Multimedia Tools and Applications (2021) 80:24925–24948
https://doi.org/10.1007/s11042-021-10905-0

An encrypted speech authentication and tampering recovery method based on perceptual hashing

Qiu-yu Zhang [1] · Deng-hai Zhang [1] · Fu-jiu Xu [1]

Received: 27 April 2020 / Revised: 23 December 2020 / Accepted: 1 April 2021 /
Published online: 12 April 2021

Abstract
With the progress of speech retrieval technology in the cloud, it brings a lot of conveniences for speech user. Yet, the inquiry encrypted speech results from the speech retrieval system are faced with some secure issues to settle, such as integrity authentication and tampering recovery. In this paper, an encrypted speech authentication and tampering recovery method based on perceptual hashing is proposed. Firstly, the original speech is scrambled by Duffing mapping to construct an encrypted speech library in the cloud, through extracting product of uniform sub-band spectrum variance and spectral entropy of encrypted speech and constructing a perceptual hashing sequence to generate the hashing template of the cloud. From this, a one-to-one correspondence between the encrypted speech and perceptual hashing sequence is established. Secondly, the authentication digest of encrypted speech is extracted according to the inquiry result during the retrieval. Then, the authentication digest and the perceptual hashing sequence of the hashing template in the cloud are matched by the Hamming distance algorithm. Finally, for encrypted speech that fails authentication, tampering detection and location are performed, and the tampered samples are recovered by the least square curve fitting method. The simulation results show that the proposed method can extract the authentication digest directly in the encrypted speech, and the authentication digest not only has good discrimination and robustness, but it accurately locates the tampered area for malicious substitution and mute attacks. In addition, the proposed method can recover tampered speech signals in high quality without any extra information.

Keywords Encrypted speech authentication · Perceptual hashing · Least square curve fitting · Tampering location · Tampering recovery

1 Introduction

With the vigorous advancement of the cloud industry, more and more users choose the convenient third-party cloud servers to store multimedia data, such as speech. However, the

✉ Qiu-yu Zhang
zhangqylz@163.com

Extended author information available on the last page of the article

security of data in the cloud surrounding has hindered its promotion and application [6, 10]. In order to protect the privacy of sensitive speech, many companies and users usually encrypt the speech before uploading it to the cloud storage center. Authorizers retrieve or download the encrypted data from cloud storage when they need these data [28]. If these encrypted speech were maliciously attacked by illegal attackers, and the encrypted speech would lose its original meaning, causing immeasurable consequences and losses to users [22]. Therefore, in the application of the encrypted speech retrieval system, how to ensure the authenticity of the encrypted speech and recover the tampered speech content after a malicious attack are the current research hotspots in the field of multimedia information security and retrieval. And the solution of these problems will also promote the popularization and application of cloud storage or encrypted speech retrieval technology.

Most of the existing privacy protection schemes usually used traditional encryption algorithms [11, 15, 25]. These encryption methods had a large amount of calculations, complicated operations. Other methods was digital steganography for privacy protection [1, 9, 26, 32], which realized the protection and transmission of private information by hiding secret speech in unimportant positions of the transform domain. Additionally, existing encryption methods based on chaos and pseudo-random number generators achieved the purpose of encryption by pseudo-random sequences, and replacing speech coefficients to lose speech characteristics [5, 20, 27]. But, these methods lost a lot speech features in exchange for the encryption performance. Consequently, none of the above encrypted methods can guarantee the privacy and security of the speech, but also can directly extract the speech authentication digest from the encrypted speech.

Many scholars had given different schemes for speech integrity authentication. In [3, 8, 13, 23, 31], the authentication methods mainly achieve the purpose of authentication by embedding additional authentication information into the speech signal, and then extracting the watermark and comparing it with the original watermark information. Such an authentication method is that the process of embedding and extracting the watermark is added, which reduces the audible quality of the speech and increases the complexity of the authentication system.

At present, some scholars proposed many schemes in digital watermarking systems for recovering maliciously tampered samples as much as possible. In [14, 16–19, 24], the authors embed additional reference information into the original speech signal. If the speech was tampered with, the reference information was extracted to restore the tampered speech content. These methods cannot guarantee acceptable speech recovery quality, and the recovery method were more complicated. Such as Liu et al. [14] proposed a block-based large-capacity embedding watermarking scheme, which embedded the compressed signal of the speech signal as a watermark signal in the speech. The corresponding compressed signal was extracted for reconstruction and recovery when the speech was tampered with. But, the reconstructed speech signal still had a large error from the real value.

By analyzing the above research, many authentication and tampering recovery methods based on digital watermarking without considering the privacy and security of speech. Accordingly, these digital watermarking methods required to embed additional information into the speech signal for authentication and recovery, which leaded to the decline of speech quality and complicated the entire authentication system. Compared with digital watermarking methods, speech authentication methods based on perceptual hashing [2, 12, 33] directly extracted perceptual features from speech to construct perceptual hashing sequences, thus, through comparing hashing sequences performed authentication. These methods all extracted perceptual features from original domain without considering the privacy and security of

speech, which were not conducive to the transmission of speech in the cloud. In [4, 35], the perceptual hashing was extracted from the original domain as a digital watermark embedded into encrypted speech, which increased the complexity of the speech retrieval system. Currently, although some scholars had proposed perceptual hashing from encrypted speech signals for retrieval [34], but, there is relatively little research on integrity authentication and tampering recovery of encrypted speech. Thence, the traditional original domain speech authentication scheme cannot meet the privacy and security of current speech storage and transmission in the cloud. Furthermore, encryption may make the speech lost its speech characteristics, making it relatively difficult to extract authentication information directly in the encrypted speech. Moreover, none of the existing speech content authentication schemes based on perceptual hashing gave how to recover the sampled values that had been tampered with as much as possible if the speech had been tampered with.

In the above speech authentication schemes, combined with the practical application of content based encrypted speech retrieval system in the cloud environment, this paper uses speech perceptual hashing technology to realize secure and recoverable encrypted speech content authentication. Hence, an encrypted speech authentication and tampering recovery method based on perceptual hashing is proposed. The main contributions of this paper can be summarized as follows:

1) A time-frequency domain scrambled speech encryption method based on Duffing mapping is designed in this paper. It realizes the privacy and security of speech stored in the cloud. At the same time, the speech authentication digest can be extracted directly from the encrypted speech.
2) An encrypted speech perceptual hashing scheme is designed, which has high authentication efficiency. The scheme also can accurately detect and locate tampering position when the encrypted speech is subjected to malicious tampering.
3) The least square curve fitting method can be used to recover the tampered samples with high quality in this paper.

Compared with the existing speech content authentication methods based on perceptual hashing and digital watermarking, the proposed method can directly extract the speech authentication digest from encrypted speech, and the speech authentication system model is simple and efficient. Simultaneously, the proposed method can accurately detect and locate tampering position, and the tampered speech can be recovered with high quality after malicious attacks.

The major symbols used in this paper are summarized in Table 1 for easy reference.

Table 1 Notations and symbols

Symbol	Definition	Symbol	Definition
XX	the sub-band	*DH*	product of uniform sub-band variance and spectrum entropy
E	the sub-band mean	SNR	signal to noise ratio
H(i)	spectral entropy	SegSNR	segment signal to noise ratio
W(:,:)	bit error rate	*X(n)*	Speech signal
FAR	false accept rate	*FRR*	false reject rate

The rest of this paper is organized as follows: Section 2 describes the related theories. Section 3 gives tampering recovery principle based on the least square curve fitting. Section 4 presents the encrypted speech content authentication model of this paper, and describes in detail the processing processes of speech encryption algorithm and perceptual hashing construction. Section 5 gives the experimental results and performance analysis as compared with other related methods. Finally, the conclusions are described in Section 6.

2 Related theory

2.1 Duffing mapping

Duffing mapping [7] was widely used in image encryption. In order to improve the key security of the speech scrambling encryption method, the Duffing chaotic system is introduced into the encryption algorithm to generate keys in this paper. The Duffing mapping is defined as follows:

$$T:\begin{cases} x_{i+1} = y_i \\ y_{i+1} = -bx_i + ay_{i+1} - y_i^3 \end{cases} \tag{1}$$

As can be seen from Eq. (1), the Duffing chaotic system iterates the entire chaotic equation by two variables x and y. The two constants a and b are usually set to $a = 2.75$ and $b = 0.2$, and the system is in a chaotic state. The Duffing chaotic system not only has an extremely sensitive dependence on the initial value, but also has excellent pseudo-randomness. It meets the requirements of various statistical characteristics, and is fit to generate an encryption key. In summary, Duffing mapping is easy to implement, more complex, highly secure, and it is suitable as a key generator for speech encryption schemes.

2.2 Uniform sub-band spectrum variance

The existing band variance calculation is to calculate the variance of each spectral line, which has large fluctuations and low stability. Therefore, the uniform sub-band separation band variance [29] is extracted in the encrypted speech of lower signal to noise ratio (SNR) in this study.

Firstly, the fast Fourier transform (FFT) is performed on each frame of data N, and there are $(N/2 + 1)$ lines in the positive frequency domain. The $(N/2 + 1)$ DFT post-amplitude spectrum $X_i = \{X_i(1), Xi(2), \ldots, X_i(N/2 + 1)\}$ is divided into q sub-bands (i denotes the i-th frame), each sub-band contains $p = fix[(N/2 + 1)/q]$ line (where $fix[\cdot]$ indicates its integer part), the sub-band $XX_i\ (m)$ 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.

Let $XX_i = \{XX_i(1), XXi(2), \ldots, XX_i(q)\}$, the sub-band mean $E_{i,1}$ is

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

Sub-band variance $D_{i,1}$ is

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

Since the band variance can account for the undulation of the band and the involved energy. In the speech of lower SNR, the band variance of the uniform sub-band separation can better distinguish the noise and the speech segment than the existing band variance method. Therefore, the proposed method extracts the uniform sub-band band variance separation as the perceptual feature in the encrypted speech to construct a perceptual hashing sequence.

2.3 Spectral entropy

Spectral entropy [30] mainly detects the flatness of the spectrum. The calculation of spectral entropy includes the following two steps:

Step 1: After the time domain speech signal $x(t)$ is subjected to windowing framing and FFT transformation, where the energy spectrum of the k-th spectral line frequency component is $Y_i(k)$. Then the normalized spectral probability density function of each frequency component is defined as:

$$P_i(k) = \frac{Y_i(k)}{\sum_{k=0}^{N/2} Y_i(k)} \tag{5}$$

where $P_i(k)$ is the probability density corresponding to the k-th frequency component f_k of the i-th frame, and N is the length of the FFT.

Step 2: The spectral entropy $H(i)$ of the i-th frame is expressed as:

$$H(i) = \sum_{k=0}^{N/2} P_i(k) \log_{10}[P_i(k)] \tag{6}$$

2.4 Calculation of uniform sub-band variance spectral entropy product

In order to directly extract the authentication digest from the encrypted speech with low SNR for speech content authentication and tampering location, a fusion feature that constructs a uniform sub-band variance spectral entropy product was constructed. In low SNR speech, it is major to distinguish between noise and speech as much as possible. Because the features of speech and noise are significantly different in the spectral domain. Generally, the energy of a speech segment has varies greatly with the frequency band. And it has a larger peak at the formant, but these energy are smaller in the noise segment. From the uniform sub-band separation band variance of Eq. (4), it can be seen that it reflects the degree of fluctuation among bands, and also shows the short-term energy of the speech signal. For speech segment, the greater the energy. For noise segment, the smaller the energy. This shows that the uniform sub-band separation band variance can well distinguish between speech segments and noise segments in the frequency domain. And the amplitude of the speech segment has a large dynamic range relative to the noise, it contains a large amount of average information, which

indicates the entropy is large. Oppositely, the amplitude of the silent segment is small, and the distribution is relatively concentrated, which demonstrates the entropy is large. Therefore, the speech entropy is robust to noise.

The separate uniform sub-band separation band variance and entropy cannot distinguish the speech and noise well when the speech signal SNR is comparatively low. For this purpose, the method of combining uniform sub-band separation band variance and spectral entropy is used to improve the accuracy of distinguishing between speech and noise segments.

$$DH_i = (1 + |D_i * H_i|)^{1/2} \quad i = 1, 2, \ldots, n \tag{7}$$

where i is the number of frames of speech, D_i is the uniform sub-band frequency band variance calculated for each frame, H_i is the spectral entropy value for each frame, and DH_i is the calculated product of uniform sub-band frequency band variance and spectrum entropy for each frame.

2.5 Least square curve fitting

The least square method [21] is to find the best function match by minimizing the sum of the squares of the errors, which can be used to obtain unknown data and minimize the sum of the squares of the errors. When the distribution of sample points is not a straight line, polynomial curve fitting can be used to obtain the lost points. The fitting curve equation is defined as the n-th order polynomial as follows:

$$y = \sum_{i=0}^{n} a_i x_i = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 \tag{8}$$

Eq (8) is expressed as $Y = X_0 A$ in matrix form, where $X_0 = \begin{bmatrix} x_1^n & \cdots & x_1 \\ \vdots & \ddots & \vdots \\ x_k^n & \cdots & x_k \end{bmatrix}$. The A is the coefficient vector to be calculated, then $A = [a_n, a_{n-1}, \ldots, a_2, a_1, a_0]^T$. In order to find the equation coefficient A of the fitted curve, the left side of the $Y = X_0 A$ is multiplied by X_0 to obtain the $X_0^T Y = X_0^T X_0 A$. At the same time, the left side of the new equation is simultaneously multiplied by the inverse matrix of $X_0^T X_0$, and the equation coefficient $A = (X_0^T X_0)^{-1} X_0^T Y$ of the fitted curve is obtained.

The average value of between adjacent samples is used as the recovery value of the tampered samples in [24], which has a large recovery error. In this paper, the least square curve fitting is used to decrease the error between the recovered value and the actual value.

3 Principle of tampered speech recovery based on least square curve fitting

In the real cloud environment, malicious attackers may tamper with the speech to change its original meaning, owing to the cloud server is a third-party unreliable service provider. For purpose of recovering the tampered speech to further reduce the speech user's losses, an efficient tampering speech recovery technology was proposed in [24]. The receiver firstly performed tampering location for tampered encrypted speech. Then, the received user reversed encrypted speech that tampered with, which leaded the tampered speech samples to the entire

speech. Finally, the tampered samples was recovered by calculating the average value of high correlation adjacent samples. This restoration method is simple, and can approximately recover the tampered samples without additional information. However, the recovered samples in this way has a large error from the real value. Especially, in the case of large speech fluctuations, the error is greater. Therefore, in order to recover the tampered samples with high quality, the proposed model adopts the least square method curve fitting to decrease the restoration errors.

The SNR and the segment signal to noise ratio (SegSNR) are used to objectively measure the quality of the recovered speech. These indexes are statistical difference measures, the calculation formulas are given in Eqs. (9) and (10), respectively. The SNR is calculated from the entire speech signal, and SegSNR is the SNR between segments after the speech is segmented. Thence, both indicators can be used to evaluate the recovered speech quality.

$$SNR = 10\log_{10}\frac{\sum_{l=1}^{L}x^2(l)}{\sum_{l=1}^{L}(x(l)-y(l))^2} \tag{9}$$

$$SegSNR = \frac{10}{I}\sum_{i=1}^{I}\log_{10}\sum_{j=1}^{J}\frac{x^2(j)}{(x(j)-y(j))^2} \tag{10}$$

where L is the total number of samples of speech, and l is each samples; I is the total number of frames of speech, and i is each frame; J is the samples of each frame, and j is each sample in a frame; x represents the original speech signal, and y represents the recovered speech signal.

Figure 1 shows the principle of recovering speech samples based on least square curve fitting method.

In Ref. [24], the recovery error is relatively large using the adjacent samples to recover the tampered samples. In order to make the reconstructed speech value closer to the original value,

Original signal

336	412	486	525	564	503	476	579	581	564	589	434	434	433	432	331

Tampered signal

336	412	486	525	564	7	476	579	581	564	589	434	434	433	20	331

As input to find the optimal solution

336	441	468	525	500	?	571	528	571	585	499	511	511	433	?	331

Order	503------>?			Order	432----->?			Ref. [24] SNR(db)	Our method SNR(db)
	True sampling point	Ref. [24]	Our method		True sampling point	Ref. [24]	Our method		
1	503	520	518	1	432	382	461	31.3	35.5
2	*	*	517	2	*	*	433	*	42
3	*	*	532	3	*	*	403	*	33.5
4	*	*	538	4	*	*	416	*	34
5	*	*	536	5	*	*	435	*	35
6	*	*	536	6	*	*	434	*	35.2
7	*	*	535	7	*	*	434	*	38
8	*	*	533	8	*	*	429	*	37
9	*	*	536	9	*	*	463	*	32

Fig. 1 An example of tampering recovery

this paper uses the least square curve fitting to recover the tampered samples. For purpose of obtaining the best match function of the least squares curve fitting method, the average of two adjacent sample points is firstly used to obtain new speech samples. Then, the new speech samples as the input to acquire the best match function coefficient and estimate the tampered samples. The coefficient of the matched function is optimal when the input error is close to stationary data. The experiments show that the recovered samples value is closer to the real value when the order of the polynomial is 2–5. Therefore, our study selects a polynomial of 2-th order to estimate the tampered samples value.

Figure 2 shows a diagram of curve fitting of different orders of 1–9 order to obtain the best matching function. The input is the average value of the adjacent sampling points of the tampered sampling point. The output is the restored tampered sampling point.

4 The proposed authentication algorithm of encrypted speech

4.1 The authentication model of encrypted speech

Figure 3 shows the encrypted speech content authentication model for specific applications of the encrypted speech retrieval system. The model builds an encrypted speech content

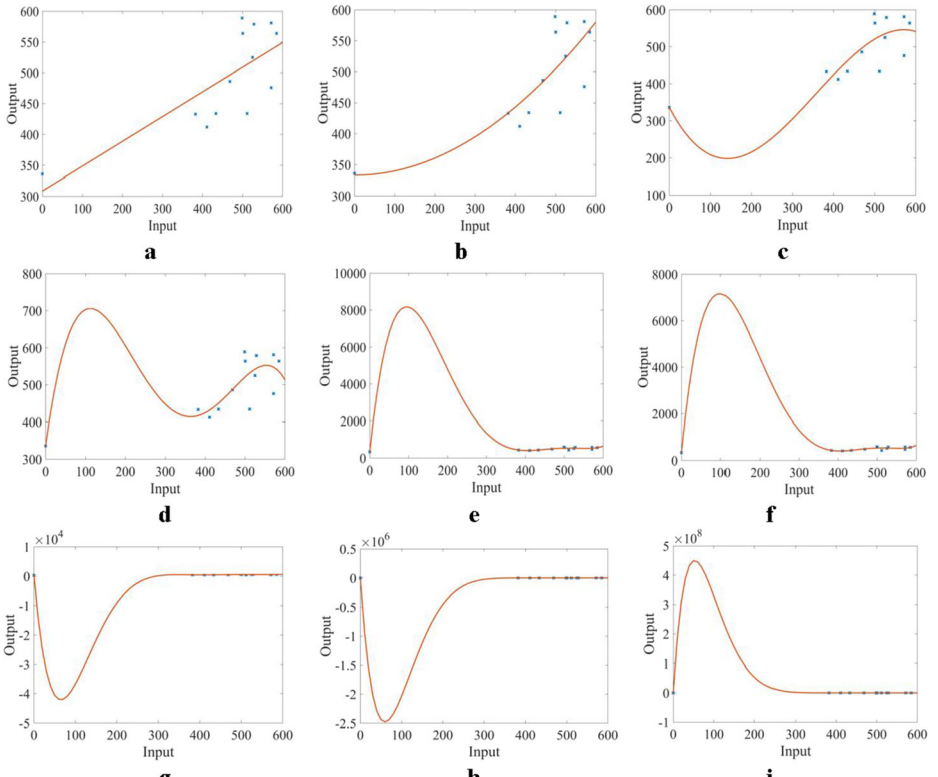

Fig. 2 Curve of 1–9 order fitting results: (**a**) 1-th (**b**) 2-th (**c**) 3-th (**d**) 4-th (**e**) 5-th (**f**) 6-th (**g**) 7-th (**h**) 8-th (**i**) 9-th

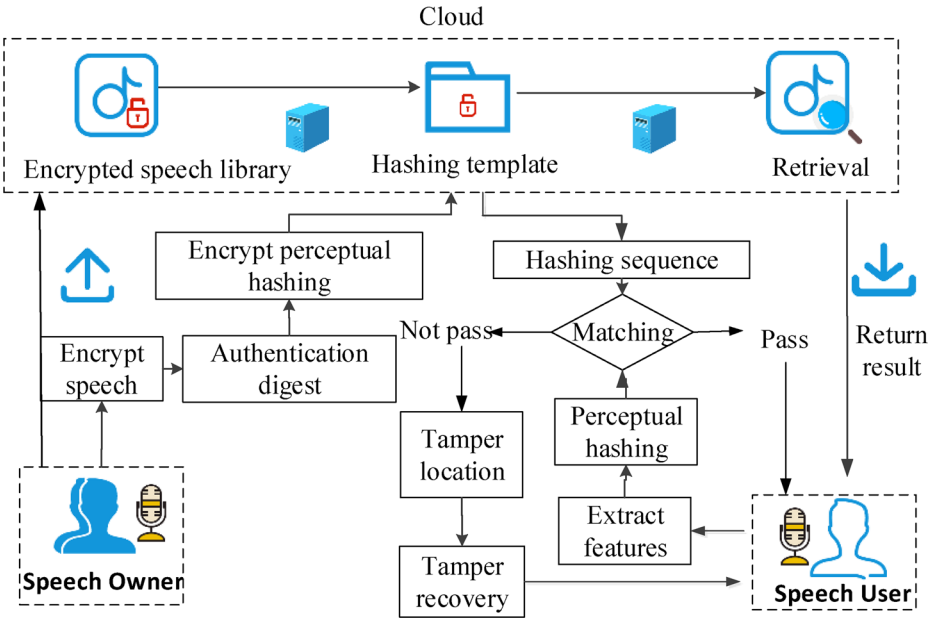

Fig. 3 The authentication model of encrypted speech based on perceptual hashing

authentication system that can directly extract the speech authentication digest and tampering recovery under the premise of ensuring the privacy of the speech.

As shown in Fig. 3, the speech holder firstly encrypt the speech by the Duffing map to construct an encrypted speech library in the cloud, while extracting the perceptual hashing sequence (authentication digest) of the encrypted speech. Secondly, the authentication digest is encrypted by Duffing map, while constructing a hashing sequence to generate a hashing template of the cloud. Finally, a one-to-one correspondence between the encrypted speech and the hashing sequence is established. After the speech users retrieve the speech, the same perceptual hashing scheme executes encrypted speech content authentication. Among them, content authentication and tampering location are mainly accomplished by an efficient perceptual hashing algorithm; recovering tampered speech is efficiently recovered by the least square curve fitting method of Section 3.

The main innovation of the proposed model is the speech can efficiently complete authentication and precisely tampering location in the encrypted speech without any additional information. And the proposed model can high quality recover tampered speech content by the least square curve fitting.

4.2 Speech encryption

In actual speech encryption, multiple rounds of operations would be performed on the speech to enhance the encryption performance. But, this would lead to excessive loss of speech features. Therefore, the proposed encryption method only adopts scrambling operations. The generated random sequence serves as a key when the Duffing map is in a chaotic state. It has strong randomness and high security. That is why this paper employs Duffing mapping to generate a random sequence of keys, and scramble each samples in the time domain. After

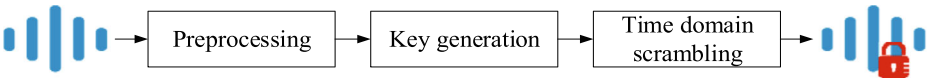

Fig. 4 Speech encryption algorithm

encryption, the speech still has some perceptual features, which can directly extract the authentication digest from the encrypted speech.

Figure 4 shows the specific encryption flow diagram and encryption steps.

4.3 The construction of encryption speech perceptual hashing sequence

In order to directly extract a better robust and discrimination authentication digest (speech perception hashing sequence) from encrypted speech with low SNR, the proposed method applies the product of uniform sub-band frequency band variance and spectral entropy as the fusion feature to construct perception hashing sequence. Figure 5 shows the construction process of speech fusion features.

The specific extraction steps are as follows:

Step 1: Speech preprocessing. Let the speech signal be $x(n)$, Using Eq. (11), Eliminate the DC component by using the Eq. (11), because the existence of the DC component will affect the calculation of the uniform sub-band spectrum variance.

$$x(n) = x(n) - mean\ (x(n)\ ) \tag{11}$$

where *mean* is the mean of the speech signal.

Step 2: Framing the windowing. Due to the spectrum is leaked by truncated the speech directly, in order to improve the occurrence of this situation, the speech signal $x(n)$ is processed by the Hamming Windowing function. 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, the i-th frame speech signal uses $x_i(m)$ expression.

Step 3: Uniform sub-band separation. According to Eq. (2), one sub-band is composed of $p = 4$ points, and $x_i(m)$ is divided into $q = 32$ sub-bands. The sub-band spectrum $XX_n(k)$ is obtained, where $1 \le k \le 32$. Define the sub-band spectrum of all frames as $XX_i = \{XX_i(1), XX_i(2), \ldots, XX_i(q)\}$, and the sub-band mean valu $E_{i,1}$ of the amplitude is calculated according to the Eq. (3). Then, using the Eq. (4), the variance of the band after each sub-band separation is calculated, which generate the feature matrix parameter $\boldsymbol{D_w}$ (1, Z_m), where Z_m represents the number of feature vectors.

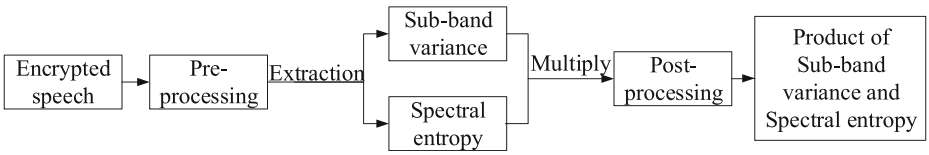

Fig. 5 The fusion feature extraction process

Step 4: Spectral entropy feature extraction. Firstly, $x_i(m)$ is subjected to FFT transformation, and according to Eq. (5), a normalized spectral probability density function $p_i(k)$ of each frequency component f_k is obtained. Then, the spectral entropy feature is obtained as H_i using Eq. (6), and a feature parameter matrix $\boldsymbol{H_w}(1, Z_m)$ is generated.

Step 5: Constructing fusion features. The extracted uniform sub-band frequency band variances $\boldsymbol{D_w}$ and spectral entropy $\boldsymbol{H_w}$ are used to obtain the fusion feature parameter $\boldsymbol{DH_w}(1, Z_m)$ according to Eq. (7).

Step 6: Hashing construction. A binary hashing construction is performed on the fusion characteristic parameter matrix $\boldsymbol{DH_w}(1, Z_m)$ to generate a hashing sequence $\boldsymbol{H}$ (1, 2, ..., q_m) of 0 and 1. The specific construction method is as follows:

Calculate the mean T of all the features in the feature parameter matrix $\boldsymbol{DH_w}(1, Z_m)$. If the feature $\boldsymbol{DH_w}$ is greater than T, it is 1. Otherwise it becomes 0;

$$H_i(x) = \begin{cases} 1, & DH_w(j+1) > T \\ 0, & Others \end{cases} \quad j = 1, 2 \cdots Z_m \tag{12}$$

where j is the j-th feature vector in the feature parameter matrix $\boldsymbol{DH_w}$.

Step 7: Hashing template encryption. To improve the security of authentication system, the proposed method scramble the generated hashing sequence. The hashing sequence H_i is scrambled using a pseudo-random sequences $V = [v_1, v_2, v_3, \ldots, v_M]$, where M is the number of perceptual hashing. And the encrypted hashing sequence is H_i^*.

4.4 Speech authentication

For speech user, the speech stored and transmitted in the cloud need to be authenticated, this is a powerful measure to prevent misunderstandings. Firstly, the perceptual hashing sequence of the speech to be authenticated is extracted. Then, the extracted hashing sequence matches with hashing template by the normalized Hamming distance to confirm whether the authentication speech has been tampered with. Suppose S_1 and S_2 represent two different speech segments, H_1 and H_2 are respectively hashing sequences generated by two speech segments, M represents a hashing sequence length. The normalized Hamming distance $W(:,:)$ represents the occurring bit error rate (BER).

$$W(S_1, S_2) = D(H_1, H_2) = \frac{1}{M} \sum_{k=1}^{M} |H_1(k) - H_2(k)| \tag{13}$$

This paper makes the following two assumptions, A and B:

A: if $W(S_1, S_2) < \tau$ is established, the authentication is passed;
B: if $W(S_1, S_2) \geq \tau$ is established, the authentication is not passed;

where τ is the hashing sequence matching threshold.

According to the above hypothesis, the authentication result is passed, if the Hamming distance is less than or equal to the matching threshold of the authentication model. Oppositely, if the Hamming distance greater than the matching threshold, the authentication result is

not pass. Then, the speech user executes tampering location for the speech that has not passed the authentication, and prepares for tampering recovery.

5 Experimental results and performance analysis

The experimental hardware platform is Itel(R) Core(TM) i5-5200U CPU @2.20GHz, RAM 4 GB, and the experimental software platform is MATLAB R2016a under Windows10 system. The experimental data uses the speech in the TIMIT (Texas Instruments and Massachusetts Institute of Technology) and TTS (Text to speech) speech libraries, which is encrypted to generate an encrypted speech library in the cloud. The speech library contains 640 segments of audio frequency recorded by men and women, 1280 segments in total. Each speech segment is 4 s long and has a WAV format. The speeches adopt 16 kHz sampling frequency with 16 bits sampling accuracy. In the experimental discussion, the encryption performance will be evaluated by the PESQ-MOS score and SNR, the performance of the perceptual hashing will be evaluated by the robustness and distinguishability, and the speech recovery quality will be evaluated by the SNR and SegSNR.

5.1 Encryption performance analysis

In the experiment, a 4 s long speech segment is selected from the speech database with a sampling frequency of 16 kHz and a total of 64,000 samples. Firstly, scramble the 64,000 samples in the time domain by Duffing map. Then, output the encrypted speech. Figure 6 shows the speech waveforms before and after encryption, decryption successful and decryption failure.

As can be seen from Fig. 6, Fig. 6(a) is completely inconsistent with Fig. 6(b). It indicates the encrypted speech waveform loses the features of the original speech waveform, making it a noisy waveform. Figure 6(c) is the speech waveform after the correct decryption, which is almost the same as the waveform of Fig. 6(a). It demonstrates encryption method can correctly decrypt the encrypted speech. From an intuitive point of view, it is verified the proposed encryption algorithm hides the speech information well, which greatly reduces the possibility that an illegal person can directly obtain the information from the encrypted speech. According to the length of the speech samples, the length of key is 64,000!, this is enough to resist general

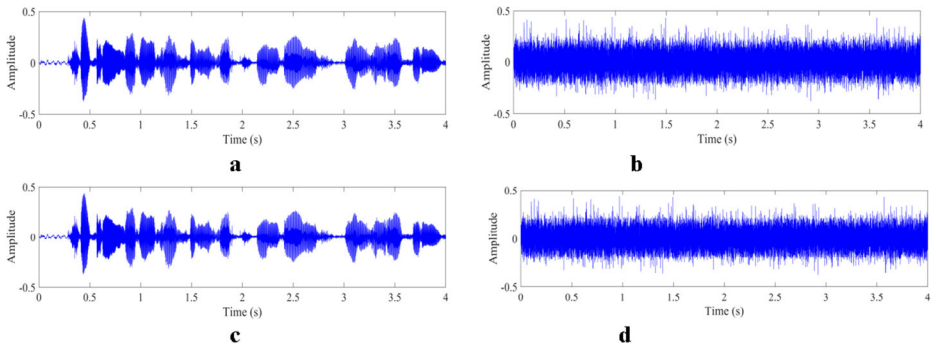

Fig. 6 The speech waveforms: (**a**) The original speech (**b**) The encrypted speech (**c**) Successful decryption speech (**d**) Failed decryption speech

key attacks. If a position in the key changes, the speech waveform will not be decrypted correctly, which is reflected in Fig. 6(d).

The Fig. 6 analyzes the encryption performance from the perspective of the time domain, and the spectrogram explains the security of the encryption algorithm in the frequency spectrum.

The spectrogram before and after speech encryption is shown in Fig. 7. Figure 7(a) is a spectrogram before speech encryption, it can be seen that the energy (darker color) of the speech signal is concentrated in the 0-1 kHz frequency band, and the energy change of the speech can be read more clearly. In Fig. 7(b), the speech energy has been scattered, and it is not concentrated in the 0-1 kHz frequency band, any speech information cannot be obtained. Figure 7(c) is the encrypted speech spectrum of the Ref. [34], which shows the clearly speech changes. Therefore, the encryption method in this study is more secure than the Ref. [34].

The above illustrates the performance of the encryption algorithm from the perspective of time and frequency domains. Then from an objective perspective (using the Perceptual Evaluation of Speech Quality (PESQ) recommended objective mean opinion score (Mean Opinion Score, MOS) and SNR) to evaluate the encryption algorithm. Generally, the score of PESQ-MOS ranges from 1 to 4.5. The PESQ-MOS value of encrypted speech hoped approach 1 or less than 1, which means the encryption performance is better. And the quality of the recovered speech after decryption is expected to be above 2.5 or closer to 4.5, it means that the quality of speech recovery is excellent.

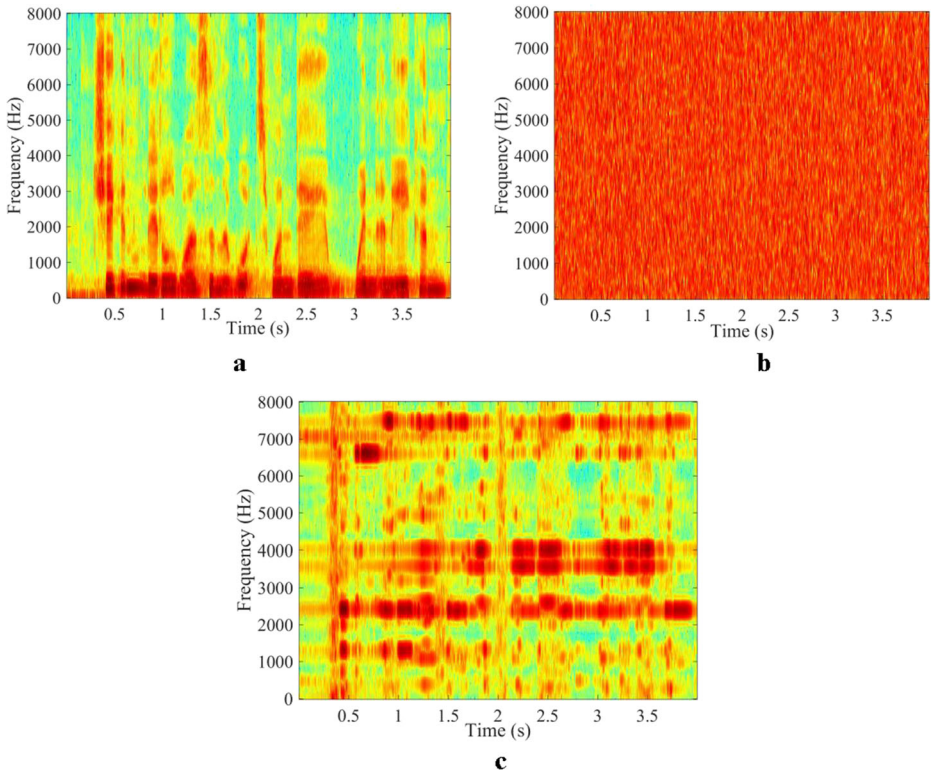

Fig. 7 The speech spectrogram before and after encryption comparison result: (**a**) the original speech spectrogram (**b**) the encrypted speech spectrogram (**c**) encrypted speech spectrogram of Ref. [34]

Table 2 PESQ-MOS/SNR of encrypted and decrypted sample speech

Method	PESQ-MOS	SNR
Ref. [34]	1.0305/4.5000	−2.5062
Our method	0.5339/4.5000	−3.0215

Table 2 shows the comparison results of the calculated average PESQ-MOS / SNR of the 10 speeches in the speech library after encryption and decryption with Ref. [34].

As shown in Table 2, the PESQ-MOS value of the proposed method is less than 1.0, which indicates the speech signal is completely noisy. In such speech noise, the illegal person cannot obtain any speech information. Besides, the SNR of the original speech and the encrypted speech signal is also calculated in this paper, which is also lower than 0. It proves that the proposed encryption algorithm can effectively ensure the privacy and security of the speech in the cloud. Compared with Ref. [34], the objective evaluation score of PESQ-MOS and SNR are higher than our method. This shows that the security of our method is higher than Ref. [34]. And the score of the decrypted speech quality PESQ-MOS is 4.5, which declares the decrypted speech quality is wonderful.

5.2 Discrimination performance analysis

To validate the performance of the proposed perceptual hashing algorithm, the false accept rate (FAR) is used to explain the perceptual hashing performance. The discrimination of the algorithm is measured by calculating the BER obtained from two different speech segments. A total of 816,004 BER values can be obtained in this study. The normal distribution of the BER of different content speech is as shown in Fig. 8.

According to the De Moiver-Laplace center limit theorem, the Hamming distance approximates obeys the normal distribution of $\mu = p, \delta = \sqrt{p(1-p)/N}$, N is the perceptual hashing sequence, μ is the BER mean, and δ is the BER, p is the probability of the perceptual hashing

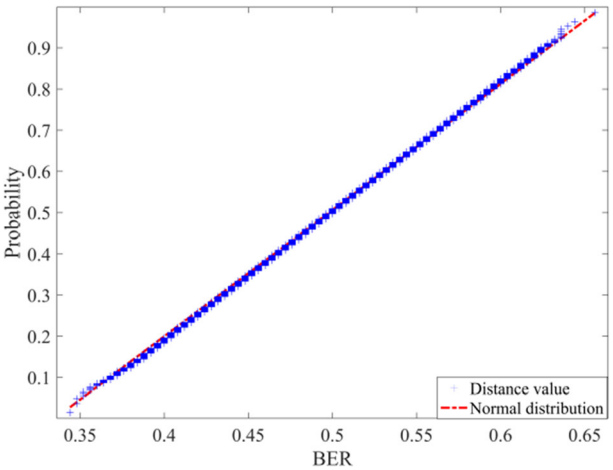

Fig. 8 BER distribution of discriminant analysis of different speech

sequence 0, 1 occurring. In this paper, $N = 250$. The theoretical normal value $\mu = 0.5$ and standard deviation $\delta = 0.0316$ are calculated, according to the De Moiver-Laplace center limit theorem. The experimental value is standard deviation $\delta = 0.0319$, $\mu = 0.4990$. It can find the experimental values μ and δ is very close the theoretical values. As shown in Fig. 8, the calculated BER distribution map of 1278 speech segments almost coincide with the standard normal distribution line, which also shows the BER distribution map follows an approximate normal distribution. It indicates our method has better discrimination.

The FAR further illustrates the differentiation of the algorithm to better and quantitatively demonstrate the discrimination of perceptual hashing in this paper. The FAR is calculated as Eq. (14):

$$\mathrm{FAR}(\tau) = \int_{\infty}^{\tau} f(x|\mu,\delta)dx = \int_{\infty}^{\tau} \frac{1}{\delta\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\delta^2}}\mathrm{dx} \tag{14}$$

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

Table 3 shows the comparison results of the FAR of the proposed algorithm and the existing algorithms.

The FAR value is calculated by Eq. (14). Under the condition of a certain matching threshold, the smaller the FAR, the lower the misrecognition rate of the authentication system. As shown in Table 3, the FAR value of the proposed perceptual hashing is much smaller than the calculated FAR of Ref. [2, 4, 12, 33, 34], which leads to a greatly reduced probability of judging different speech segments as the same speech. It indicates that our method has better discrimination, and is more suitable for a speech authentication system based on perceptual hashing. In this paper, the sampling frequency is 16 kHz and the number of frames is 360, hence, the FAR value decreases because the performance of the MDCT-NMF algorithm proposed in [2] depended on the number of frames. In Ref. [12], the linear prediction analysis is used to approximate with the minimum mean square error. There is a certain error, which makes the FAR value decrease. Ref. [33] uses an improved spectral entropy method to construct perceptual hashing, which affects the size of its spectral entropy value in the noise section. It leads to cannot accurately represent the speech signal, so the FAR value reduced. In Ref. [4], speech syllables was used to construct the perceptual hashing, which will cause misjudgment when distinguishing between noise and speech segments, it causes the FAR value decreases. Ref. [34] 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 this study, only 2.21 speech fragments are misjudged for every 10^{10} speech fragments when matching threshold is $\tau = 0.3$, which is lower than the comparison algorithm.

Table 3 Comparison results of FAR values under different thresholds

τ	Our method	Ref. [2]	Ref. [12]	Ref. [33]	Ref. [4]	Ref. [34]
0.1	3.38×10^{-36}	2.94×10^{-21}	2.95×10^{-22}	7.75×10^{-32}	1.59×10^{-29}	5.25×10^{-27}
0.15	3.69×10^{-28}	1.14×10^{-16}	3.42×10^{-17}	9.49×10^{-25}	5.56×10^{-23}	4.11×10^{-21}
0.20	3.52×10^{-21}	1.11×10^{-12}	1.38×10^{-13}	1.33×10^{-13}	2.63×10^{-17}	5.45×10^{-16}
0.25	2.96×10^{-15}	2.75×10^{-09}	1.94×10^{-10}	2.15×10^{-13}	1.69×10^{-12}	1.23×10^{-11}
0.30	2.21×10^{-10}	1.68×10^{-06}	9.69×10^{-07}	4.05×10^{-09}	1.51×10^{-08}	4.77×10^{-08}

5.3 Robustness performance analysis

During the daily use of speech, some conventional content preserving operation (CPO) processing may be performed on the speech, such as volume increase and decrease, resampling, and compression et al. In order to test the robustness of the proposed perceptual hashing algorithm, eight conventional CPO operations are selected to test the robustness of the perceptual hashing algorithm under CPO. Table 4 shows the CPO operations.

Table 5 shows the comparison results of the robustness between the Ref. [2, 4, 12, 33, 34] and the proposed method.

As can be seen from Table 5, the proposed perceptual hashing algorithm is less robust in the encrypted domain than in the original domain. It certificates that encryption has affected the speech extraction features. Compared with the Ref. [2], our method is more robust in the operation of Gaussian white noise. It is slightly worse in volume increase, decrease, compression and resampling, but it is better than the Ref. [2] in the original domain. Compared with Ref. [12], it is obviously better in resampling, compression, and white noise operations. The volume increase in the original domain is better than Ref. [12]. Compared with the Ref. [33], the volume decrease, compression, and white noise are more robust. And the volume is increased and worse, but it is better than the Ref. [33] in the original domain. Compared with the Ref. [4], the volume reduction operation is more robust and the resampling operation is better in the original domain, but the volume and compression operations are slightly worse. Compared with the Ref. [34], the operation of adding white noise is more robust, the volume, resampling, and compression operations are less robust, but they are better in the original domain. Because the performance of the proposed encryption algorithm is much better than the Ref. [34], this greatly affects the robustness of the algorithm. In summary, although some operations are less robust in encrypted speech, this is on account of encryption makes the speech features less and affects its robustness. Because the authentication digests in [2, 12, 33] are extracted in the original domain, the authentication abstracts extracted in our paper are directly extracted in the encrypted speech. Moreover, the operations of less robust compared with comparative literature, the difference is small. Hence, the proposed perceptual hashing in our study can meet the requirements for robustness of speech authentication.

The false reject rate (FRR) can also be used to evaluate the robustness of the hash algorithm, it refers to the probability of judging two identical contents as different contents. For evaluating the overall performance of the hashing algorithm, the FAR-FRR curve is shown in Fig. 9. The formula for calculating FRR is shown in Eq. (15).

Table 4 CPO types

Type	Parameters	Abbreviation
Volume adjustment 1	−50%	V.↓
Volume adjustment 2	+50%	V.↑
Resampling 1	16–8-16 kHz	R8–16
Resampling 2	32–8-16 kHz	R32–16
MP3 compression 1	32 kbps	M.32
MP3 compression 2	192 kbps	M.192
Add white noise 1	Add 5 dB narrowband Gaussian white noise	W.N1
Add white noise 2	Add 40 dBnarrowband Gaussian white noise	W.N2

Table 5 Comparison of average BER

Type	Our method	Application to the original speech	Ref. [2]	Ref. [12]	Ref. [33]	Ref. [4]	Ref. [34]
V.↓	0.0041	0.0107	0.0040	0.0016	0.0385	0.0042	0.0038
V.↑	0.0686	0.0198	0.0256	0.0415	0.0604	0.0039	0.0160
R.8→16	0.0068	0.0015	0.0012	0.0260	0.0032	0.0026	0.0033
R.32→16	0.0484	0.0119	0.0098	0.1219	0.0423	–	0.0223
M.32	0.0651	0.0081	0.0218	0.1147	0.2761	0.0016	0.0090
M.192	0.0192	0.0050	0.0035	0.0727	0.2600	–	0.0086
W.N1	0.0136	0.0027	0.1049	0.4257	0.2755	–	0.0964
W.N2	0.0634	0.0087	0.2633	0.4578	0.2934	–	0.1394

$$FRR(\tau) = 1 - \int_{-\infty}^{\tau} f(x|\mu,\delta) = 1 - \int_{\infty}^{\tau} \frac{1}{\delta\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\delta^2}} \mathrm{dx} \tag{15}$$

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

Figure 9 shows the FAR-FRR curve of the perceptual hashing algorithm. The X-axis is the critical value for judging correct as wrong. The Y-axis represents the probability that the correct speech is misjudged as a tampered speech.

As can be seen from Fig. 9 that the two curves of FAR and FRR do not intersect. And the proposed scheme in this study also have a large threshold decision interval, which shows it has good discrimination and robustness. Thence, the proposed method can be applied to content authentication of the encrypted speech.

5.4 Authentication efficiency analysis

For expounding the authentication efficiency of this algorithm, 100 clips 4 s speeches are selected from the speech libraries. The total time that generate a perceptual hashing sequence and hashing match is calculated to measure the authentication efficiency. Table 6 shows the comparison of algorithm authentication time of the proposed method and existing Ref. [2, 4, 12, 33–35].

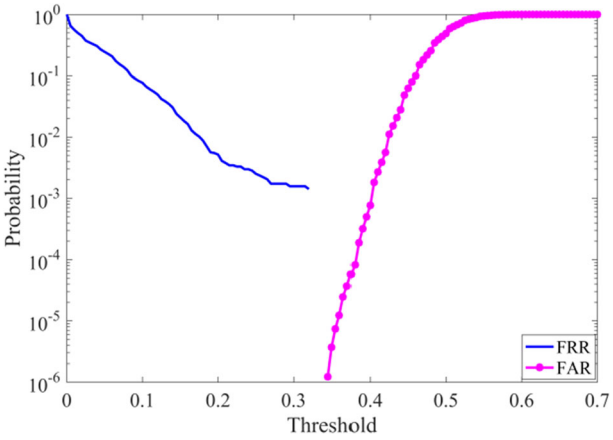

Fig. 9 The FAR-FRR curve of proposed method

Table 6 Comparison of authentication efficiency of different algorithms

Algorithm	Ref. [2]	Ref. [12]	Ref. [33]	Ref. [34]	Our method
Work frequency (GHz)	2.5GHz	3.3GHz	2.2GHz	2.2GHz	2.2GHz
Total(s)	360	360	266	250	250
Length of hashing sequence (bit)	130.4	12.47	3.78	13.84	2.08

It can be seen from Table 6 that the authentication efficiency of the proposed method is 62 times faster than Ref. [2], 6 times faster than Ref. [12], 1.8 times faster than Ref. [33], and 6 times faster than Ref. [34]. Because the algorithms in [2, 12, 34] adopted the NMF dimensionality reduction method for data dimensionality reduction processing, which had resulted in a significant decrease for the efficiency of matching authentication. Through the above analysis, it is shown that the proposed method has high authentication efficiency. Additionally, the digest length is 250, which has good abstractness. Therefore, the algorithm designed in this paper is very simple and easy to implement, which is suitable for the real-time authentication needs in practical applications.

5.5 Tampering location and tampering recovery

The encrypted speech needs to be located and recovered when the speech authentication fails, so as to make up for the loss of the user as much as possible. In order to explain the proposed method's ability of tampering location, two kinds of forgeries are performed for speech. Then, the SNR and SegSNR are used to evaluate tampered speech recovery quality. Figures 10 and 11 show the results of tampering location and recovery for mute and replace attacks.

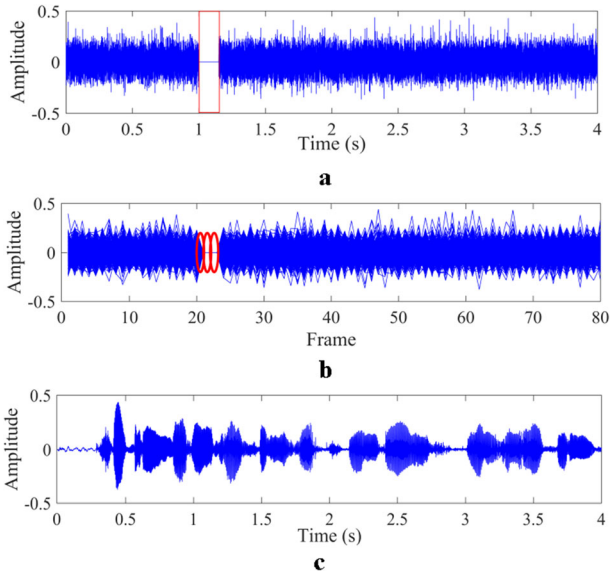

Fig. 10 Tamper location and recovery results from mute attacks: (**a**) encrypted speech under mute attack (**b**) the location of tampered speech frames (**c**) the decrypted and recovered speech subjected to mute attacks

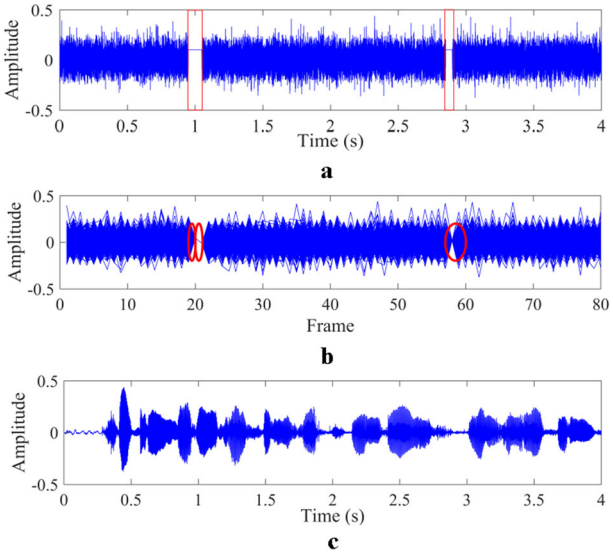

Fig. 11 Tamper location and recovery results from replacement attacks: (**a**) encrypted speech undergo multiple tamper attacks (**b**) the location of tampered speech frames (**c**) the decrypted and recovered speech subjected to substitution attacks

5.5.1 Location and recovery of mute attack

In the real attacks, the attackers often implement mute attack to reset the significant speech information to zero. In this study, the speech samples (16,000-18,400) are randomly reset to zero, as shown in Fig. 10(a). Figure 10(a), (b), and Fig. 10(c) respectively show the speech waveform of the encrypted speech after the mute attack, the tampering location of the mute attack, and the speech waveform after decryption and recovery. It can be seen from Fig. 10(b) that the proposed method can accurately locate the tampered area, there is the basic work for the speech recovery. When the marked tampered speech is decrypted by inverse scrambling diffusion, it is difficult to observe the difference between the decrypted tampered speech and the original speech on the waveform. It indicats the tampered samples are diffused into the entire speech, which is beneficial to recover the tampered speech using the least square curve fitting. In Fig. 10(c), the recovered speech waveform is very similar to the original speech waveform in Fig. 6(a). It also shows that the proposed method can recover high quality tampered speech.

5.5.2 Location and recovery of substitute attacks

Many malicious attackers often employ tampering attacks to change the meaning of original speech, such as substitute. Therefore, the samples (14,400-16,000) and samples (46,400-47,200) of the encrypted speech are replaced, as shown in Fig. 11(a). Figure 11(a), (b), and (c) respectively show the speech waveform of the encrypted speech after multiple replacement attacks, the localization digram of replacement attacks, and the speech waveform after decryption and recovery. As can be seen from Fig. 11(b), the authentication method can accurately locates multiple tampered position. In Fig. 11(c), the recovered speech waveform is

very similar to the original decrypted speech waveform in Fig. 6(a), which shows that the recovery method in this paper can recover the tampered samples value to the greatest extent.

5.5.3 Tamper recovery performance analysis

As can be seen from Figs. 10 and 11, the proposed method can precisely locates mute and substitute two malicious attacks. When the proposed method locates the tampered position of the encrypted speech, the speech users mark these tampered samples. Then, the marked speech signal is diffused inversely to obtain tampered decrypted speech. This paper uses least curve fitting to recover the tampered samples for recovering the tampered samples with high quality. If the entire speech as the input is used to find the optimal solution, it may occur overfitting. This is resulting in poor recovery speech quality, because the amplitude of the speech samples fluctuate differently in different speech segments. Therefore, this study believes that speech fragments with as consistent speech fluctuations as possible are input, which makes the obtained fitting coefficients optimal. In the experiment, this paper selects a frame as a segment, and obtains the average value of untampered neighboring samples as the input of the least square curve fitting to find the optimal solution of the curve fitting function. Thus, using the optimal solution to recover the tampered samples with high quality.

Figures 10(c) and 11(c) show the speech waveforms after decryption and recovery of the mute and substitute attacks. It can be seen from Figs. 10(c) and 11(c) that the speech waveform recovered by the least square curve fitting is very similar to the decrypted original speech waveform. It demonstrates that the proposed method can expensive quality the recover tampered samples value. Next, the SNR and SegSNR are used to describe performance of the proposed recover method.

Table 7 is the recovery quality results of the tampered speech.

As can be seen from Table 6, the recovery quality of the proposed method is obviously higher than the recovery quality of the Ref. [14, 24]. This is because the least square curve fitting method reduces the error between the recover value and the original value, making the recovered samples more reasonable. Compared with the DCT-based compression reconstruction in Ref. [14], the recovery quality of this paper is higher and the recovery method is easier. Because the compression reconstruction based on the DCT sparse basis has a large amount of compressed data, it is very troublesome for the speech to carry these compressed data. This will undoubtedly cause speech distortion and speech quality degradation, which contraries to the original intention of recovering the speech signal with high quality. Compared with the Ref. [24], the proposed method further enhances the quality of speech recovery, and reduces the recovery error. It raises the SNR and the peak signal-to-noise ratio (PSNR) of the recovered speech. Based on the above analysis, these show the proposed method can be applied to the entire encrypted speech authentication model. Compared with digital watermark authentication models, the authentication steps become simpler. Moreover, the proposed method does not

Table 7 Tampered speech recovery quality results for different attacks

Attack forms	SNR			SegSNR		
	Ref. [14]	Ref. [24]	Our method	Ref. [14]	Ref. [24]	Our method
Mute attack	18.3	28.1	30.2	39.43	77.13	89.72
Substitute attack	17.5	28.2	29.8	38.38	76.81	99.18

decrease the speech quality because of embedding additional speech information, and does not need to consider embedding capacity.

Table 8 shows the comparison results between the authentication model in this study and the authentication model proposed in the comparative reference.

As can be seen from Table 8, our proposed the authentication model settles the issues of privacy protection, speech authentication, tampering location, and tampering recovery. Compared with the Ref. [2, 4, 12, 33], the perceptual features are directly extracted from the encrypted speech in this study, and the speech authentication and tampering location can be directly operated in the encrypted speech. It greatly increases the speech privacy and security in the cloud. Furthermore, the proposed perceptual hashing scheme constructed can be used for speech retrieval and content authentication. Compared with the digital watermarking scheme proposed in [14, 24], our research can complete the speech authentication, tampering location and recovery without additional information; Compared with the existing digital watermark authentication model, the speech authentication model based on perceptual hashing does not affect the quality of the speech by embedding the watermark information, which simplifies the overall authentication steps.

5.6 Discussion

In this paper, we show that the solution of encrypted speech authentication and tampering recovery. We confirm that our research method has breakthroughs in the following aspects compared with the existing research methods. Firstly, the extraction method of the authentication abstract in our scheme is innovative. The existing authentication scheme extracts the authentication abstract from the original speech, while our scheme extracts the authentication abstract from the encrypted speech. This ensures the privacy and security of speech while authenticating. Secondly, other authentication schemes based on perceptual hashing had not solved the problem of how to recover the speech after tampering. However, our plan put forward a solution to the tamper recovery problem. Finally, compared with the digital watermarking scheme, our scheme reduces the embedding and extraction steps of the digital watermark, and simplifies the complexity of the entire authentication system. The results indicate that our scheme is indeed better than the schemes compared in the article in terms of encryption performance, perceptual hashing performance, and speech recovery quality.

Although there are important discoveries revealed by these studies, there are also limitations. First, although the encryption method mentioned in our plan can ensure general speech security, it may be insufficient for some places with higher security. This is because a certain

Table 8 Comparison of performance for different methods

	Encryption	Methods	Feature extraction	Application	location	Recovery
Ref. [2]	No	Perceptual hashing	Original speech	Authentication	No	No
Ref. [12]	No	Perceptual hashing	Original speech	Authentication	No	No
Ref. [33]	No	Perceptual hashing	Original speech	Authentication	No	No
Ref. [4]	Yes	Perceptual hashing	Original speech	Retrieval	No	No
Ref. [34]	Yes	Perceptual hashing	Encrypted speech	Retrieval	No	No
Ref. [14]	No	Digital watermarking	–	Authentication	Yes	Yes
Ref. [24]	Yes	Digital watermarking	–	Authentication	Yes	Yes
Our method	Yes	Perceptual hashing	Encrypted speech	Retrieval and Authentication	Yes	Yes

encryption performance is sacrificed in exchange for the ability to extract the authentication digest directly from the encrypted speech. Second, the performance of our proposed performance of speech hashing is based on our proposed encryption method. For other encryption schemes, further demonstration is needed. Overall, our speech recovery program may not have good recovery quality for too large tampering area. Not with standing its limitation, this study does suggest a better advantage.

In summary, we have identified that we will further combine deep learning technology to research on the basis of the limitations of existing solutions to better solve the problems caused by the limitations of this solution.

6 Conclusions and future work

In specific application scenario of the encrypted speech retrieval system, the queried speeches are confronted with some threats, such as privacy leak and speech tampering. In order to solve these problems, this paper using speech perceptual hashing, an encrypted speech authentication and tampering recovery method based on perceptual hashing has been presented. Firstly, the proposed method designed a scrambling encryption algorithm based on Duffing mapping for solving the privacy and security of speech stored in the cloud; Moreover, a perceptual hashing algorithm in the encrypted speech is constructed by using the product of uniform subband frequency band variance and spectral entropy. Then, through perceptual hashing verifying the integrity of encrypted speech. Finally, a speech tampering recovery method based on the least square curve fitting method is proposed for recovering the tampered speech as much as possible. Theoretical analysis and experiments show that the proposed encryption algorithm has better encryption performance, and can directly extract the perceptual hashing sequence (authentication digest) from the encrypted speech. And the designed perceptual hashing algorithm has better discrimination and robustness, high authentication efficiency, and good abstractness. In addition, the proposed method can precisely locate the tampered area when the encrypted speech is implemented two kinds of malicious tampering and replacement, and it can recover the tampered samples with high quality by the least square curve fitting method. In short, the proposed method is suitable for speech privacy protection, efficient integrity authentication and tampering recovery in the cloud environment.

As future work, the scheme plan to enhance the robustness of encrypted speech authentication based on perceptual hashing, and research on tampering location and recovery for malicious attacks, such as speech insertion and deletion.

Acknowledgments This work is supported by the National Natural Science Foundation of China (No. 61862041, 61363078). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

1. Ali AH, George LE, Zaidan AA, Mokhtar MR (2018) High capacity, transparent and secure audio steganography model based on fractal coding and chaotic map in temporal domain. Multimed Tools Appl 77(23):31487–31516. https://doi.org/10.1007/s11042-018-6213-0

2. Chen N, Wan WG (2010) Robust speech hash function. ETRI J 32(2):345–347. https://doi.org/10.4218/etrij.10.0209.0309
3. Chen J, Zheng P, Guo J, Zhang W, Huang JW (2018) A privacy-preserving multipurpose watermarking scheme for audio authentication and protection. In 17th IEEE international conference on trust, security and privacy in computing and communications (IEEE TrustCom) / 12th IEEE international conference on big data science and engineering (IEEE BigDataSE). IEEE 86-91. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00023
4. He SF, Zhao H (2017) A retrieval algorithm of encrypted speech based on syllable-level perceptual hashing. Comput Sci Inf Syst 14(3):703–718. https://doi.org/10.2298/CSIS170112024H
5. Jawad AK, Abdullah HN, Hreshee SS (2018) Secure speech communication system based on scrambling and masking by chaotic maps. In 2018 international conference on advance of sustainable engineering and its application (ICASEA), 2018 international conference on. IEEE 7-12. https://doi.org/10.1109/ICASEA.2018.8370947
6. Jin X, Yu S, Liang Z, Chen Z, Pei J (2018) Video logo removal detection based on sparse representation. Multimed Tools Appl 77(22):29303–29322. https://doi.org/10.1007/s11042-018-5959-8
7. Jithin KC, Sankar S (2020) Colour image encryption algorithm combining, Arnold map, DNA sequence operation, and a Mandelbrot set. J Inform Secur Appl 50:102428. https://doi.org/10.1016/j.jisa.2019.102428
8. Kaur A, Dutta MK (2018) High embedding capacity and robust audio watermarking for secure transmission using tamper detection. ETRI J 40(1):133–145. https://doi.org/10.4218/etrij.2017-0092
9. Kocal OH, Yürüklü E, Dilaveroğlu E (2016) Speech steganalysis based on the delay vector variance method. Turkish J Electric Eng Comput Sci 24(5):4129–4141. https://doi.org/10.3906/elk-1411-167
10. Kumar R, Goyal R (2019) On cloud security requirements, threats, vulnerabilities and countermeasures: a survey. Comput Sci Rev 33:1–48. https://doi.org/10.1016/j.cosrev.2019.05.002
11. Kusuma EJ, Indriani OR, Sari CA, Rachmawanto EH (2017) An imperceptible LSB image hiding on edge region using DES encryption. In 2017 international conference on innovative and creative information technology (ICITech). IEEE 1-6. https://doi.org/10.1109/INNOCIT.2017.8319132
12. Li JF, Wang HX, Jing Y (2015) Audio perceptual hashing based on NMF and MDCT coefficients. Chin J Electron 24(3):579–588. https://doi.org/10.1049/cje.2015.07.024
13. Liu Z, Wang H (2014) A novel speech content authentication algorithm based on Bessel–Fourier moments. Digital Signal Process 24:197–208. https://doi.org/10.1016/j.dsp.2013.09.007
14. Liu ZH, Zhang F, Wang J, Wang HX, Huang JW (2016) Authentication and recovery algorithm for speech signal based on digital watermarking. Signal Process 123:157–166. https://doi.org/10.1016/j.sigpro.2015.10.023
15. Liu Y, Tang S, Liu R, Zhang L, Ma Z (2018) Secure and robust digital image watermarking scheme using logistic and RSA encryption. Expert Syst Appl 97:95–105. https://doi.org/10.1016/j.eswa.2017.12.003
16. Lu W, Chen Z, Li L, Cao X, Wei J, Xiong N, Dang J (2018) Watermarking based on compressive sensing for digital speech detection and recovery. Sensors 18(7):2390–1-22. https://doi.org/10.3390/s18072390
17. Luo XR, Xiang SJ (2014) Fragile audio watermarking with perfect restoration capacity based on an adapted integer transform. Wuhan Univ J Nat Sci 19(6):497–504. https://doi.org/10.1007/s11859-014-1044-y
18. Menendez-Ortiz A, Feregrino-Uribe C, Garcia-Hernandez JJ, Guzman-Zavaleta ZJ (2017) Self-recovery scheme for audio restoration after a content replacement attack. Multimed Tools Appl 76(12):14197–14224. https://doi.org/10.1007/s11042-016-3783-6
19. Menendez-Ortiz A, Feregrino-Uribe C, Garcia-Hernandez JJ (2018) Self-recovery scheme for audio restoration using auditory masking. PloS one 13(9):e0204442, 1-23. https://doi.org/10.1371/journal.pone.0204442
20. Mostafa A, Soliman NF, Abdalluh M, El-samie FE (2015) Speech encryption using two dimensional chaotic maps. In Computer Engineering Conference (ICENCO), 2015 Computer engineering conference on. IEEE 235–240. https://doi.org/10.1109/ICENCO.2015.7416354
21. Mustafa I, Abbas Z, Arif A, Javed T, Ghaffari A (2020) Stability analysis for multiple solutions of boundary layer flow towards a shrinking sheet: analytical solution by using least square method. Physica A-StatisticMechan Applications 540:123028. https://doi.org/10.1016/j.physa.2019.123028
22. Qian Q, Wang H, Shi C, Wang H (2016) An efficient content authentication scheme in encrypted speech based on integer wavelet transform. In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA). IEEE 1–8. https://doi.org/10.1109/APSIPA.2016.7820814
23. Qian Q, Wang HX, Hu Y, Zhou LN, Li JF (2016) A dual fragile watermarking scheme for speech authentication. Multimed Tools Appl 75(21):13431–13450. https://doi.org/10.1007/s11042-015-2801-4
24. Qian Q, Wang HX, Sun XM, Cui YH, Wang H, Shi CH (2018) Speech authentication and content recovery scheme for security communication and storage. Telecommun Syst 67(4):635–649. https://doi.org/10.1007/s11235-017-0360-x

25. Rihan SD, Khalid A, Osman SEF (2015) A performance comparison of encryption algorithms AES and DES. Int J Eng Res Technol (IJERT) 4(12):151–154. https://doi.org/10.1109/ICICT.2005.1598556
26. Shahadi HI (2018) Covert communication model for speech signals based on an indirect and adaptive encryption technique. Comput Electric Eng 68:425–436. https://doi.org/10.1016/j.compeleceng.2018.04.018
27. Sheela SJ, Suresh KV, Tandur D (2017) Chaos based speech encryption using modified Henon map. In 2017 second international conference on electrical, computer and communication technologies (ICECCT). 2017 international conference on. IEEE 1-7. https://doi.org/10.1109/ICECCT.2017.8117918
28. Sun F, Li Y, Liu Z, Qi C (2019) Speech forensics based on sample correlation degree. Advances in computer communication and computational sciences. Springer, Singapore, pp 173–183. https://doi.org/10.1007/978-981-13-0344-9_14
29. Wang W, Hu GM, Yang L, Huang DF, Zhou Y (2016) Research of endpoint detection based on spectral subtraction and uniform sub-band spectrum variance. Audio Eng 40(5):40–43. https://doi.org/10.16311/j.audioe.2016.05.09
30. Wu XZ, Xia LX, Zhang X, Zhou C (2019) Voice activity detection method based on MFPH. J Beijing Univ Posts Telecommun 42(2):83–89. https://doi.org/10.13190/j.jbupt.2018-228
31. Xiang S, He J (2017) Database authentication watermarking scheme in encrypted domain. IET Inf Secur 12(1):42–51. https://doi.org/10.1049/iet-ifs.2017.0092
32. Yang WX, Tang SY, Li MQ, Zhou BB, Jiang YJ (2018) Markov bidirectional transfer matrix for detecting LSB speech steganography with low embedding rates. Multimed Tools Appl 77(14):17937–17952. https://doi.org/10.1007/s11042-017-5505-0
33. Zhang QY, Hu WJ, Huang YB, Qiao SB (2018) An efficient perceptual hashing based on improved spectral entropy for speech authentication. Multimed Tools Appl 77(2):1555–1581. https://doi.org/10.1007/s11042-017-4381-y
34. Zhang QY, Zhou L, Zhang T, Zhang DH (2019) A retrieval algorithm of encrypted speech based on short-term cross-correlation and perceptual hashing. Multimed Tools Appl 78(13):17825–17846. https://doi.org/10.1007/s11042-019-7180-9
35. Zhao H, He SF (2016) A retrieval algorithm for encrypted speech based on perceptual hashing. In IEEE 2016 natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD). IEEE 1840-1845. 10.1109 / FSKD.2016.7603458

Affiliations

Qiu-yu Zhang[1] · **Deng-hai Zhang**[1] · **Fu-jiu Xu**[1]

Deng-hai Zhang
zdhler@sina.com

Fu-jiu Xu
xufujiu@163.com

[1] School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China