Multimedia Tools and Applications (2020) 79:24889–24909
https://doi.org/10.1007/s11042-020-09211-y

Multi-format speech BioHashing based on spectrogram

Yi-bo Huang[1] · Yong Wang[1] · Qiu-yu Zhang[2] · Wei-zhao Zhang[1] · Man-hong Fan[1]

Received: 7 October 2019 / Revised: 30 April 2020 / Accepted: 11 June 2020 /
Published online: 26 June 2020

Abstract
In order to solve the security problem of speech perception hash authentication, the application scope of speech authentication algorithm, and improve the robustness, discrimination and real-time authentication in the process of authentication, a multi-format speech BioHashing algorithm based on spectrogram is proposed. Firstly, the speech signal to be processed is converted into spectrogram and feature extraction is carried out by two-dimensional discrete cosine transform. Then, the dimensionality of the eigenvector is reduced by non-negative matrix factorization, and generation of BioHashing sequences by inner product of reduced dimension eigenvectors and orthogonal normalized random matrices. Finally, the BioHashing is encrypted by equal-length scrambling using Henon chaotic map. The algorithm also validates the unidirectionality of BioHashing with trapdoor by comparative difference method. The experimental results show that the proposed algorithm has the characteristics of good security, strong robustness, high real-time performance and wide application range.

Keywords Speech content authentication · BioHashing · Spectrogram · Henon map · Comparative difference method

1 Introduction

In recent years, with the rapid development and increasing popularity of mobile communication technologies, data storage technologies and multimedia home collection devices, multimedia information has become the main medium for transmitting information. Authentication, retrieval, and recognition are extremely important applications in multimedia information processing. How to realize it quickly and accurately has attracted social attention and research. Speech signal is an important information carrier, how to achieve its fast authentication and ensure the security of data transmission has become a hot issue [23, 26, 43]. In recent years, with the introduction of BioHashing, 1) The key is difficult to forge

✉ Yi-bo Huang
huang_yibo@foxmail.com

[1] College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China

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

or distribute. 2) Guessing the key is very difficult. 3) The key will not be lost or forgotten, and it is also difficult to copy or share [2], which makes the BioHashing technology significantly improved in the robustness and discrimination of the traditional hash, and the security is also significantly improved. this makes BioHashing widely applied to the field of identity authentication [6, 9, 12, 13, 15].

At present, the main speech feature extraction methods based on hash include double-tree complex wavelet Transform (DT-CWT) [24, 28], frequency band variance [39], linear prediction coefficients (LPC) [20], modulated complex lapped transform (MCLT) [14], linear prediction minimum mean square error (LP-MMSE) [34, 40], discrete cosine transform (DCT) [10, 17, 18, 33, 37], discrete wavelet transform (DWT) [5, 41, 42], Mel frequency cepstral coefficients (MFCC) [3, 8], spectrogram [1, 29, 32, 45] and model [7, 44]. In [45], a high-efficiency speech identification perceptual hash algorithm based on spectrogram is proposed. The algorithm has good robustness and efficiency, but poor discrimination, and the research on the security of the hash function is lacking. Chen et al. [4] proposed a perceptual hash algorithm based on linear prediction and non-negative matrix factorization. The algorithm extracts linear prediction coefficients into perceptual features and uses non-negative matrix factorization to reduce dimensionality of the feature matrices. The algorithm has good efficiency, but it is not effective in the robustness of content preservation operation, and the security of the hash function is not involved. In [35], a multi-format speech perception hash algorithm based on dual-tree complex wavelet transform is proposed. The algorithm has high computational efficiency, can realize speech authentication of five different speech formats in the original domain and the compressed domain, but does not consider security issues and has weak robustness. Zhang et al. [36] proposed a multi-format speech perception hash algorithm based on energy to zero ratio. The algorithm has good efficiency and security, but its robustness and discrimination are poor.

Zhu et al. [46] proposed a novel unsupervised visual hashing algorithm, the so-called semantic assisted visual hashing (SAVH).The algorithm automatically extracts semantics from noisy associated text, which improves the discrimination of hash codes. Xie et al. [30] proposed online cross-modal hashing (OCMH) for fast retrieval of streaming web images. The algorithm has good efficiency. Plapous et al. [22] proposed a low-complexity audio fingerprinting technique for embedded applications.The algorithm has good robustness. Xie et al. [31] proposed an unsupervised hashing method: multi-graph cross-modal hashing (MGCMH) for large-scale multimedia search. MGCMH integrates multi-graph learning and hash function learning into a joint framework.

Jin et al. [11] proposed a BioHashing algorithm for two-factor revocable biological features based on sorting-based locally sensitive hash excitation. The robustness and security of the algorithm are relatively high. Zhang et al. [38] proposed a biometric data encryption and matching algorithm. The algorithm has good discrimination and security. Lacharme et al. [16] proposed a biological template protection algorithm. The algorithm can accurately identify the original fingerprint and the counterfeit fingerprint by reconstructing the fingerprint. Lumini et al. [21] proposed an improved BioHashing algorithm for identity authentication. The algorithm has good robustness and discrimination, but it is difficult to meet the timeliness requirements. Teoh et al. [27] proposed a BioHashing algorithm for revocable biometrics and annotations. The algorithm has good security and discrimination.

To sum up, aiming at the problem of poor security of speech perception hash authentication, the problem of small application scope and low operation efficiency of speech authentication algorithm, the robustness and discrimination of traditional hash authentication need to be improved, this paper proposes a multi-format speech BioHashing algorithm based on spectrogram. This algorithm solves the problems of low efficiency and low

security of traditional algorithms. At the same time, for common speech formats, such as WAV, MP3, FLAC, OGG, M4A, M4R, APE and AIFF. It has good robustness and discrimination. The paper also validates the unidirectionality of BioHashing with trapdoor by comparative difference method.

2 Related theory

2.1 Pre-processing

Set the speech signal is $x(n)$, after the pre-processing to obtain the i-th frame speech signal $x_i(n)$, satisfies:

$$x_i(n) = w(n) * x((i-1) * T + n) \quad 1 \le n \le L, 1 \le i \le N \tag{1}$$

Where, $w(n)$ is a window function, generally a rectangular window or a Hamming window. $x_i(n)$ is the value of the i-th frame. L is the frame length. T is the frame shift length. N is the total number of frames after framing.

When the window function is a rectangular window, its time domain and frequency domain formulas are as shown in (2) and (3):

$$w(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & else \end{cases} \tag{2}$$

$$W(e^{jw}) = W_R(w)e^{j\Theta(w)} = e^{-j(\frac{N-1}{2})w}\frac{sin\left(\frac{wN}{2}\right)}{sin\left(\frac{w}{2}\right)} \tag{3}$$

Where, $W_R(w) = \dfrac{sin\left(\frac{wN}{2}\right)}{sin\left(\frac{w}{2}\right)}$.

When the window function is a Hamming window, its time domain and frequency domain formulas are as shown in (4) and (5):

$$w(n) = \begin{cases} 0.54 - 0.46 * cos(\frac{2\pi n}{N-1}), & 0 \le n \le N-1 \\ 0, & else \end{cases} \tag{4}$$

$$W(w) = 0.54 * W_R(w) + 0.23 * \left[W_R\left(w - \frac{2\pi}{N-1}\right) + W_R\left(w + \frac{2\pi}{N-1}\right)\right] \tag{5}$$

2.2 Two-dimensional discrete cosine transform

Discrete cosine transform (DCT) has many advantages, such as abundant spectral components, concentrated energy, and does not need to estimate the speech phase. It can achieve better speech enhancement effect with low computational complexity. The horizontal ordinate of the spectrogram is time, the ordinate is frequency, and the coordinate point is the energy of speech data. It reflects the dynamic time spectrum characteristics of speech signal. It can be regarded as two-dimensional speech. As a two-dimensional spatial signal, there is also two-dimensional discrete cosine transformation (2D-DCT), which is defined as (6) and inverse transformation as (7).

$$F(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)cos\left(\frac{(2x+1)u\pi}{2M}\right)cos\left(\frac{(2y+1)v\pi}{2N}\right) \tag{6}$$

$$f(x,y)=\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\alpha(u)\alpha(v)F(u,v)cos\left(\frac{(2x+1)u\pi}{2M}\right)cos\left(\frac{(2y+1)v\pi}{2N}\right) \tag{7}$$

$$\alpha(u)=\begin{cases}\frac{1}{\sqrt{M}}, & u=0\\ \sqrt{\frac{2}{M}}, & u\neq 0\end{cases} \tag{8}$$

$$\alpha(v)=\begin{cases}\frac{1}{\sqrt{N}}, & v=0\\ \sqrt{\frac{2}{N}}, & v\neq 0\end{cases} \tag{9}$$

Where, $F(u,v)$ is a coefficient matrix of two-dimensional discrete cosine transform, and the size of the spectrogram is $M*N$, $u=0,1\cdots M-1$, $v=0,1\cdots N-1$, $x=0,1\cdots M-1$, $v=0,1\cdots N-1$. $\alpha(u)$, $\alpha(v)$ are a conversion constant.

2.3 Non-negative matrix factorization

In the process of decomposing the matrix, non-negative matrix factorization can not only reduce the dimension of the data, but also mine the local features of the data, and add non-negative constraints to the base matrix and the coefficient matrix, which makes the factorization result sparse. It is defined as:

$$V_{m*n}\simeq W_{m*r}H_{r*n} \tag{10}$$

Where, V_{m*n} is the matrix to be decomposed (m, n are the number of rows and columns of the matrix, respectively), W_{m*r} is the base matrix, H_{r*n} is the coefficient matrix, and r is the number of base vectors, and satisfies:

$$r\leq min(m,n) \tag{11}$$

In order to make the result of the product of W and H close to V, the Euclidean metric is used as the objective function, and its expression is:

$$\begin{cases}minf(W,H), & \frac{1}{2}||V-WH||_F^2\\ W, & H\geq 0\end{cases} \tag{12}$$

Where, $||\cdot||_F^2$ represents the Frobenius norm.

2.4 Henon map

The Henon map is a two-dimensional nonlinear discrete chaotic map [25]. It is sensitive to initial values, unpredictable, and traversed. Therefore, it is often used in encryption algorithms with low time overhead and high security [19]. It is defined as follows:

$$\begin{cases}x_{n+1}=1-ax_n^2+y_n\\ y_{n+1}=bx_n\end{cases} \tag{13}$$

Where, n represents the number of iterations. x, y represent the iteration value. a, b are two control parameters, $a\in(0,1.4)$, $0.2<b\leq 0.314$.

3 Multi-format speech BioHashing construction

The block diagram of the multi-format speech BioHashing algorithm based on the spectrogram is shown in Fig. 1:

3.1 User terminal

Step 1: Pre-processing Firstly, the input time waveform signal $x(n)$ is pre-emphasized to enhance the high frequency part, and then the pre-emphasis signal is divided into a speech signal with a frame length of L, a frame shift of T, and a total number of frames of N. Then, the speech signal is smoothed through a Hamming window to obtain a speech signal $x_i(n)$. Where, $x_i(n)$ is the n-th sample value of the i-th frame.

Step 2: Feature extraction The construction block diagram of feature extraction is shown in Fig. 2. Firstly, $x_i(n)$ extracts the biological features of speech signals, then two-dimensional discrete cosine transform is performed, and finally, $F(u, v)$ is dimensionally reduced.

a: Fast Fourier transform of $x_i(n)$ is used to get $X(n, k)$. Then the energy spectral density of short-term amplitude spectrum estimation ($|X(n, k)|$) at M is calculated according to (15). Finally, $P(n, k)$ is expressed as a two-dimensional graph composed of gray levels, which is a spectrogram.

$$X(n,k) = \sum_{m=0}^{N-1} x_n(m) e^{-j\frac{2\pi km}{N}} \tag{14}$$

$$P(n,k) = |X(n,k)|^2 \tag{15}$$

where, $0 \le k \le N-1$.

b: Perform 2D-DCT on the spectrogram to obtain $F(u, v)$.

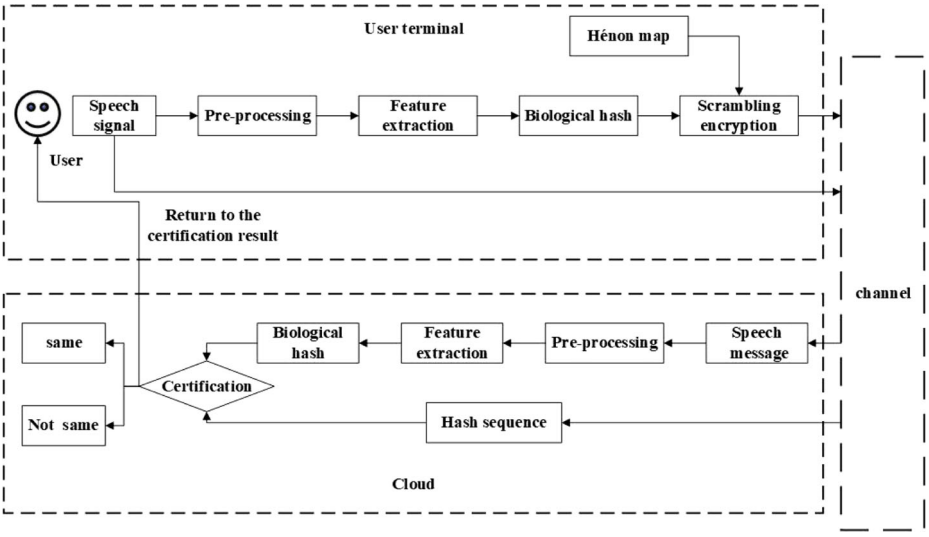

Fig. 1 Speech authentication flow chart based on BioHashing

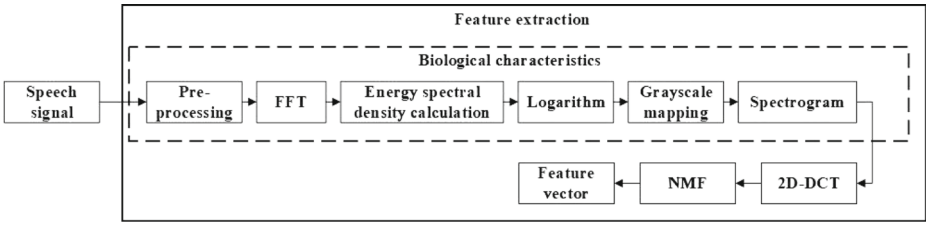

Fig. 2 Block diagram of biometrics

c: Non-negative matrix factorization of $F(u, v)$'s coefficient matrix to obtain one-dimensional feature vector $V = V((i)|i = 1, 2, \cdots, N)$.

Step 3: BioHashing construction The basic flow diagram of the BioHashing construction is shown in Fig. 3:

a: set the key k so that it produces a random sequence $s(i)(1 \leq n \leq N)$ of equal length to the feature vector V.

b: convert $s(i)$ to the standard orthogonal set $S(i)(1 \leq n \leq N)$ and perform a scalar product with the feature vector V to get $D = D((i)|i = 1, 2, \cdots, N)$.

$$D(i) = S(i) \times V(i) \tag{16}$$

c: binarize D, which is the BioHashing sequence $h = h((i)|i = 1, 2, \cdots, N)$.

$$h = [h(1) \quad h(2) \quad \cdots \quad h(i)] \tag{17}$$

Binary processing: Set the hash sequence $h(1)$ to 0. If the i-th data of the vector $D(i)$ is larger than the $(i-1)$-th data, the i-th data of the hash sequence is 1, otherwise it is 0, $i = 2, \cdots, N$.

$$h(i) = \begin{cases} 0, & h(i) \leq h(i-1) \\ 1, & else \end{cases} (i = 2, \cdots N) \tag{18}$$

Step 4: Scrambling encryption Firstly, set a = 1.4, b = 0.3, let the henon chaotic map generate a chaotic sequence $l(i)$ with the same length of $h(i)$. Then arrange $l(i)$ in descending order to get $l'(i)$, and there is a one-to-one mapping between $l'(i)$ and $h(i)$. Finally, assigning $h(i)$ to $l'(i)$ through the mapping relationship between the two, and then restoring $l'(i)$ to the unsorted state, the scrambling encryption sequence is $H(i)$ for the hash sequence.

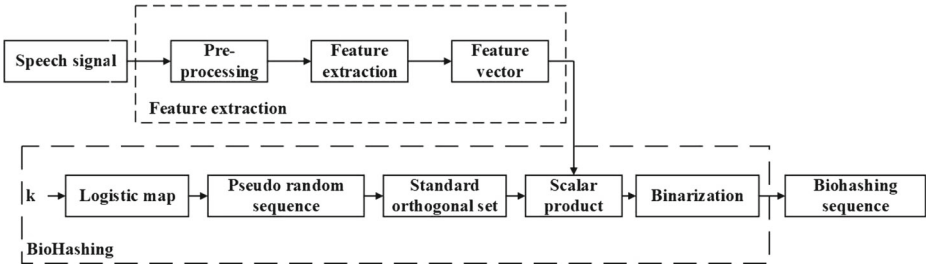

Fig. 3 Basic block diagram of a BioHashing

3.2 Cloud

Step 1: The original speech $x(n)$, according to Step 1–4 of the 3.1 user terminal, generates a scrambled encrypted BioHashing sequence H_2.

Step 2: The BioHashing sequence H_1 generated by the user terminal is hash-matched with the BioHashing sequence H_2 generated by the cloud.

In the authentication process, for the BioHashing sequences H_1 and H_2, the normalized Hamming distance $D(:,:)$ between the two can be regarded as the bit error rate (BER), and it is defined as:

$$BER = D(H_1, H_2) = \frac{1}{N}\sum_{i=1}^{N}|H_1(i) - H_2(i)| \tag{19}$$

Where, D is the bit error rate.

In order to measure the overall performance of the algorithm, this paper uses the hypothesis test of BER to describe the hash match.

P_0: If the two speech segments H_1 and H_2 are the same content, then:

$$D \leq \tau \tag{20}$$

P_1: If the two speech segments H_1 and H_2 are not the same content, then:

$$D > \tau \tag{21}$$

Where, τ is the perceived authentication threshold. When the mathematical distance is less than or equal to the authentication threshold, it is passed, otherwise, it is not passed.

Step 3: The matching result is fed back to the user.

To further measure this algorithm, this paper defines the false accept rate (FAR) and false reject rate (FRR), which are:

$$R_{FAR} = \int_{-\infty}^{\tau} f(x|\mu,\sigma)d\alpha = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}}d\alpha \tag{22}$$

$$R_{FRR} = \int_{-\infty}^{\tau} f(x|\mu,\sigma)d\alpha = 1 - \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}}d\alpha \tag{23}$$

Where, R_{FAR} represents the false accept rate. R_{FRR} represents the false reject rate. τ represents the perceived authentication threshold. μ represents mean of the BER. σ represents standard deviation of the BER.

4 Experimental results and analysis

The speech data used in the experiment are the speech signal in TIMIT (texas instruments and massachusetts institute of technology) and TTS (text to speech) speech library, and the duration is 4 s. In order to verify the generality, discrimination, security, efficiency of the algorithm, a total of nine speech libraries were established in this experiment. Eight of the speech libraries each contain a speech signal (speech libraries I, II, III, IV, V, VI, VII and $VIII$ formats are WAV, MP3, FLAC, OGG, M4A, M4R, APE and AIFF, respectively), each speech library is composed of 450 speech signals composed of Chinese men and women and English men and women, a total of 3600 speech signals. Speech library TOATL is composed of the above eight formats, a total of 3600 speech signals.

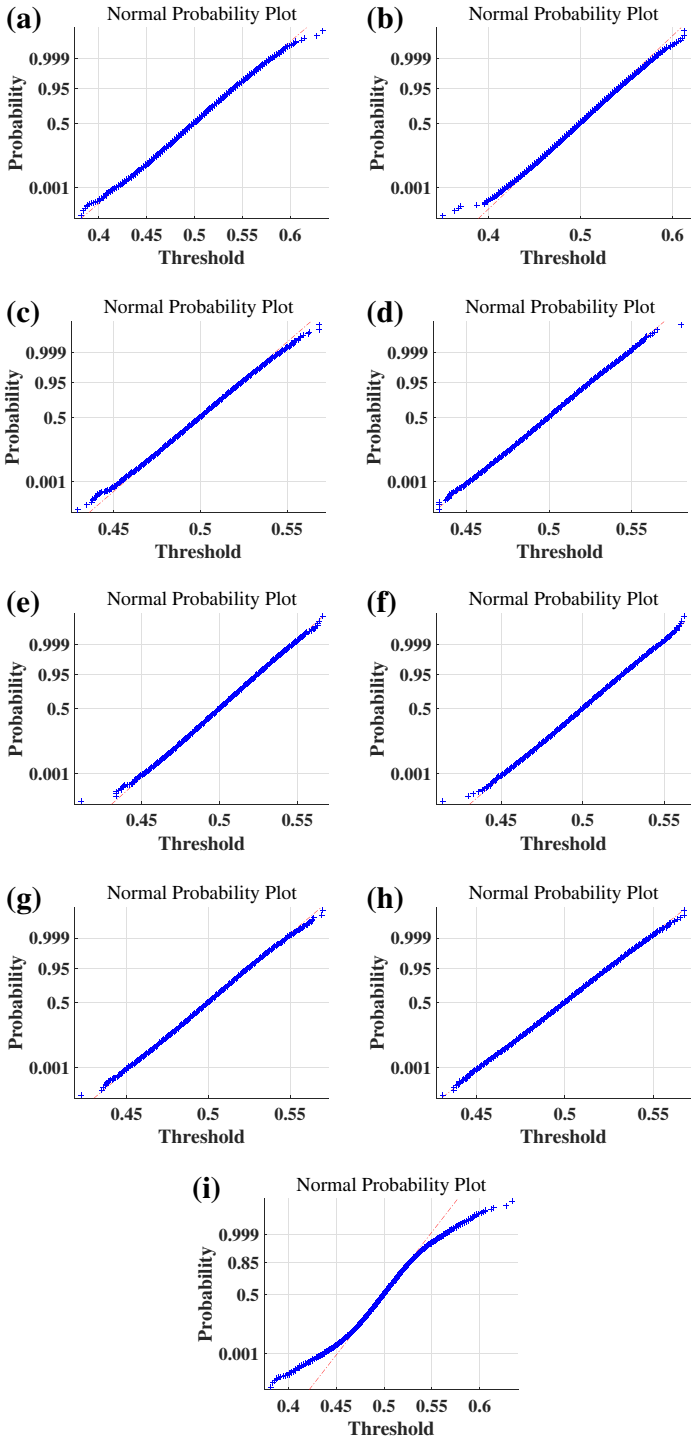

Fig. 4 BER of speech library

Table 1 Normal distribution parameter values of each speech library

Speech library	Theoretical values		Experimental values	
	μ	σ	μ	σ
I	0.5	0.0233	0.4990	0.0251
II	0.5	0.0233	0.4989	0.0250
III	0.5	0.0144	0.4996	0.0135
IV	0.5	0.0141	0.4978	0.0151
V	0.5	0.0140	0.4996	0.0150
VI	0.5	0.0140	0.4996	0.0150
VII	0.5	0.0141	0.4996	0.0151
$VIII$	0.5	0.0141	0.4995	0.0150

The experimental hardware platform is Intel(R) Core(TM) i5-7500M CPU, 3.40 GHz, 4 GB of memory, the software environment is Matlab R2018b under Windows 10 operating system. The main parameters of the experiment are as follows: frame length $L = 200$, frame shift $T = 140$ ms, window function is Hamming window.

4.1 Discrimination analysis

The BER of the BioHashing value of different content speech signals basically obeys the normal distribution. The normal distribution of each speech library data in this experiment are shown in Fig. 4:

According to the De Moivre-Laplace center limit theorem, the normalized Hamming distance of hash sequence obeys the normal distribution of $(\mu = p, \sigma = \sqrt{p(1-p)/N})$, where $p$represents the probability of occurrence of 0 or 1 in the BioHashing sequence, N represents the total number of frames. In this experiment, the normal distribution parameters of the speech libraries I-$VIII$ are shown in Table 1:

From Table 1, it can be seen that the normal distribution parameters of the proposed algorithm for eight different speech formats are very close to the theoretical values, so the algorithm has good randomness and anti-collision.

In order to verify the normal distribution parameters of the same algorithm in different total frames, the speech library I is selected for experimental analysis. The normal distribution parameters of speech library I are shown in Table 2:

It can be seen from Table 2 that in the total number of different frames of the same algorithm, the normal distribution parameter obtained by the algorithm of this paper (N = 459) is closer to the theoretical value, so the algorithm of this paper has better randomness and anti-collision.

Table 2 Normal distribution parameter values of speech library I

N	Theoretical values		Experimental values	
	μ	σ	μ	σ
$N = 459$	0.5	0.0233	0.4990	0.0251
$N = 401$	0.5	0.0250	0.4985	0.0275
$N = 357$	0.5	0.0267	0.4987	0.0295
$N = 321$	0.5	0.0279	0.4981	0.0318

Table 3 Comparison of FAR of different frame shifting algorithms

τ	T = 140 ms	T = 160 ms	T = 180 ms	T = 200 ms
0.10	2.3958×10^{-57}	1.0732×10^{-50}	1.5318×10^{-45}	1.3566×10^{-40}
0.20	4.2749×10^{-33}	2.4077×10^{-29}	1.9523×10^{-26}	1.2235×10^{-23}
0.25	1.5121×10^{-23}	6.1806×10^{-21}	6.5741×10^{-19}	5.8839×10^{-17}
0.30	1.0381×10^{-15}	4.9632×10^{-14}	1.0018×10^{-13}	1.8220×10^{-11}
0.35	1.4120×10^{-9}	1.2723×10^{-8}	7.0476×10^{-8}	3.7047×10^{-7}

It can be seen from Table 3 that when the frame shift of the algorithm is 160, 180 and 200 respectively, the FAR is about 4.4×10^6, 6.4×10^{11} and 5.7×10^{16} times when the frame shift is $T = 140$. This is because when the speech segment is preprocessed, only the internal data of the window function is weighted, and the data outside the window is set to 0. Therefore, as the frame shift is gradually increased, the overlapping frames between the speech segments are less, and the data are less, anti-collision ability also decreases.

From Table 3, It can be seen that even in the same algorithm, the FAR is different at different frame shifts, i.e. the FARs of different hash sequences obtained by different frame shifts are different in the same speech segment, so in the experiment, it is not sufficient to measure the discrimination of the algorithm only by FAR.

Entropy rate (ER) is a measure to measure the uncertainty of random events. It is also an ideal method to measure the discrimination of hash sequences of different lengths, defined as:

$$ER = -p \log_2 p - (1-p) \log_2 (1-p) \tag{24}$$

Where, $p = \frac{1}{2}(\sqrt{(\frac{\sigma^2 - \sigma_0^2}{\sigma^2 + \sigma_0^2})} + 1)$.

It can be seen from Tables 2 and 4 that the algorithm in this paper (N = 459) has better discrimination under different total frames of the same algorithm.

It can be seen from Fig. 5 and Table 5 that the curves obtained by the speech library experiments of 8 different speech formats and the theoretical curves are not only approximately coincident, but also have a high entropy rate, which further proves that compared with other algorithms, the algorithm of this paper has a better discrimination.

It can be seen from Table 6 that compared with the other three algorithms, the proposed algorithm with overlapping frames has better anti-collision capability. Among them, when $\tau = 0.1$, the number of misjudged words per 1×10^{57} speech segments is 2.3958, and under the same conditions, it is 8.9×10^9 times smaller than Ref. [36] , 1.3×10^{19} times smaller than Ref. [45], 1.5×10^{15} times smaller than Ref. [41], 3.4×10^{12} times smaller than Ref. [40].

4.2 Robustness verification and analysis

In order to test the robustness of the proposed algorithm, eight kinds of content preservation operations as shown in Table 7 are first performed for each speech library, and then the

Table 4 Entropy rate comparison of speech library I

N	$N = 459$	$N = 401$	$N = 357$	$N = 321$
ER	0.9457	0.9303	0.9271	0.9040

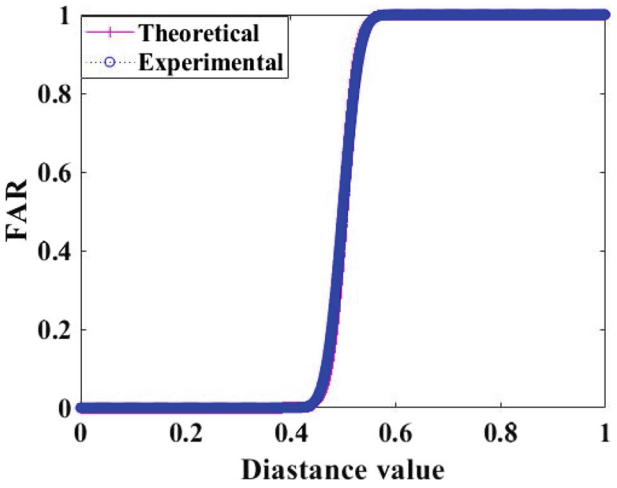

Fig. 5 FAR

average after the content preservation operation is calculated. The average of the speech libraries *I*-*VIII* is shown in Fig. 6:

It can be seen from Fig. 6 that the average values of the speech libraries *I*–*VIII* are concentrated in the interval of (0.02, 0.185), which indicates that the proposed algorithm has good robustness to multi-format speech segments. Because the spectrogram of speech is used as input, and the coordinate point value in the spectrogram represents the energy of the speech data, and the discrete cosine transform has the advantages of rich signal spectrum components and energy concentration, combining with the average BER distribution of each speech database in Fig. 6 after content preservation operation, it can be concluded that for low-pass filtering operation, because the operation filters out a part of the speech signal, in the time domain: the intensity of the part of the speech signal is reduced. In the frequency domain: it will reduce the voiceprint and lighten the color, so its robustness is the worst. For adding echo operations, adding echoes will superimpose normal speech and echo speech. In the time domain: since the amplitude of the echo is attenuated by 60%, the speech signal strength will not increase. In the frequency domain: it will increase the voiceprint and darken the color, so its robustness to content preservation operation is poor. For narrowband noise operation, the effect in the time domain is small at low signal-to-noise ratios. In the frequency domain: noise can increase voiceprint, and the color becomes

Table 5 Entropy rate comparison of different algorithms

Proposed algorithm	ER	Algorithm	ER
I	0.9457	Ref. [35]	0.8633
II	0.9487	Ref. [4]	0.6730
III	0.9530	Ref. [41]	0.9187
IV	0.9501	Ref. [37]	0.9212
V	0.9497	Ref. [33]	0.9445
VI	0.9497	Ref. [18]	0.5449
VII	0.9501	Ref. [34]	0.9745
VIII	0.9550	Ref. [40]	0.9297

Table 6 Comparison of FAR of different frame shifting algorithms

τ	Proposed	Ref. [36]	Ref. [45]	Ref. [41]	Ref. [40]
0.10	2.3958×10^{-57}	2.1420×10^{-47}	3.0310×10^{-38}	3.6540×10^{-42}	8.2137×10^{-45}
0.20	4.2749×10^{-33}	1.9220×10^{-27}	2.6890×10^{-22}	1.4500×10^{-24}	7.8089×10^{-26}
0.25	1.5121×10^{-23}	1.3754×10^{-18}	5.1740×10^{-16}	1.2150×10^{-17}	2.2030×10^{-18}
0.30	1.0381×10^{-15}	3.8423×10^{-13}	7.5420×10^{-11}	6.1660×10^{-12}	2.7579×10^{-12}
0.35	1.4120×10^{-9}	4.2761×10^{-8}	8.4900×10^{-7}	1.8740×10^{-7}	1.5648×10^{-7}

darker, so its robustness is poor. For volume adjustment operation, in the time domain: for volume increase or decrease, the intensity of speech signal will increase or decrease. In the frequency domain, the influence is small, so its robustness is better. For the resampling operation, since the operation does not affect the spectrum of the speech signal, the value does not fluctuate greatly in the time domain and the frequency domain, so it is the most robust in the content retention operation.

By comparing the eight content preservation operations of the proposed algorithm with those of Refs. [34, 36, 40, 42], Table 8 shows that the average BER of the proposed algorithm is lower overall, which further shows that the proposed algorithm has good robustness.

In the content preservation operations of the speech library: through the pairwise comparison of the BioHashing values, the speech libraries I-$VIII$ each obtained 101025 data, the speech library TOTAL obtained 6478200 data, get the FAR and FRR of each speech library, and draw the FRR-FAR curves of Fig. 7.

Figure 8a–c are the curves of Refs. [18, 36, 41]. References [18, 41] use the WAV format speech signal, Ref. [36] uses the WAV, MP3, FLAC, OGG and M4A formats.

Comparing Figs. 7 and 8, it can be seen that the $FRR - FAR$ curves of Refs. [36]. References [18, 41] have obvious intersections, Ref. [36] has no intersections, the curves of the speech libraries I–$VIII$ and the speech library TOTAL not only do not cross, and the decision interval is above 0.1, which shows that the proposed algorithm not only has good robustness, but also has a good balance between robustness and discrimination.

Table 7 Content preservation operation

Operation means	Operation method	Abbreviation
Volume adjustmean1	Volume up 50%	V. ↑
Volume adjustmean2	Volume down 50%	V. ↓
Resampling 1	Sampling frequency decrease to 8 kHz, and then increase to 16 kHz	R. ↑
Resampling 2	Sampling frequency increase to 32 kHz, and then droppsed to 16 kHz	R. ↓
Adding echo	Echo attenuation 25%, delay 300 ms	E.A
Narrowband noise	SNR = 30 dB narrowband Gaussian noise, center frequency distribution in 0–4 kHz	G.N
Low-pass filtering1	12 order Butterworth low-pass filtering, Cutoff frequency of 3.4 kHz	B.W
Low-pass filtering2	12 order FIR low-pass filtering, Cutoff frequency of 3.4 kHz	F.I.R

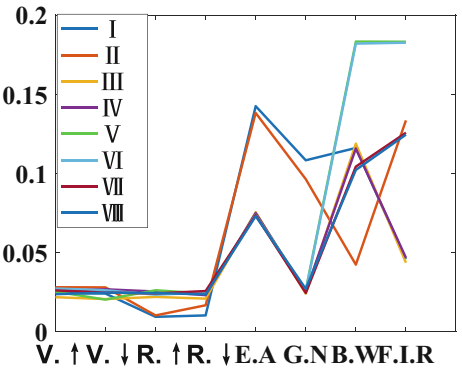

Fig. 6 Average value of content preservation operations of each speech library

In conclusion, the proposed algorithm has good robustness and discrimination.

4.3 Security analysis

In order to improve the security of authentication system in open mobile channel and semi-open cloud storage, and ensure the security of users uploading speech and hash sequence from cloud server, this paper proposes an encryption algorithm based on Henon map.

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

T_0: If the position number before and after the scrambling of a speech segment has not changed, then:

$$\Delta_i = l(i) - l'(i) = 0 \tag{25}$$

T_1: If the position number before and after the scrambling of a speech segment changes, then:

$$\Delta_i = l(i) - l'(i) \neq 0 \tag{26}$$

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 Fig. 9a that Δ_i has few intersection with the line $y = 0$, which proves that the henon mapping encryption algorithm has good disorder.

Table 8 The average BER comparison results

Operation means	Proposed	Ref. [40]	Ref. [34]	Ref. [42]	Ref. [36]
V. ↑	0.0387	0.0356	0.0455	0.0349	0.0319
V. ↓	0.0371	0.0235	0.0047	0.0102	0.0478
R. ↑	0.0380	0.0025	0.0074	0.0083	0.0703
R. ↓	0.0398	0.0394	0.0910	0.0614	0.1358
E.A	0.1005	0.0965	0.1109	0.3026	0.2087
G.N	0.0821	0.0834	0.0682	0.1150	0.1411
B.W	0.0400	0.0705	0.1248	0.1500	0.1543
F.I.R	0.1720	0.0752	0.1410	0.1754	0.1812

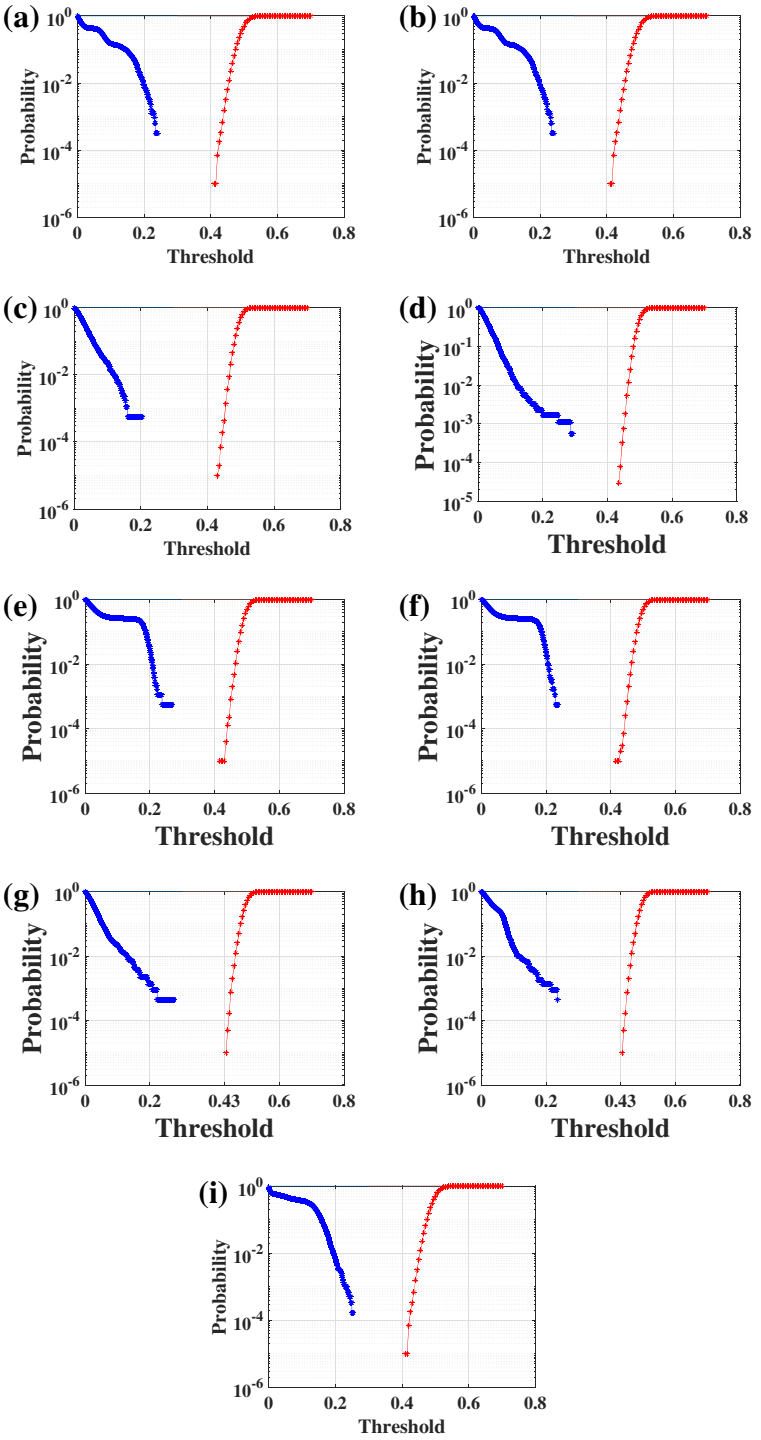

Fig. 7 FRR-FAR curves of speech library

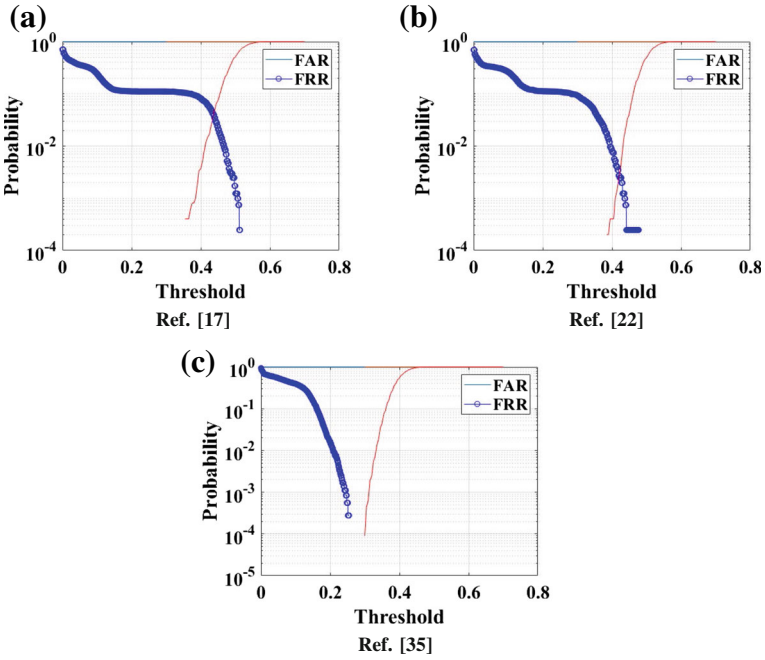

Fig. 8 FRR-FAR curves of the Ref. [18, 36, 41]

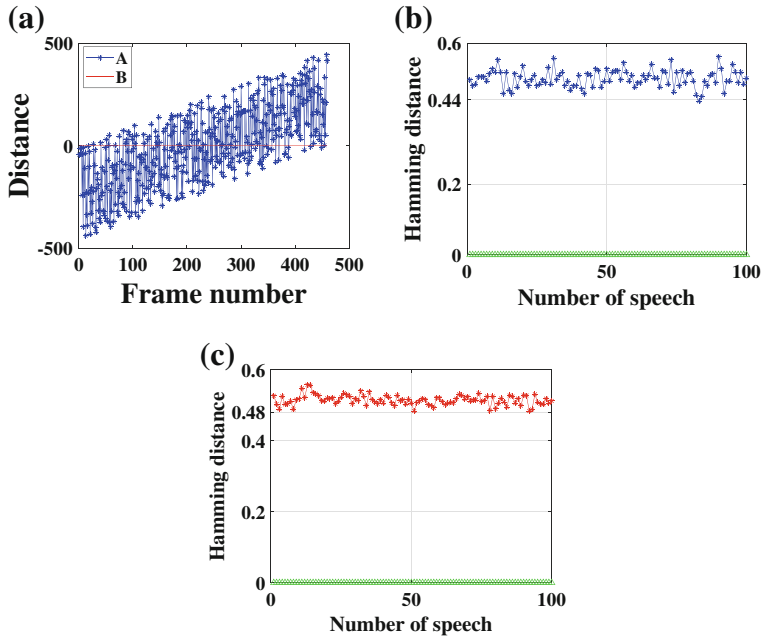

Fig. 9 **a** The intersection of Δ:difference of position number before and after scrambling and y = 0. **b** Distribution of Hamming distance before and after speech scrambling in the same group. **c** Distribution of Hamming distance between correct key and error key

In order to measure the security of the algorithm during transmission, this experiment randomly extracts 100 speeches from the speech library TOTAL. Then calculate the Hamming code distance of the unencrypted hash sequence of the same speech after two feature extractions, the hash sequence before scrambling and the Hamming code distance of the hash sequence after scrambling. Finally, the hash sequence using the correct key is calculated and the normalized Hamming code distance from the original hash sequence is used when the error key is used.

It can be seen from Fig. 9b–c that regardless of the Hamming code distance between the scrambling and the scrambling, or when using the correct key, the Hamming code distance when using the wrong key is in the upper part of the decision interval $(0.249, 0.41)$, the normalized Hamming code of the Hamming code distance of the unencrypted hash sequence of the same speech after two feature extractions is distributed on the straight line $y = 0$, the normalized Hamming code of the hash sequence using the correct key and the original hash sequence is also distributed on the line $y = 0$. Henon map based encryption algorithm proposed in this paper has good security in the transmission process.

To sum up, Henon mapping encryption algorithm increases the disorder of hash sequence by scrambling it, and then improves the security of the algorithm in the transmission process.

4.4 Unidirectional analysis

In order to verify that the BioHashing has the unidirectionality of the trapdoor, a unidirectional algorithm with trapdoor based on the Comparative Difference Method (CDM) is proposed.

For the random speech segment x, when its speech feature V obtains a BioHashing through the A direction of Fig. 10, then the speech original feature V is obtained through the B direction, and finally, the difference between the two sequences is calculated, and the CDM between the two sequences is calculated, which is defined as:

$$CDM(i) = \begin{cases} 1, & V'(i) - V(i) = 0 \\ 0, & V'(i) - V(i) \neq 0 \end{cases} \tag{27}$$

Where, CD represents the CDM of the i-th frame, and the form of the matrix is as shown in (26):

$$CDM(i) = [CDM(1) \quad CDM(2) \quad \cdots \quad CDM(i)] \tag{28}$$

As shown in Fig. 10: the A direction is a BioHashing generation. The B direction is the reverse direction of the BioHashing.

In order to verify that the BioHashing has the unidirectionality of the trapdoor, this paper randomly extracts a speech clip from the speech library TOTAL, and makes the speech first obtain the BioHashing D according to Fig. 11, Then the original feature V_1' under the correct key and the original feature V_2' under the wrong key are extracted according to Fig. 12. Finally, the CDM between the V_1' and the V under the correct key and the difference

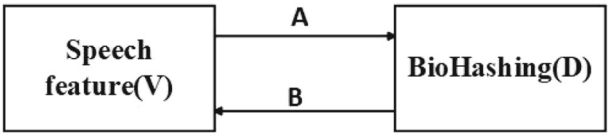

Fig. 10 BioHashing

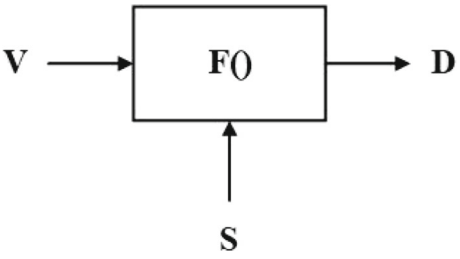

Fig. 11 A direction: BioHashing generation block diagram

between the two sequences are calculated, and the CDM between the V_2' and the V under the wrong key and the difference between the two sequences are calculated. the results of which are shown in Figs. 13 and 14.

By comparing Fig. 13a with Fig. 14a, it can be seen that the original feature V_1' extracted when using the correct key is only slightly different from the original feature V, and the distance between the two is distributed at $(-1\times10^{-17}, 1.5\times10^{-17})$. The original feature V_2' extracted when using the wrong key is completely different from the original feature V, and the distance between the two is distributed at $(-60, 0)$. This indicates that the BioHashing has unidirectional with trapdoor.

In order to further verify that the BioHashing B direction is unidirectional with trapdoors, this paper first randomly extracts 100 speeches from the speech library TOTAL, then performs BioHashing construction, and finally extracts the original features and calculates them. Calculate the Hamming code distances of V_1', V_2' and V respectively, and the Hamming code distance is as shown in Fig. 15.

It can be seen from Fig. 15 that the Hamming code distance between the original feature V and the original feature V_2' extracted using the error key is distributed between $(0.2, 0.8)$, and the original feature V and the original feature V_1' extracted using the correct key is distributed between $(2\times10^{-19}, 1\times10^{-18})$, this further proves that the BioHashing B direction is unidirectional with a trapdoor.

4.5 Efficiency analysis

In order to test the running efficiency of the algorithm, this experiment randomly extracts 100 speeches from the speech library TOTAL, the running time of the statistical algorithm under different frame shifts and the average running time of feature extraction, as shown in Tables 9 and 10, respectively:

It can be seen from Table 9: As the frame shift increases, the running time also decreases. This is because the sampling point and frame length of the speech are fixed. When the frame

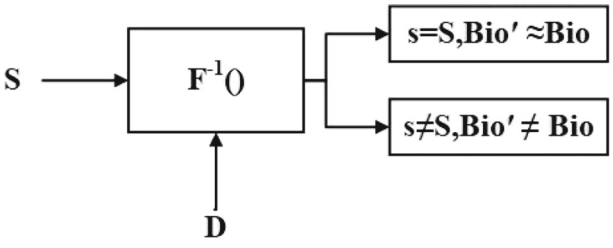

Fig. 12 B direction: extraction block diagram of original features

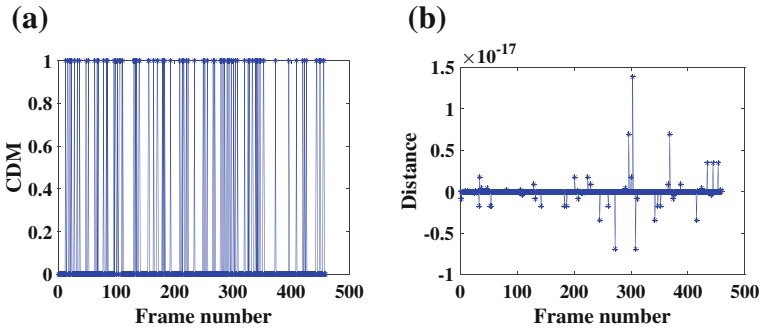

Fig. 13 Feature extraction of the correct key

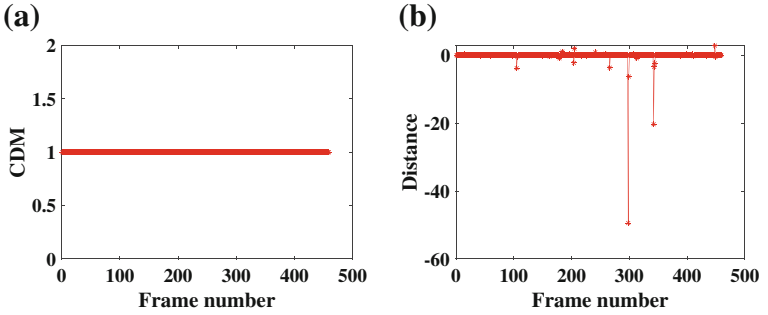

Fig. 14 Feature extraction of error key

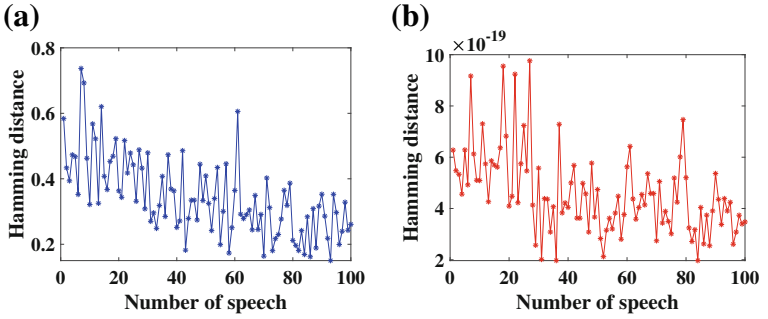

Fig. 15 Hamming code distances of V_1', V_2' and V

Table 9 Running time of different frame shifts

T	Overlap frame	N	Running time/s
140	60	459	9.75
160	40	401	8.62
180	20	357	7.73
200	0	321	6.96

Table 10 Comparing the operating efficiency of algorithms

Algorithm	Main frequency/GHz	Overlap frame	The average running time/s
Proposed	3.4	60	0.0975
Ref. [5]	3.2	0	0.5323
Ref. [41]	3.2	0	0.0848
Ref. [17]	2.5	0	0.4194
Ref. [18]	2.5	0	0.1304
Ref. [40]	2.3	0	0.0378

shift increases, the overlapping frames and the total number of frames of each speech are reduced, i.e. a large number of mathematical operations are simplified, so the running time is decreases with the frame shift increases .

From Table 10, it can be concluded that the computational efficiency of this algorithm is higher than Ref. [5, 17, 18], but lower than Ref. [40, 41]. This is due to the complexity of the structure of the algorithm itself and the large amount of computation, but the algorithm can meet the requirements of real-time communication authentication for multi-format speech.

5 Conclusions and future work

The multi-format speech BioHashing algorithm based on the spectrogram proposed in this paper solves the problem of commonality between eight speech formats of common WAV, MP3, FLAC, OGG, M4A, M4R, APE and AIFF. The experimental results show that the proposed algorithm has good robustness and discrimination, and it has a good balance between the two. This paper proves that the BioHashing has the unidirectionality of the trapdoor by the Comparative Difference Method. The encryption of the hash sequence also effectively improves the security of the algorithm authentication system in open mobile channels and semi-open cloud storage.

Because the structure of the algorithm is relatively complex, the efficiency of the algorithm needs to be improved, and the small-scale tamper detection and localization of speech can not be achieved. So the next work is to further improve the efficiency of the algorithm through the structure of the optimization algorithm, and achieve the small-scale tamper detection and localization of speech.

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

References

1. Alpar O, Krejcar O (2018) Online signature verification by spectrogram analysis. Appl Intell 48(5):1189–1199
2. Amin R, Biswas GP (2015) A secure three-factor user authentication and key agreement protocol for tmis with user anonymity. J Med Syst 39(8):78
3. Awais A, Kun S, Yue Y, Hayat S, Ahmed A, Tu T (2018) Speaker recognition using mel frequency cepstral coefficient and locality sensitive hashing. In: International conference on artificial intelligence and Big data (ICAIBD). IEEE, p 2018
4. Chen N, Wan W (2010) Robust speech hash function. ETRI J 32(2):345–347

5. Chen N, Wan W, Xiao H-D (2010) Robust audio hashing based on discrete-wavelet-transform and non-negative matrix factorisation. IET Commun 4(14):1722–1731
6. Hammad M, Luo G, Wang K (2019) Cancelable biometric authentication system based on ecg. Multimed Tools Appl 78(2):1857–1887
7. Huang Y-B, Zhang Q-Y (2017) Strong robustness hash algorithm of speech perception based on tensor decomposition model. J Softw Eng 11:22–31
8. Huang Y-B, Zhang Q-Y, Hu W-J (2018) Robust speech perception hashing authentication algorithm based on spectral subtraction and multi-feature tensor. IJ Netw Secur 20(2):206–216
9. Jiang Q, Chen Z, Li B, Shen J, Yang L, Ma J (2018) Security analysis and improvement of bio-hashing based three-factor authentication scheme for telecare medical information systems. J Ambient Intell Human Comput 9(4):1061–1073
10. Jiao Y, Ji L, Niu X (2009) Robust speech hashing for content authentication. IEEE Signal Process Lett 16(9):818–821
11. Jin Z, Hwang JY, Lai Y-L, Kim S, Teoh ABJ (2017) Ranking-based locality sensitive hashing-enabled cancelable biometrics: Index-of-max hashing. IEEE Trans Inf Forensics Secur 13(2):393–407
12. Kanak A, Sogukpinar I (2017) Biotam: a technology acceptance model for biometric authentication systems. IET Biom 6(6):457–467
13. Kaur H, Khanna P (2018) Random slope method for generation of cancelable biometric features. Pattern Recognit Lett 126:31–40
14. Kim H-G, Cho H-S, Kim JY (2016) Robust audio fingerprinting using peak-pair-based hash of non-repeating foreground audio in a real environment. Clust Comput 19(1):315–323
15. Kumari S, Li X, Wu F, Das AK, Choo K-KR, Shen J (2017) Design of a provably secure biometrics-based multi-cloud-server authentication scheme. Future Gener Comput Syst 68:320–330
16. Lacharme P (2013) Revisiting the accuracy of the biohashing algorithm on fingerprints. IET Biom 2(3):130–133
17. Li J, Wu T (2015) Perceptual audio hashing using rt and dct in wavelet domain. In: 2015 11th international conference on computational intelligence and security (CIS). IEEE, pp 363–366
18. Li J, Wang H, Jing Y (2015) Audio perceptual hashing based on nmf and mdct coefficients. Chin J Electron 24(3):579–588
19. Liu J, Li J, Ma J, Sadiq N, Bhatti UA, Ai Y (2019) A robust multi-watermarking algorithm for medical images based on dtcwt-dct and henon map. Appl Sci 9(4):700
20. Lotia P, Khan DMR (2013) Significance of complementary spectral features for speaker recognition. IJRCCT 2(8):579–588
21. Lumini A, Nanni L (2007) An improved biohashing for human authentication. Pattern Recognit 40(3):1057–1065
22. Plapous C, Berrani S-A, Besset B, Rault J-B (2018) A low-complexity audio fingerprinting technique for embedded applications. Multimed Tools Appl 77(5):5929–5948
23. Qian Q, Wang H, Sun X, Cui Y, Wang H, Shi C (2018) Speech authentication and content recovery scheme for security communication and storage. Telecommun Syst 67(4):635–649
24. Sharma M, Sharma P, Pachori RB, Gadre VM (2019) Double density dual-tree complex wavelet transform-based features for automated screening of knee-joint vibroarthrographic signals. In: Machine intelligence and signal analysis. Springer, pp 279–290
25. Sheela SJ, Suresh KV, Tandur D (2018) Image encryption based on modified henon map using hybrid chaotic shift transform. Multimed Tools Appl 77(19):25223–25251
26. Siddavatam I, Khatri D, Ashar P, Parekh V, Sharma T (2019) Authentication using dynamic question generation. In: Integrated intelligent computing, communication and security. Springer, pp 293–300
27. Teoh ABJ, Kuan YW, Lee S (2008) Cancellable biometrics and annotations on biohash. Pattern Recognit 41(6):2034–2044
28. Wang NF, Jiang DX, Yang WG (2019) Dual-tree complex wavelet transform and svd-based acceleration signals denoising and its application in fault features enhancement for wind turbine. J Vib Eng Technol 7(4):311–320
29. Wodecki J, Kruczek P, Bartkowiak A, Zimroz R, Wyłomańska A (2019) Novel method of informative frequency band selection for vibration signal using nonnegative matrix factorization of spectrogram matrix. Mech Syst Signal Process 130:585–596
30. Xie L, Shen J, Zhu L (2016) Online cross-modal hashing for web image retrieval. In: Proceedings of the thirtieth AAAI conference on artificial intelligence (AAAI-16), pp 294–300
31. Xie L, Zhu L, Chen G (2016) Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval. Multimed Tools Appl 75(15):9185–9204
32. Yenigalla P, Kumar A, Tripathi S, Singh C, Kar S, Vepa J (2018) Speech emotion recognition using spectrogram & phoneme embedding. In: Interspeech, pp 3688–3692

33. Zhang Q, Xing P, Huang Y, Dong R, Yang Z-P (2015) An efficient speech perceptual hashing authentication algorithm based on wavelet packet decomposition. J Inf Hiding Multimed Signal Process 6(2):311–322
34. Zhang Q, Hu W, Qiao S, Zhang T (2016) An efficient speech perception hash authentication algorithm based on the linear prediction minimum mean squared error. J Huazhong Univ Sci Technol (Nat Sci Edition) 44(12):127–132
35. Zhang Q-Y, Xing P-F, Huang Y-B, Dong R-H, Yang R-H (2016) Perception hashing algorithm for multi-format audio. J Beijing Univ Posts Telecommun 39(4):77–82
36. Zhang Q, Qiao S, Zhang T, Huang Y (2017) Perception hashing authentication algorithm for multi-format audio based on energy to zero ratio. J Huazhong Univ Sci Technol (Nat Sci Ed) 45(9):33–38
37. Zhang Q, Qiao S, Zhang T, Huang Y (2017) A fast speech feature extraction method based on perceptual hashing. In: 2017 13th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD). IEEE, pp 1295–1300
38. Zhang C, Zhu L, Xu C (2017) Ptbi: an efficient privacy-preserving biometric identification based on perturbed term in the cloud. Inf Sci 409:56–67
39. Zhang Q-Y, Ge Z-X, Qiao S-B (2018) An efficient retrieval method of encrypted speech based on frequency band variance. J Inf Hiding Multimed Signal Process 9:1452–1463, 11
40. Zhang Q, Hu W, Huang Y, Qiao S (2018) An efficient perceptual hashing based on improved spectral entropy for speech authentication. Multimed Tools Appl 77(2):1555–1581
41. Zhang Q, Qiao S, Huang Y, Zhang T (2018) A high-performance speech perceptual hashing authentication algorithm based on discrete wavelet transform and measurement matrix. Multimed Tools Appl 77(16):21653–21669
42. Zhang Q, Xing P, Huang Y, Dong R, Yang Z (2018) An efficient speech perceptual hashing authentication algorithm based on dwt and symmetric ternary string. Int J Inf Commun Technol 12(1–2):31–50
43. Zhang X, Zhang J, He T, Chen Y, Shen Y, Xu X (2018) A speech and lip authentication system based on android smart phone. In: Proceedings of the 6th international conference on information technology: iot and smart city. ACM, pp 110–114
44. Zhang Q, Zhang T, Wu D-F, Ge Z-X (2018) Strong robust speech authentication algorithm based on quasi-harmonic model. J Huazhong Univ Sci Technol (Nat Sci Ed) 46(3):58–64
45. Zhang Q, Zhang T, Qiao S-B, Wu D-F (2019) Spectrogram-based efficient perceptual hashing scheme for speech identification. Int J Netw Secur 21(2):259–268
46. Zhu L, Shen J, Xie L (2017) Unsupervised visual hashing with semantic assistant for content-based image retrieval. IEEE Trans Knowl Data Eng 29(2):472–486