A Hyperchaotic Encrypted Speech Perceptual Hashing Retrieval Algorithm Based on 2D-Gabor Transform

Yi-bo Huang[1], Shi-hong Wang[1], Yong Wang[1], Yuan Zhang[1], and Qiu-yu Zhang[2]
(Corresponding author: Yi-bo Huang)
College of Physics and Electronic Engineering, Northwest Normal University[1]
967 Anning E Rd, Anning District, Lanzhou 730070, Gansu, China
School of Computers and Communication, Lanzhou University of Technology[2]
Lanzhou 730050, China
Email: huang_yibo@nwnu.edu.cn
(Received July 15, 2020; Revised and Accepted Apr. 24, 2021; First Online Aug. 17, 2021)

Abstract

A hyperchaotic encrypted speech perceptual hashing retrieval algorithm based on 2D-Gabor transform and PCA dimension reduction has been proposed in this paper. The proposed algorithm first uses 2D-Gabor transform to extract speech features. Then use PCA to reduce the dimension of the extracted feature. Finally, the proposed algorithm uses the extracted feature to construct a hash sequence, then uploads the hash sequence to the cloud to establish a hash sequence table. At the same time, use the four-dimensional hyperchaotic encryption method to encrypt the spech, and then upload it to the cloud and establishes a phonetic table with a one-to-one correspondence with the hash sequence table. When the user needs to retrieval speech, compare the generated hash sequence of target speech with the hash sequence table. After the matching is completed, the speech corresponding to the matching result is returned to the user. Thus, the proposed retrieval method can achieve successful matching without downloading and decrypting speech. Experimental results show the proposed algorithm in this paper improves the algorithm's accuracy compared with the previous retrieval algorithm and has high discrimination and nice robustness.

Keywords: 2D-Gabor Feature Extraction; Encrypted Speech Retrieval; Four-Dimensional Hyperchaotic System; Perceptual Hashing; Principal Components Analysis

1 Introduction

With the continuous development of multimedia technology, people put higher requirements in speech storage and retrieval, so how to effectively retrieval target speech from massive speech in the cloud has become a challenging tasks [3, 6]. However, the third-party cloud services are not a place you can rest assured. Therefore, it is necessary to strengthen the security of speech during transmission [9,22,24] under the premise of continuously improving retrieval accuracy and speed. However, there is not much research now about how to reteieval in encrypted speech. Therefore, how to retrieve encrypted speech has become an important field in speech retrieval research.

The current mainstream speech retrieval scheme is content-based encryption speech retrieval technology [2, 5, 12]. The content-based speech retrieval scheme realizes the retrieval of speech through the physical characteristics of speech, such as speech amplitude, speech spectrum and other characteristics, thereby greatly improving the discrimination and robustness of speech features. The perceptual hash sequence calculated by this way can have a high accuracy during retrieval. The content-based speech retrieval solution can also retrieve encrypted speech without downloading and decrypting. Therefore, this scheme not only ensures the security of speech data but also improves the efficiency and accuracy of retrieval.

Content-based encrypted speech retrieval technology mainly includes three aspects: Speech feature extraction, Speech encryption technology, Speech retrieval technology. The main speech feature extraction method now includes speech fingerprints [15, 17] and perceptual hash [1, 4], *etc.* Speech encryption technology mainly includes chaotic map encryption [14], DNA encoding encryption [11] and Haar Transform and Permutation encryption [13], *etc.* Speech retrieval technology mainly includes feature matching [20], example speech search [8], *etc.*

In 2013, Wang *et al.* [18] proposed a retrieval method for encrypted speech using perceptual hashing. It encrypted speech with Chus's chaotic circuit and piecewise linear(PWL), and used the speech zero-crossing rate to

extract speech features for retrieval. This scheme has good robustness and high retrieval speed, but due to poor discrimination, the retrieval accuracy is low. He *et al.* [6] proposed a perceptual hashing based on syllable-level to encrypt speech. Although this scheme has nice retrieval speed and improves the retrieval efficiency, but the characteristics of the algorithm are too simple, the discrimination is not high. Zhang *et al.* [23] proposed a perceptual hashing based on short-time zero-crossing rate to retrieve encrypted speech. Although it has good retrieval accuracy and robustness, the algorithm has low discrimination and the encryption effect is not enough. Zhang *et al.* [21] proposed a perceptual hashing based on IFFT and measurement matrix to retrieve encrypted speech. Although this scheme has better encryption and discrimination, the robustness and retrieval accuracy of this algorithm are not high. Zhang *et al.* [19] proposed a perceptual hashing based on Chirp-Z and second feature extraction to retrieve encrypted speech. Although this algorithm has good discrimination and retrieval efficiency, the robustness of the algorithm is still not high, and the accuracy of retrieval is not enough.

These studies show that there are still many problems and shortcomings in the current encrypted speech retrieval scheme, such as discrimination and security are not enough. In order to solve these problems, we propose a hyperchaotic encrypted speech perceptual hashing retrieval algorithm based on 2D-Gabor transform and PCA dimension reduction. The proposed algorithm first use Gabor transform to extract speech feature, then use PCA on extracted feature to reduce the dimensional. At the same time, the algorithm use four-dimensional hyperchaotic to encrypt the speech files. After that, the algorithm operate the dimensional-reduced feature to generate perceptual hash sequence, and store them in the hash sequence table. In the retrieval process, the algorithm first extract the feature of the speech to be retrieved and generate a perceptual hash, then match it with the hash sequence table, finally return the matching result to the user.

The main contributions of our approach can be summarized as follow:

1) The Gabor feature extraction used in this paper can effectively generate a hash sequence that can well represent the feature information of speech, and has good discrimination and nice robustness;
2) This paper uses PCA to reduce a number of feature datas, which can greatly improve the efficiency of retrieval;
3) This paper uses four-dimensional hyperchaotic encryption for speech encryption, which not only has a large key space and can not easily brute-forced, but also greatly reduces the correlation between each frame of speech;
4) This paper uses Minimum code distance to tamper detection,which can not only detect that the speech has been tampered, but also accurately locate the tampered location.

The remaining part of this apper is organized as follows. Section 2 introduces the related theory, including Gabor Feature Extraction, Four-dimensional hyperchaotic map system. Section 3 describes in detail the specific implementation process of the proposed algorithm in this paper. Section 4 gives the experimental results and performance analysis as compared with other related algorithms. Finally, we conclude our paper and give the future perspectives in Section 5.

2 Related Theory

2.1 Gabor Feature Extraction

Fourier transform is a mathematical analysis method widely used in the field of signal processing, however it is mainly used to analyze stationary signal, but the characteristics of the signal in the local area cannot be processed well. Gabor transform [10] was proposed to solve this problem:

Generally for any $f(t) \in L^2(R)$, the Gabor transform defined as Equation (1):

$$G_f(b,w) = \int_{-\infty}^{+\infty} f(t)e^{-jwt}s(t-b)dt. \tag{1}$$

When the Window function $s(x)$ is Gaussian function, thus $s_a(x) = \frac{1}{2\sqrt{\pi a}} \exp\left(-\frac{x^2}{4a}\right)$, $a > 0$. So there is one-dimensional Gabor core function as Equation (2):

$$g(a,b,w,t) = e^{jwt}s_a(t-b). \tag{2}$$

Although one-dimensional Gabor transform have many improvements when dealing with local features compared with Fourier transform, but the one-dimensional Gabor transform cannot completely describe the characteristics of the signal. In order to better describe the signal characteristics, 2D-Gabor [7] was proposed to solve this problem. Expanding the one-dimensional Gabor core function into two-dimensional space, we can get the 2D-Gabor core function as Equation (3):

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right) * \exp\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \tag{3}$$

The real part as Equation (4):

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right) * \cos\left(2\pi\frac{x'}{\lambda}+\psi\right) \tag{4}$$

And the imaginary part as Equation (5):

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right) * \sin\left(2\pi\frac{x'}{\lambda}+\psi\right) \quad (5)$$

Among them $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$. And λ is the wavelength of sine function, θ is the direction of the core function, φ is the phase shift, σ is the Gaussian standard deviation, and γ is the aspect ratio of the two directions: x, y (That is, the ellipticity of the Gabor function is specified).

Gabor Feature Extraction is to process the signal S(x,y), but if we directly operate the obtained data, because the dimension is too high, which is not conducive to subsequent processing. So we generally block the signal first, for example: Take 16 equal divisions in the horizontal and vertical directions respectively, divide the signal into 16 × 16 sub-signal blocks, then calculate the energy corresponding to each block as Equation (6):

$$e(k) = \sum_{i=1}^{16}\sum_{j=1}^{16}|a(k)|^2; k = 1,2,\cdots,64 \quad (6)$$

Finally, we can get the frequency energy matrix E.

2.2 Four-Dimensional Hyperchaotic System

A currently accepted definition of Chaotic is Li-Yorke chaotic [16], the defination is as follow:

If there is a closed interval I, and $f(x)$ is a continuous self-mapping on I, if the following conditions are met, it is considered chaotic:

1) Continuous self-mapping function $f(x)$ is unbounded for any period;

2) In a closed interval I there has an uncountable subset S, and meet the following conditions:

   a. For any x and y, and $x \in S$, $y \in S$, there is: $\liminf_{n\to\infty}|f^n(x) - f^n(y)| = 0$;

   b. For any x and y, and $x \in S$, $y \in S$, and satisfy $x \neq y$ there is: $\limsup_{n\to\infty}|f^n(x) - f^n(y)| > 0$;

   c. For any x and y, and$x \in S$, $y \in S$, among them y is any period of $f(x)$, there is: $\limsup_{n\to\infty}|f^n(x) - f^n(y)| < 0$.

However, the dynamic equations of low-dimensional chaotic system are too simple, and the key sensitivity is not high enough. Therefore, this paper proposes a new four-dimensional hyperchaotic system as Equation (7):

$$\begin{cases} x_1(n+1) = \frac{2\alpha_1\sin(\beta_1 x_1(n))}{\gamma_1\sin(x_4(n))^2+w_1} \\ x_2(n+1) = \frac{2\alpha_2\sin(\beta_2 x_2(n))}{\gamma_2\sin(x_1(n))^2+w_2} \\ x_3(n+1) = \frac{2\alpha_3\sin(\beta_3 x_3(n))}{\gamma_3\sin(x_2(n))^2+w_3} \\ x_4(n+1) = \frac{2\alpha_4\sin(\beta_4 x_4(n))}{\gamma_4\sin(x_3(n))^2+w_4} \end{cases} \quad (7)$$

Among them $x(n)$ represent the chaotic sequence, $\alpha_i, \beta_i, \gamma_i, w_i$ are the parameters of chaotic sequence, and satisfy $\alpha_i, \beta_i, \gamma_i, w_i \neq 0$, $i = 1,2,3,4$.

The Lyapunov exponent is a quantitative description of chaotic system, which reflects the overall effect of the movement trajectories generated by the nonlinear mapping being close to or separated from each other, and describes the sensitivity of the system to the initial value when the parameters change in the chaotic motion system and the local instability in changing process. Therefore, having positive Lyapunov exponent can be used as the basis for discriminating chaotic system. The formula of Lyapunov exponent as Equation (8):

$$\lambda = \lim_{n\to\infty}\frac{1}{n}\sum_{n=0}^{n-1}\ln\left|\frac{df(x_n,\mu)}{dx}\right| \quad (8)$$

For a low-dimensional chaotic systems, at least there have one positive Lyapunov exponent, but for a high-dimensional hyperchaotic system, there must be at least two positive Lyapunov exponents, so the behavior is more complicated than general chaotic system, making it more difficult to predict.

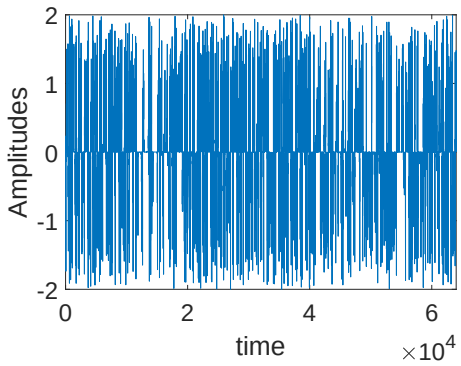

Figure 1: The waveform of ecryption system

According to Equation (8), Equation (7) have two positive Lyapunov exponents, $\lambda_1 = 2.7161$ and $\lambda_2 = 0.2445$, so this hyperchaotic system has sufficiently complex chaotic property. The time-domain waveform of hyperchaotic system as Figure 1. It can be clearly seen that the time-domain waveform of this hyperchaotic system is sufficiently complex to hide the waveform of target signal in the hyperchaotic waveform.

3 The Proposed Algorithm

3.1 System Model

The system model of the scheme mainly includes three parts: Sever terminal, Client terminal and Speech retrieval. As shown in Figure 2, we first generate perceptual hashing of all original speech files on the server side, store all the perceptual hashing sequences in the hash sequence table, then store them in the cloud. At the same time,

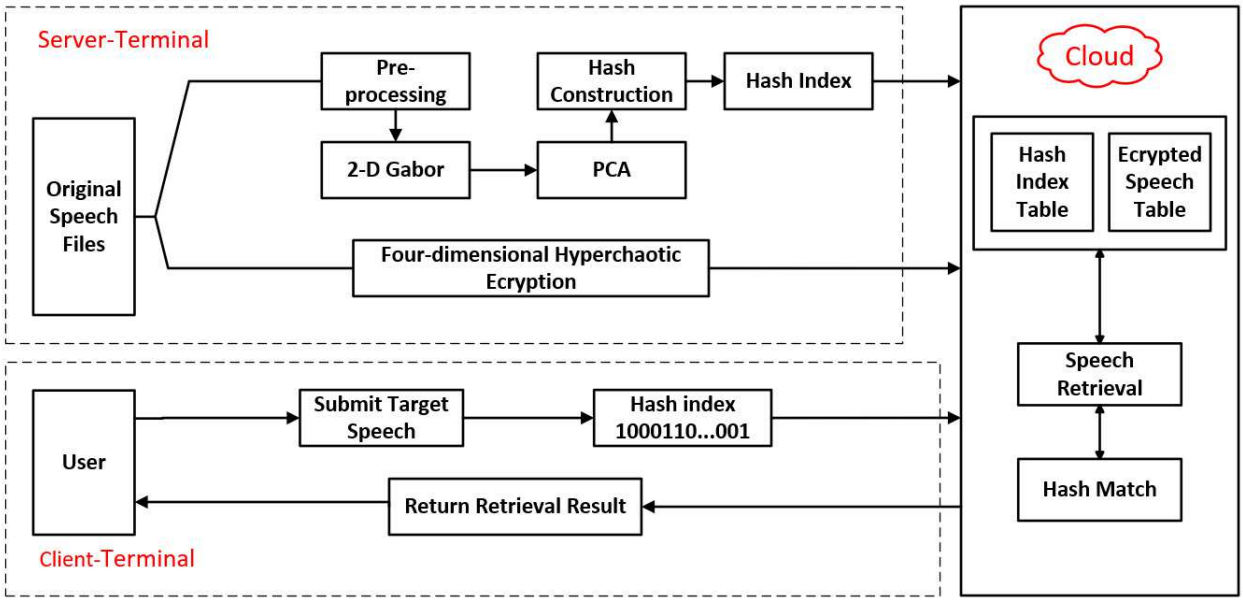

Figure 2: The flow chart of the proposed speech retrieval algorithm

we encrypt all the speech files, upload to the cloud, and realize the one-by-one mapping relationship with the corresponding speech hash sequence. When the user needs to retrieve the speech, first submit the speech to be retrieved to the client, then the system will process the retrieved speech to generate a hash sequence and upload it to the cloud. In the cloud, the system will match the hash sequence generated by the speech to be retrieved with the hash sequence table. If the same sequence is matched, the corresponding encrypted speech is found, and the result of a successful match is returned. If not found, the result of the failed match is returned to the user.

3.2 Speech Encryption Process

Assume the size of the speech to be encrypted is $1 \times m$ (The original speech has been processed into mono), the encryption steps are as follow:

1) Given the initial conditions and system parameters, repeatedly iterate Equation (7) N times, remember the result of the $N-th$ time is $X_N = [x_1(N), x_2(N), x_3(N), x_4(N)]$ (This step is to keep the initial conditions random enough to minimize human factors);

2) Use X_N as the initial conditions, then iterate m times, get the chaotic sequence: $\{H(k)|k = 1, 2, \cdots, m\}$, and satisfy $H(k) = x_1(k)$;

3) We arrange the chaotic sequence in ascending order $\boldsymbol{H} = \{h_1, h_2, h_3, \ldots, h_n\}$ to get a new sequence $\boldsymbol{K} = \{k_1, k_2, k_3, \ldots, k_j\}$, $j = 1, 2, 3, \ldots, M$;

4) Chaotic sequence $\{H(k)|k = 1, 2, \cdots, m\}$ Meet the mapping relationship with the encrypted speech E: $E(j) = H(i)$, scrambling the original speech signal according to the mapping relationship.

In the algorithm proposed in this paper, given system parameters, initial conditions as follow:

$$\text{keys} = \left\{ \begin{array}{c} \alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4 \\ \gamma_1, \gamma_2, \gamma_3, \gamma_4, w_1, w_2, w_3, w_4 \\ x_1(1), x_2(1), x_3(1), x_4(1), N \end{array} \right\}$$

Upload the encrypted speech after the above operation to the cloud.

3.3 Feature Extraction and Hash Sequence Construction

The steps of feature extraction and generating hash sequence algorithm are as follows:

1) Pre-processing: Pre-emphasis the input signal $s(t)$ to get $s(t)'$, Pre-emphasis can increase the features of the speech signal's high-frequency components. Then, the processed signal is framed, $s(t)'$ is divided into m frame, and get $f_i = \{f_i(n)|n = 1, 2, \ldots, L/m, i = 1, 2, \ldots, m\}$. L is the length of speech, m is the total number of frames, and $f_i(n)$ is the $n-th$ frame;

2) Feature extraction: According to Equation (3) to process $f(n)$ with 2D-Gabor transform, then use Equation (3) to get feature vector $\{V(k)|k = 1, 2, \cdots, m\}$;

3) Dimensional reduction: We use PCA to reduce the feature vector $\{V(k)|k = 1, 2, \cdots, m\}$ dimensional to get vector H;

4) Hash sequence generation: We use Equation (9) to get the hashing sequence:

$$\boldsymbol{h}(i) = \begin{cases} 1, & \text{if } H(i+1) > H(i) \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

where $\boldsymbol{h}(1) = 0$, thus we get the hash sequence $\boldsymbol{h} = \{h(i)|i = 1, 2, \ldots, m\}$.

4 Experimental Results and Analysis

4.1 Experimental Environment and Main Parameter Settings

We conducted a series of experiments to evaluate our approach using speech samples from the standard Texas Instrument and Massachusetts Institute of Technology (TIMIT) and Text to Speech (TTS) speech library as the test speech. The library consists of 1200 speech clips stored as 16 bits 16kHz mono recordings. The operating experimental hardware platform is Intel(R) core(TM) i5-7500 CPU @ 3.40 GHz, with memories of 4G. The operating system is window 7. And the simulation platform is MATLAB R2018b.

4.2 Performance Analysis of Perceptual Hashing

In this section, we use discrimination and robustness for evaluate the performance of the extracted speech perceptual hashing. Whether to extract the perceptual hashing sequence with good performance is the important part of speech retrieval. At the same time, we also analyzed Encryption effect and retrieval effect.

4.2.1 Discrimination Analysis

Discrimination is one of the important indicators for evaluating the speech hash sequence. The discrimination of perceptual hash is used to determine the degree of similarity between two speeches. We determine the similarity between two speechs by calculating the hamming distance between two speech hash sequences(Also named bit error rate, BER), the calculating formula of the normalized Hamming distance $D(H_{\mathrm{x}}, H_{\mathrm{q}})$ is shown in Equation (10):

$$\begin{aligned} D(H_x, H_q) &= \frac{1}{N}\sum_{p=1}^{N}(|H_x(p) - H_q(p)|) \\ &= \frac{1}{N}\sum_{p=1}^{N} H_x(p) \oplus H_q(p). \end{aligned} \tag{10}$$

Where H_{q} is the perceptual hashing sequence of query speech, H_{x} is perceptual hashing sequence in the hash sequence table, N is the length of perceptual hashing value and $p = 1, 2, \ldots, N$, setting T as the similarity threshold.

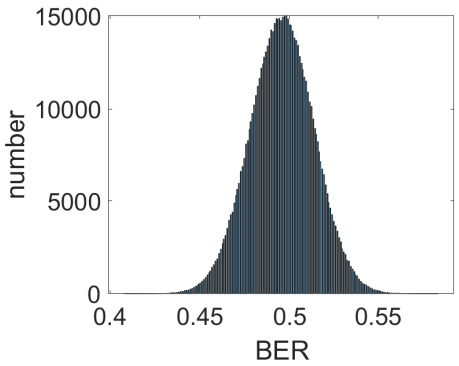

Figure 3: Statistics histogram of 1200 speech clips matching results

If $D(H_{\mathrm{x}}, H_{\mathrm{q}}) < T$, then the two corresponding hash sequences match successfully, otherwise the match is wrong, and the accuracy of retrieval is related to the threshold. The statistic histogram of BERs of the matching results is shown in Figure 3.

As shown in Figure 3, The hash BER of different speech contents basically conforms to the normal distribution. That shows the perceptual hash sequence algorithm proposed in this paper has a nice randomness and collision resistance perforamnce. The probability of the BER normal distribution is shown in Figure 4.

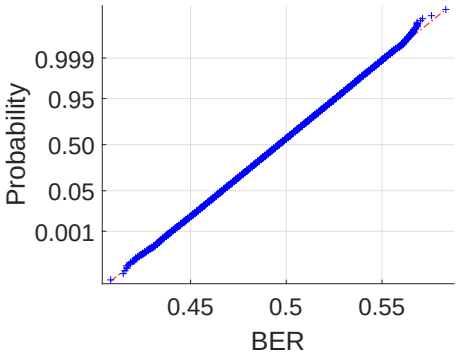

Figure 4: Probability distributions of 1200 speech clips matching results

As can be seen in Figure 4, the probability distributions of BER values of different speech basically conforms to the standard normal distribution. This indicates that the Hamming distance between different speechs is approximately normal distribution.

In order to better quantify the discrimination of the proposed algorithm, the False Accept Rate (FAR) and False Reject Rate (FRR) are mentioned. Their calculation formula are as Equation (11) and Equation (12):

$$FAR(\tau) = \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{11}$$

$$FRR(\tau) = 1 - \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \tag{12}$$

where τ is the similarity threshold, μ is the expected value, σ is the standard deviation. FAR and FRR are used to evaluate the discrimination and the robustness of the algorithm. The lower FAR means better discrimination, and the lower FRR means better robustness. According to the De Moivre-Laplace central limit theorem, the Hamming distance(also named BER) approximately obey the normal distribution ($\mu = p, \sigma = \sqrt{p(1-p)/N}$, Nis the number of bits in hashing sequence, p represents the probability of 0 or 1). In this paper, the length of the hash sequence of the speech clips is N = 1068. According to the De Moivre-Laplace central limit theorem, the mean value of normal distribution is $\mu = 0.5$, the variance is $\sigma = 0.0153$. The mean value of the experimental is $\mu_0 = 0.4960$, the variance is $\sigma_0 = 0.0179$. It can be seen from Figure 5, that the values of μ and σ measured in this paper are very close to the theoretical value. This shows that the hash sequence generated by this algorithm has high randomness and collision resistance, so as to ensure that each speech has its own unique hash sequence. In Table 1, we compare the FAR value under different

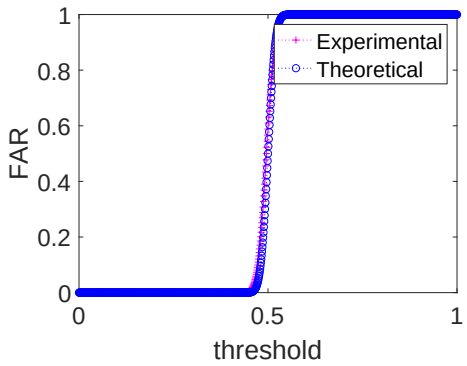

Figure 5: The FAR curve of hashing sequence

thersholds of the proposed algorithm with those existing algorithms [19–21, 23].

It can be seen from Table 1, the smaller the matching threshold τ is, the smaller the FAR value is. In the proposed algorithm, when the threshold $\tau = 0.16$ is set, about 1.3 of each 10^{80} speech clips are false accepted. This indicates that the algorithm proposed in this paper has high discrimination. When $\tau = 0.16$, about 1.8 of each 10^{-33} speech clips in [20] false accepted, about 2.3 of each 10^{-22} speech clips in [23] false accepted, about 3.3 of each 10^{-29} speech clips in [21] false accepted, about 2.5 of each 10^{-29} speech clips in [19] false accepted.

All in all, compare with the proposed algorithm in [19–21, 23], the proposed algorithm in this paper have lower FAR. What this means is that compare with the proposed algorithm in [19–21, 23], the proposed algorithm have higher discrimination.

4.2.2 Robustness Analysis

Robustness is to judge the same speech under different Content Preserving Operation (CPO) the degree of change of the speech perceptual hash sequence. The lower the robustness value, the less the extracted perceptual hash under different CPOs will be affected.

Table 2 introduces the different CPOs and their operations in proposed algorithm. Under seven kinds of CPOs, 1200 speech clips paired to compare the BER, the FRR-FAR curve is obtained in Figure 6. As can be seen from

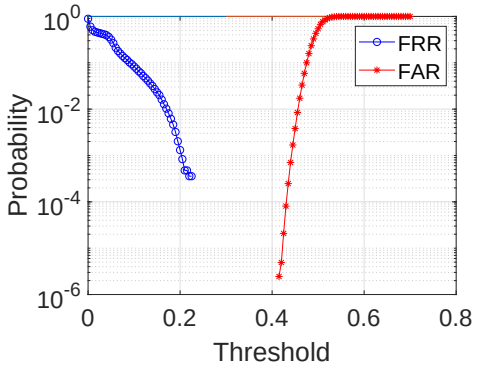

Figure 6: The FRR-FAR curve

Figure 6, there is no cross section, and the interval is still very large, indicating that the algorithm proposed in this paper has a large space to accurately determine different speech content, even after the CPOs. Obviously, when the matching threshold can be setting between 0.20 and 0.41, a good retrieval effect will be obtained. At the same time, it also shows that the algorithm proposed in this paper has both high discrimination and robustness.

Table 3 shows the BER mean value, max value and the variance of the BER value. It can be obtained from Table 3: The average BER obtained by the algorithm proposed in this paper does not exceed 0.1, and the maximum BER does not exceed 0.2. It shows that the algorithm in this paper can maintain good robustness under different CPOs. And if we do not consider the case of narrowband noise with SNR=30db, the average BER obtained in this paper does not exceed 0.05, and the maximum BER does not exceed 0.08. It shows that the algorithm in this paper can maintain nice robustness even under different CPOs. The mean BER comparison results of this algorithm is then compared with the algorithms in [19–21, 23] and the result is shown in Table 4.

As can be seen from Table 4, the average BER value of the algorithm proposed in this paper is smaller than the algorithm proposed in the [20] regardless any CPOs, which means that the algorithm in this paper is more robust than the algorithm in [20]. Our result is equal to or better than the algorithm proposed in the [19, 21, 23], indicating that the algorithm proposed in this paper is at least as good as the [19, 21, 23]. But from the previous part of this paper, we can see that the discrimination of the algorithm in this paper is much better than [19,21,23].

Table 1: Comparison of FAR values

τ	**Proposed method**	**[20]**	**[23]**	**[21]**	**[19]**
0.02	6.0115×10^{-160}	1.4486×10^{-66}	7.9324×10^{-43}	2.1743×10^{-56}	1.9366×10^{-56}
0.04	6.1132×10^{-147}	6.2274×10^{-61}	1.7718×10^{-39}	6.0859×10^{-52}	5.2700×10^{-52}
0.06	1.7184×10^{-134}	3.3854×10^{-56}	2.8523×10^{-36}	1.1039×10^{-47}	9.3200×10^{-48}
0.08	1.3354×10^{-122}	7.6358×10^{-52}	3.3141×10^{-33}	1.2978×10^{-43}	1.0713×10^{-43}
0.10	2.8699×10^{-111}	9.9134×10^{-47}	2.7782×10^{-30}	0.8909×10^{-40}	8.0059×10^{-40}
0.12	1.7058×10^{-100}	2.3020×10^{-42}	1.6830×10^{-27}	4.8877×10^{-36}	3.8902×10^{-36}
0.14	2.8053×10^{-90}	3.6436×10^{-38}	7.3382×10^{-25}	1.5665×10^{-32}	1.2295×10^{-32}
0.16	1.2768×10^{-80}	1.7923×10^{-33}	2.3147×10^{-22}	3.2571×10^{-29}	2.5281×10^{-29}

Table 2: Content preserving operation

CPO	**Operation method**	**Abbreviation**
Re-sampling	8-16kbps	R
Amplitude increase	3 db for amplitude increase	$A\uparrow$
Amplitude decrease	3 db for amplitude decrease	$A\downarrow$
Narrowband Noise 1	SNR=30db	N1
Narrowband Noise 2	SNR=50db	N2
MP3 compression	128kbps	M
Echo addition	Attenuation 50%	E

Table 3: The BER value after CPO

CPO	**Mean**	**Max**	**Variance**
R	0.0267	0.0774	4.0292×10^{-4}
$A\uparrow$	0.0010	0.0128	2.2262×10^{-6}
$A\downarrow$	0.0021	0.0088	2.3241×10^{-6}
N1	0.0941	0.2049	1.4000×10^{-3}
N2	0.0219	0.0705	2.3855×10^{-4}
M	0.0031	0.0091	3.6840×10^{-6}
E	0.0473	0.0617	2.0654×10^{-5}

It can be seen that the algorithm proposed in this paper balances the discrimination and robustness of the perceptual hash sequence, on the premise of greatly improving the discrimination of the algorithm, the robustness of the algorithm does not lost. It shows that the perceptual hash sequence in this paper will not be greatly lost in different speech environments, so it can meet the needs of speech retrieval.

4.3 Encryption Performance Analysis

We encrypted and decrypted the speech signal during transmission using the Four-dimensional hyperchaotic map system described previously. First, given the key:

$$\text{keys} = \left\{ \begin{array}{c} 1,2,3,4,2,2,4,1 \\ 3,4,2,2,4,2,2,1 \\ 3,2,5,4,500 \end{array} \right\}$$

Then we randomly select a speech from the speech library to encrypt and decrypt it, and analyze the effectiveness of the encryption algorithm proposed in this paper. Figure 7 shows the speech waveform before encryption, Figure 8 shows the the speech waveform after encryption and Figure 9 shows the speech waveform after decryption.

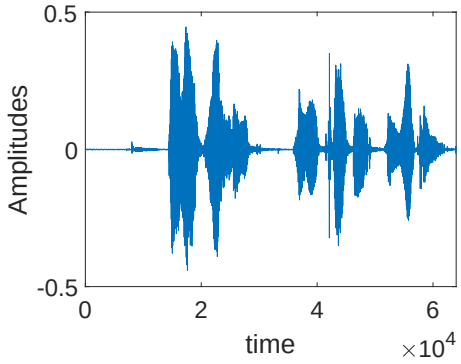

Figure 7: Original speech waveform

As we can seen from Figure 8 and Figure 9, encrypted speech has no regularity at all, and can't see any features of the original speech at all and the decrypted speech is basically the same as the original speech. For the effectiveness of an encryption algorithm, correlation analysis is also a very important criterion. For a speech clip, use Equation (13) to calculate the correlation coefficient be-

Table 4: The BER mean value of different algorithm

CPO	Proposed method	[20]	[23]	[21]	[19]
R	0.0267	0.0304	0.0033	-	0.0283
$A\uparrow$	0.0010	0.0925	0.0160	0.0052	0.0054
$A\downarrow$	0.0021	0.0089	0.0038	0.0015	0.0014
$N1$	0.0941	-	0.0248	0.0424	-
$N2$	0.0219	0.0416	-	0.0032	0.0267
M	0.0031	-	0.0090	0.2028	0.1928
E	0.0473	0.2375	-	0.1467	0.1505

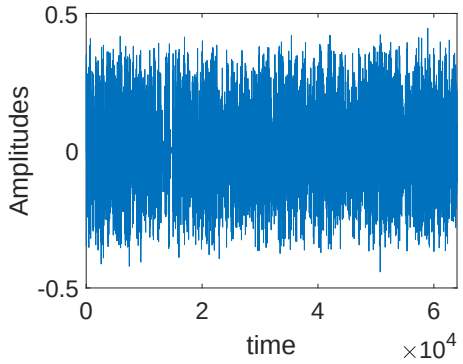

Figure 8: Encryption speech waveform

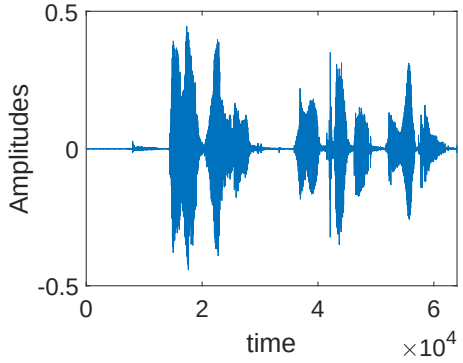

Figure 9: Decryption speech waveform

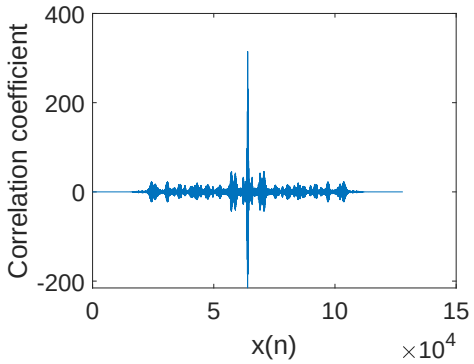

Figure 10: The correlation coefficient before encryption

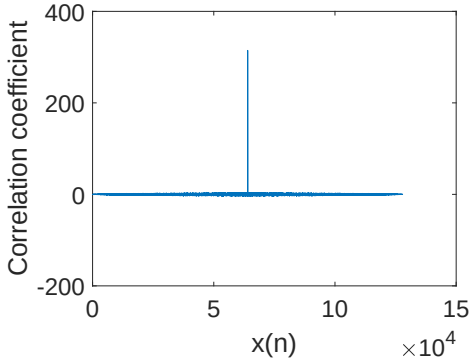

Figure 11: The correlation coefficient after encryption

tween adjacent sample points of speech.

$$\hat{R}_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n^*, & m \geq 0 \\ \hat{R}_{yx}^*(-m), & m < 0 \end{cases} \tag{13}$$

According to Equation (13), we can get the speech correlation coefficient before encryption and after encryption as Figure 10 and Figure 11. As can be seen from Figure 10 and Figure 11, the encrypted speech correlation coefficient is basically zero, indicating that the encryption effect is nice (At zero point, the speech x is exactly same as itself, and the maximum peak of the correlation coefficient appears). The original speech has a clear correlation, but the encrypted speech can't see the obvious correlation at all. This shows that the encryption algorithm proposed in this paper confuses the relevant characteristics of the original speech, so it also proves that the algorithm in this paper has high security.

In order to measure the disorder of our encryption algorithm, the position number before scrambling and the change of position number after scrambling are used to describe in this paper.

If the position number before and after the scrambling of a speech segment has not changed, there have Equation (14):

$$\Delta_i = l(i) - l'(i) = 0 \tag{14}$$

If the position number before and after the scrambling of a speech segment changes, then, there have Equa-

tion (15):

$$\Delta_i = l(i) - l'(i) \neq 0 \tag{15}$$

where, Δ_i represents the position difference.$l(i)$ represents the $i-th$ position number of the original hash sequence.$l'(i)$ represents the $i-th$ position number after scrambling. It can be seen from Figure 12 that Δ_i

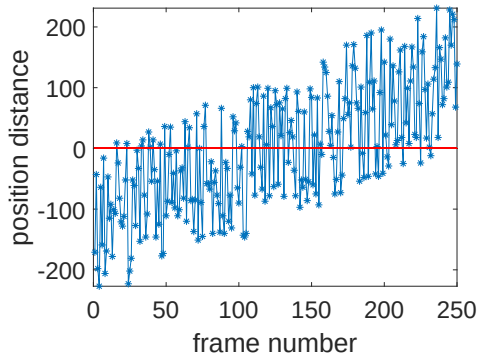

Figure 12: The intersection of Δ and the line $y = 0$

and the line $y = 0$ has few intersections, which proves that our encryption algorithm has good disorder.

4.4 Tamper Detection and Location

Aiming at the low-resolution tamper detection capability and low BER of speech segments, this paper proposes a tamper detection and localization algorithm based on minimum code distance (MCD) of Hamming code. In the detection process, for the original speech $x(n)$ and the original speech $x'(n)$ after the malicious attack, the hash sequence $h(n)$ and $h'(n)$ are obtained through the hash template. Then the MCD of the Hamming code between each frame of the two sequences is calculated. Finally, determine whether the speech has been attacked or not. defined as Equation (16):

$$MCD(i) = \begin{cases} 1, & h(i) \neq h'(i) \\ 0, & h(i) = h'(i) \end{cases} \tag{16}$$

Where, $MCD(i)$ is the MCD of the Hamming code of the $i-th$ frame, and its matrix form is as follows:

$$MCD(i) = \left[\begin{array}{cccc} MCD(1) & MCD(2) & \ldots & MCD(i) \end{array} \right]$$

As shown in Figure 13,Figure 14, the blue represents speech and red represents an area where speech content is tampered. It can be seen from Figure 15 that the algorithm can effectively detection and localization tamper. Which proves that the algorithm has good tamper detection and location capabilities for small-scale malicious attacks.

4.5 Retrieval Performance Analysis

As can be seen from above paper, according to FAR-FRR curve, we know that it is most appropriate to set the

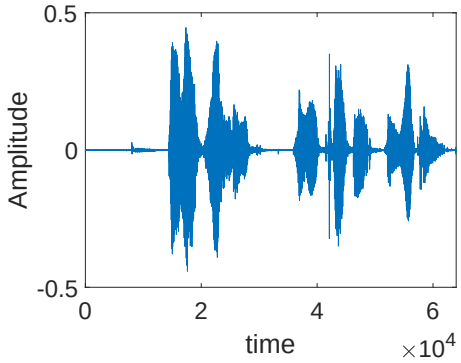

Figure 13: Original speech

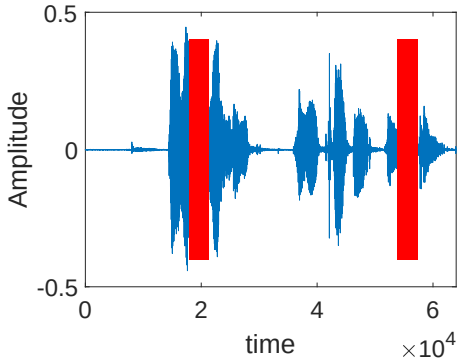

Figure 14: Tampered speech

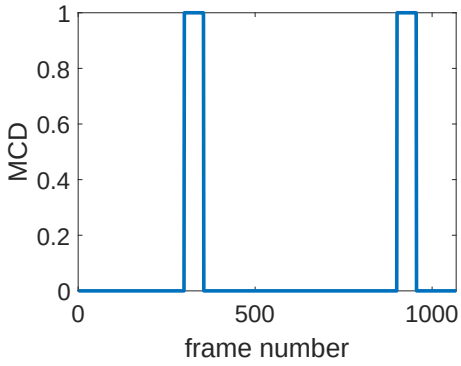

Figure 15: The minimum code distance of Hamming code

threshold between 0.25 and 0.41, so in the next experiment, we will set the threshold T=0.3, so if $D(H_x, H_q) < T$, the retrieval is successful. In order to test whether the retrieval system we proposed in this paper can accurately retrieve speech clips, we randomly select a speech clip for retrieval. The retrieval result is show in Figure 16. As show in Figure 16, our chosen threshold T=0.3 is suitable. In speech hash sequence table, except the speech to be retrieved, there is no BER value less than 0.3, which basically floats around 0.5. So the retrieval algorithm proposed in this paper is successful. The criteria for judging the performance of speech retrieval are Recall ratio (R)

Table 5: Comparison of the recall ratio after CPO

CPO	The recall ratio (%)				
	Proposed method	**[20]**	**[23]**	**[21]**	**[19]**
R	100	100	100	100	100
$A\uparrow$	100	100	100	100	100
$A\downarrow$	100	100	100	100	100
N1	100	-	100	-	-
N2	100	100	-	-	100
M	100	-	100	100	100
E	100	100	-	-	100

Table 6: Comparison of the precision ratio after CPO

CPO	The Precision ratio (%)				
	Proposed method	**[20]**	**[23]**	**[21]**	**[19]**
R	100	100	100	100	100
$A\uparrow$	100	100	100	100	100
$A\downarrow$	100	100	100	100	100
N1	100	-	100	-	-
N2	100	100	-	-	100
M	100	-	100	97	98
E	100	96	-	-	99

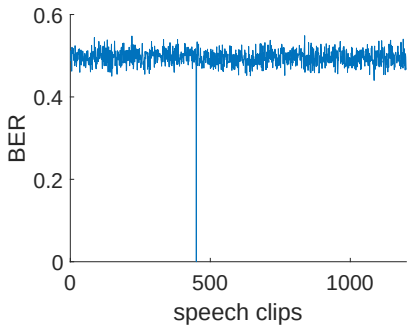

Figure 16: The matching result

and Precision ratio (P), the combination of these two criteria can describe the success rate of retrieval. The R is to describe the ability of retrieve relevant content, as show in Equation (17):

$$R = \frac{a}{a+b} \times 100\% \tag{17}$$

Where a represents the number of retrieved successfully, and b represents the number of retrieval failures. The P is to describe the ability of accurately retrieved relevant content, as show in Equation (18):

$$P = \frac{a}{a+d} \times 100\% \tag{18}$$

Where d is the number of errors retrieved.

Next, we use R and P to compare the performance of the retrieval algorithm proposed in this paper with those existing algorithms [19–21, 23]. As shown in Table 5 and Table 6, regardless of any CPOs, the recall ratio and precision ratio of the retrieval algorithm proposed in this paper can be maintained at 100% retrieval success rate, indicating that the algorithm proposed in this paper is very effective. But in [19–21, 23], due to various reasons, there will be more or less some retrieval failures.

Table 7: Efficiency comparison of different algorithms

Algorithm	**Average retrieval time**
Proposed algorithm	0.1209 s
[20]	0.1467 s
[21]	0.0649 s
[19]	0.0582 s

The efficiency of the retrieval algorithm is also an important indicator to judge the performance of the algorithm. In order to test the efficiency of our proposed retrieval algorithm, we perform a circular search on all the speechs in the speech library to ensure that each speech was retrieved once, and calculate each speech average retrieval time. The retrieval time comparison between the retrieval algorithm proposed in this paper and the algorithm in [19–21] is shown in Table 7.

As can be seen from Table 7, although the algorithm proposed in this paper has significantly improved the discrimination and robustness of the perceptual hash compared to the [19–21], but the feature extraction algorithm is too complicated, which leads to the retrieval efficiency

reduce.

5 Conclusions

In this paper, a hyperchaotic encrypted speech perceptual hashing retrieval algorithm based on 2D-Gabor transform and PCA dimension reduction has proposed. Compared with the existing algorithms, the feature extraction method of this algorithm can greatly improve the discrimination of the algorithm. Moreover, the four-dimensional hyperchaotic system reduces the correlation of speech, thereby greatly improving the security of speech in the transmission process. The experimental results show the advantages of our algorithm:

1) The generated perceptual hashing sequence has high discrimination, which can greatly improve the accuracy of retrieval.

2) The generated perceptual hashing sequence has nice robustness, which can adapt to multiple CPOs.

3) The key space of the four-dimensional hyperchaotic encryption system proposed in this paper is very large, so it can resist brute force cracking, and there is no obvious correlation between the speech that before and after encrypt, so it has high security.

The proposed algorithm has high security, retrieval accuracy and retrieval efficiency, it may be applied to cloud search, mobile speech assistant, *etc.*

The shortcoming of the proposed algorithm is when the perceptual hashing sequence is in the narrowband noise with SNR=30db, the robustness is poor and the efficiency of retrieval is not high. Therefore, improving the robustness of the perceptual hashing algorithm and further reducing the complexity of the algorithm is the next research goal.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61862041), Youth Science and Technology Found of Gansu Province of China (No.1606RJYA274).

References

[1] R. Biswas, R. A. Vasco-Carofilis, E. F. Fernandez, F. J. Martino, and P. B. Medina, "Perceptual hashing applied to tor domains recognition," *Computer Vision and Pattern Recognition*, 2020. arXiv:2005.10090.

[2] M. Blaß and R. Bader, "Content-based music retrieval and visualization system for ethnomusicological music archives," in *Computational Phonogram Archiving*, pp. 145–173, 2019.

[3] D. Dash, P. Ferrari, S. Malik, and J. Wang, "Overt speech retrieval from neuromagnetic signals using wavelets and artificial neural networks," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP'18)*, pp. 489–493, 2018.

[4] L. Du, A. T. S. Ho, and R. Cong, "Perceptual hashing for image authentication: A survey," *Signal Processing: Image Communication*, vol. 81, pp. 115713, 2020.

[5] L. Fan, "Audio example recognition and retrieval based on geometric incremental learning support vector machine system," *IEEE Access*, vol. 8, pp. 78630–78638, 2020.

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

[7] P. Huang, T. Mao, Q. Yu, Y. Cao, J. Yu, G. Zhang, and D. Hou, "Classification of water contamination developed by 2-d gabor wavelet analysis and support vector machine based on fluorescence spectroscopy," *Optics express*, vol. 27, no. 4, pp. 5461–5477, 2019.

[8] H. Kamper, A. Anastassiou, and K. Livescu, "Semantic query-by-example speech search using visual grounding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*, pp. 7120–7124, 2019.

[9] S. Kassim, O. Megherbi, H. Hamiche, S. Djennoune, and M. Bettayeb, "Speech encryption based on the synchronization of fractional-order chaotic maps," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT'19)*, pp. 1–6, 2019.

[10] J. Luo, G. Liu, Z. Huang, and S. S. Law, "Mode shape identification based on gabor transform and singular value decomposition under uncorrelated colored noise excitation," *Mechanical Systems and Signal Processing*, vol. 128, pp. 446–462, 2019.

[11] P. K. Naskar, S. Paul, D. Nandy, and A. Chaudhuri, "DNA encoding and channel shuffling for secured encryption of audio data," *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 25019–25042, 2019.

[12] G. Qian, "A music retrieval approach based on hidden markov model," in *The 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA'19)*, pp. 721–725, 2019.

[13] A. Saleh and S. B. Sadhkan, "A proposed speech scrambling based on haar transform and permutation," in *The 2nd International Conference on Engineering Technology and its Applications (IICETA'19)*, pp. 31–36, 2019.

[14] P. Sathiyamurthi and S. Ramakrishnan, "Speech encryption algorithm using FFT and 3D-lorenz–logistic chaotic map," *Multimedia Tools and Applications*, vol. 79, pp. 17817-17835, 2020.

[15] J. S. Seo, J. Kim, and H. Kim, "Audio fingerprint matching based on a power weight," *The Journal of the Acoustical Society of Korea*, vol. 38, no. 6, pp. 716–723, 2019.

[16] Y. Wang, E. Chen, and X. Zhou, "Mean li-yorke chaos for random dynamical systems," *Journal of Differential Equations*, vol. 267, no. 4, pp. 2239–2260, 2019.

[17] M. Wang, K. Li, L. Luo, X. Song, Z. Zhou, and H. Qin, "An subarea localization algorithm based on combination features using representative audio fingerprint," in *IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA'19)*, pp. 374–380, 2019.

[18] H. Wang, L. Zhou, W. Zhang, and S. Liu, "Watermarking-based perceptual hashing search over encrypted speech," in *International Workshop on Digital Watermarking*, pp. 423–434, 2013.

[19] Q. y. Zhang, Z. X. Ge, Y. J. Hu, J. Bai, and Y. B. Huang, "An encrypted speech retrieval algorithm based on chirp-z transform and perceptual hashing second feature extraction," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 6337–6361, 2020.

[20] Q. Y. Zhang, Z. X. Ge, and S. B. Qiao, "An efficient retrieval method of encrypted speech based on frequency band variance," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 9, no. 6, pp. 1452–1463, 2018.

[21] Q. Zhang, Z. Ge, L. Zhou, and Y. Zhang, "An efficient retrieval algorithm of encrypted speech based on inverse fast fourier transform and measurement matrix," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 27, no. 3, pp. 1719–1736, 2019.

[22] Q. Zhang, Y. Li, Y. Hu, and X. Zhao, "An encrypted speech retrieval method based on deep perceptual hashing and cnn-bilstm," *IEEE Access*, vol. 8, pp. 148556–148569, 2020.

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

[24] Q. Zhang, D. Zhang, and L. Zhou, "An encrypted speech authentication method based on uniform sub-band spectrum variance and perceptual hashing," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 28, no. 5, pp. 2467–2482, 2020.

Biography

Yi-bo Huang received the PhD degree from Lanzhou University of Technology in 2015, and now working as a Associate Professor in the college of physics and electronic engineering in Northwest Normal University. His research interests include multimedia information processing, information security and speech recognition.

Shi-hong Wang received the BS degree in Liren college of Yanshan University, Hebei,China, in 2018. His research interests include speech signal processing and application,

multimedia retrieval techniques.

Yong Wang received the BS degree in Henan Institute of Science and Technology, Henan, China, in 2017. His research interests include speech signal processing and application, multimedia authentication techniques.

Yuan Zhang received the B.S. degree in Wuhan Institute of Technology, Hubei, China, in 2017. His research interests include speech signal processing and application, and multimedia authentication.

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