrease for perceptual features, which reduces the overall robustness of the algorithm. Compared with Ref. [17], the volume decrease and the resampling in the original domain have better robustness, but the volume increase and the MP3 compression are less robust. The robustness of volume decrease, resampling, and 128 kbps compression is better than that of Ref. [18], but 32 kbps compression and filtering are less robust. This is because the band variance, which contains energy, is sensitive to energy fluctuations, and filtering and echo operations can change the energy. Thus, the filtering and echo operations have a great impact on speech.

The false reject rate (FRR) can also be used to evaluate the robustness; it determines the probability that two identical contents are judged as different contents. The formula for calculating FRR is shown in (9):

$$FRR(\tau) = 1 - \int_{-\infty}^{\tau} f(x|\mu,\sigma)dx = 1 - \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx, \tag{9}$$

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

To evaluate the overall performance of the hashing algorithm, the FAR–-FRR curve is drawn from the BERs of discrimination and robustness. Figure 6 compares the FAR-–FRR curves of the proposed method, Ref. [12], and Ref. [13].

In Figure 6, we can see that the FAR and FRR curves intersect because the robustness of Ref. [12] is worse and Ref. [13] has less discrimination. However,the FAR and FRR curves in our method do not intersect, which shows that the proposed method's overall performance is better than those of Refs. [12, 13]. Hence, the proposed method is suitable for speech authentication.

4.4. Authentication efficiency analysis

In an actual speech authentication scene, authentication efficiency is one of the best evaluation indicators. Therefore, for illustrating the authentication efficiency of the algorithm, 100 speeches are selected from the

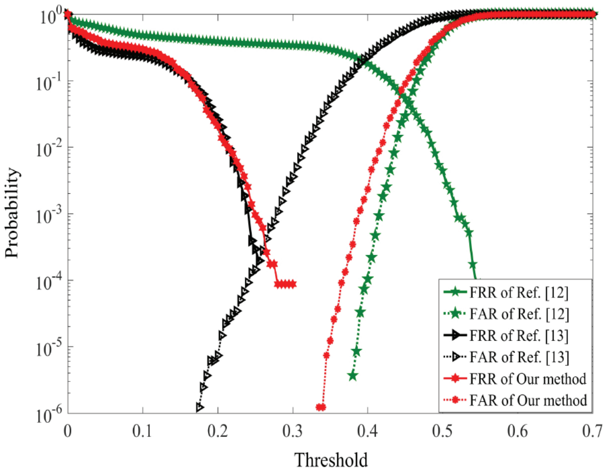


Figure 6. Comparison of the FAR–-FRR curves between the proposed method and the algorithms in [12, 13].

speech library. The total computation time for generating the perceptual hashing sequence and matching hashing sequence is calculated to measure the authentication efficiency. Table 6 compares the authentication efficiency of the proposed method and those of Refs. [12, 13, 18].

Table 6. Comparison of authentication efficiency of different algorithms.

	The proposed method	Ref. [12]	Ref. [13]	Ref. [18]
Work frequency (GHz)	2.2	3.3	2.5	2.2
Total (s)	2.08	12.47	130.4	13.84
Length of hashing sequence (bits)	250	360	360	248

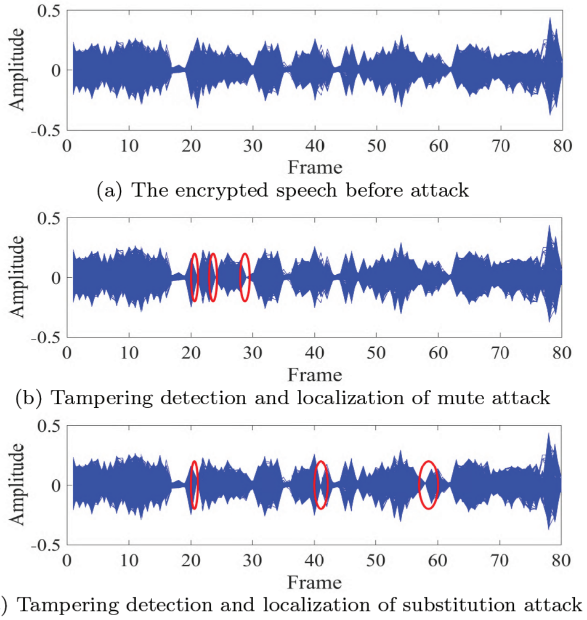
As can be seen, the authentication efficiency of the proposed method is 62 times faster than that of the Ref. [13] and 6 times faster than those of Refs. [12, 18]. This is because Refs. [12, 13, 18] adopt the NMF dimension reduction method for data dimension reduction, which leads to a significant decrease in authentication efficiency. The above analysis shows that the proposed method has the highest authentication efficiency. Moreover, the length of the hashing sequence is 250, which indicates good compactness. The proposed method is very simple and easy to implement, and it is suitable for real-time authentication in practical applications.

4.5. Tampering detection and location

It is necessary to determine the tampered location when the speech does not pass the authentication. Generally, malicious attacks affect a significant amount of speech content, changing the speech's meaning and causing misunderstanding.

In order to test the method's ability to locate the tampered location, two kinds of forgeries are performed on the encrypted speech. Figure 7 shows tampering detection and localization schematic diagrams for two kinds of attacks. Figure 7a shows the encrypted speech before attack.

1) Mute attack: Hackers often set some significant content in the speech to zeros when tampering with the speech. Hence, we set 21, 24, and 29 frames of the authentication speech to zeros. In Figure 7b, the red circles mark the falsified 21, 24, and 29 frames. The proposed method can precisely locate the tampering position of the mute attack.



(a) The encrypted speech before attack

(b) Tampering detection and localization of mute attack

(c) Tampering detection and localization of substitution attack

Figure 7. Tampering detection and localization.

2) Substitution attack: Substitution attacks also occur in tampered speech. Hackers usually replace some important information in the speech, causing misunderstandings by the recipient of the message. Therefore, in this experiment, we substitute 21, 41, and 58 frames of the authentication speech. As shown in Figure 7c, the red circles mark the falsified 21, 41, and 58 frames. The proposed method can accurately locate the tampered position of the substitution attack.

According to the above analysis, we believe that the proposed method can accurately locate the tampered area if the encrypted speech is attacked. Additionally, when the tampered speech is decrypted, the tampered frames are diffused to the entire speech. The decrypted speech waveform is smooth, and we do not find differences between the original and decrypted tampered speech. Consequently, we must detect whether the unauthenticated speech was tampered with by hackers. If the speech was tampered with by malicious attacks, we need to locate the tampered area for subsequent recovery work. In this study, our method can accomplish speech authentication well; in addition, it can detect and locate tampered areas for mute and substitution attacks.

Table 7 compares the performance of the proposed method and those of Refs. [12, 13, 17, 18].

Table 7. Comparison of performance for different methods.

Method	Encryption	Feature extraction	Application	Tamper location
Ref. [12]	No	Original speech	Authentication	No
Ref. [13]	No	Original speech	Authentication	No
Ref. [17]	Yes	Original speech	Retrieval	-
Ref. [18]	Yes	Encrypted speech	Retrieval	-
The proposed method	Yes	Encrypted speech	Authentication	Yes

As shown in Table 7, the proposed method protects speech privacy and can accurately locate tampered positions. For the authentication digest, Ref. [17] extracts it from the original speech as a watermark embedded into the encrypted speech, which undoubtedly complicates the model. However, the authentication digest is extracted directly from the encrypted speech in our method, making the entire system simple. Moreover, the perceptual hashing method in this paper is better than that of Ref. [18]. Meanwhile, our method can achieve privacy protection and content authentication, and it has a strong ability to detect and locate tampered areas. Hence, it is appropriate for speech authentication in the cloud storage environment.

5. Conclusions and future work

In this paper, for the specific application of the encrypted speech retrieval system, in order to solve the problem of the content authentication of the queried speech, an encrypted speech authentication method based on uniform subband spectrum variance and perceptual hashing is proposed. The proposed method consists of encryption, perceptual hashing authentication, and determining the locations of tampering. First, an encrypted speech algorithm using Henon mapping is proposed for speech privacy in the cloud. Moreover, a perceptual hashing algorithm based on uniform subband spectrum variance is proposed for encrypted speech authentication. The experimental results show that the proposed method has better robustness, discrimination, and high efficiency. It also has a strong ability to detect malicious attacks and precisely locate the tampered position. Meanwhile, the proposed method can directly extract the authentication digest from encrypted speech. Therefore, the proposed method is applicable to speech privacy protection and content authentication in the cloud.

In future work, we plan to research the recovery of encrypted speech after tampering.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61862041, 61363078). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] Barona R, Anita EAM. A survey on data breach challenges in cloud computing security: issues and threats. In: IEEE 2017 Circuit, Power and Computing Technologies; Kollam, India; 2017. pp. 1-8.

[2] Wang HX, Zhou LN, Zhang W, Liu S. Watermarking-based perceptual hashing search over encrypted speech. In: International Workshop on Digital Watermarking; Auckland, New Zealand; 2013. pp. 423-434.

[3] Thangavel M, Varalakshmi P, Renganayaki S, Subhapriya GR, Preethi T et al. SMCSRC—Secure multimedia content storage and retrieval in cloud. In: IEEE 2016 International Conference on Recent Trends in Information Technology; Chennai, India; 2016. pp. 1-6.

[4] Qian Q, Wang HX, Hu Y, Zhou LN, Li JF. A dual fragile watermarking scheme for speech authentication. Multimedia Tools and Applications 2016; 75 (21): 13431-13450. doi: 10.1007/s11042-015-2801-4

[5] Mossa E. Security enhancement for AES encrypted speech in communications. International Journal of Speech Technology 2017; 20 (1): 163-169. doi: 10.1007/s10772-017-9395-3

[6] Farsana FJ, Gopakumar K. Private key encryption of speech signal based on three dimensional chaotic map. In: IEEE 2017 Communication and Signal Processing; Chennai, India; 2017. pp. 2197-2201.

[7] Wang S, Miyauchi R, Unoki M, Nam S. Tampering detection scheme for speech signals using formant enhancement based watermarking. Journal Information Hiding Multimedia Signal Process 2015; 6 (6): 1264-1283.

[8] Wang J, He J. A speech content authentication algorithm based on a novel watermarking method. Multimedia Tools and Applications 2017; 76 (13): 14799-14814. doi: 10.1007/s11042-016-4027-5

[9] Liu ZH, Wang HX. A novel speech content authentication algorithm based on Bessel–Fourier moments. Digital Signal Processing 2014; 24: 197-208. doi: 10.1016/j.dsp.2013.09.007

[10] Liu ZH, Huang JW, Sun XM, Qi CD. A security watermark scheme used for digital speech forensics. Multimedia Tools and Applications 2017; 76 (7): 9297-9317. doi: 10.1007/s11042-016-3533-9

[11] Qian Q, Wang HX, Shi CH, Wang H. An efficient content authentication scheme in encrypted speech based on integer wavelet transform. In: IEEE 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference; Jeju, South Korea; 2016. pp. 1-8.

[12] Chen N, Wan WG. Robust speech hash function. ETRI Journal 2010; 32 (2): 345-347. doi: 10.4218/etrij.10.0209.0309

[13] Li JF, Wang HX, Jing Y. Audio perceptual hashing based on NMF and MDCT coefficients. Chinese Journal of Electronics 2015; 24 (3): 579-588. doi: 10.1049/cje.2015.07.024

[14] Zhang QY, Xing PF, Huang YB, Dong RH, Yang ZP. An efficient speech perceptual hashing authentication algorithm based on DWT and symmetric ternary string. International Journal of Information and Communication Technology 2018; 12 (1/2): 31–50. doi: 10.1504/IJICT.2018.089021

[15] Zhang QY, Hu WJ, Huang YB, Qiao SB. An efficient perceptual hashing based on improved spectral entropy for speech authentication. Multimedia Tools and Applications 2018; 77 (2): 1555–1581. doi: 10.1007/s11042-017-4381-y

[16] Zhao H, He SF. A retrieval algorithm for encrypted speech based on perceptual hashing. In: IEEE 2016 Natural Computation, Fuzzy Systems and Knowledge Discovery; Changsha, China; 2016. pp. 1840-1845.

[17] He SF, Zhao H. A retrieval algorithm of encrypted speech based on syllable-level perceptual hashing. Computer Science Information System 2017; 14 (3): 703-718. doi: 10.2298/CSIS170112024H

[18] Zhang QY, Zhou L, Zhang T, Zhang DH. A retrieval algorithm of encrypted speech based on short-term cross-correlation and perceptual hashing. Multimedia Tools and Applications 2019; 78 (13): 17825-17846. doi: 10.1007/s11042-019-7180-9

[19] Ping P, Xu F, Mao YC, Wang ZJ. Designing permutation–substitution image encryption networks with Henon map. Neurocomputing 2018; 283: 53-63. doi: 10.1016/j.neucom.2017.12.048

[20] Jiang N, Dong X, Hu H, Ji ZX, Zhang WY. Quantum image encryption based on Henon mapping. International Journal of Theoretical Physics 2019; 58 (3): 979-991. doi: 10.1007/s10773-018-3989-7

[21] Belkhouja T, Du X, Mohamed A, Ali A, Guizani M. Symmetric encryption relying on chaotic Henon system for secure hardware-friendly wireless communication of implantable medical systems. Journal of Sensor Actuator Networks 2018; 7 (2): 21. doi: 10.3390/jsan7020021

[22] Wang W, Hu GM, Yang L, Huang DF, Zhou Y. Research of endpoint detection based on spectral subtraction and uniform sub-band spectrum variance. Audio Engineering 2016; 40 (5): 40-43 (in Chinese). doi: 10.16311/j.audioe.2016.05.09