Journal of Computers Vol. 31 No. 2, 2020, pp. 168-179
doi:10.3966/199115992020043102015

A Safety and Efficient Speech Biological Hashing Algorithm

Qiuyu Zhang*, Gaili Li, Si-bin Qiao, Yong-bing Zhang

School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China
{zhangqylz, ligaili01, qiaosibin, gstszyb}@163.com

Received 19 July 2018; Revised 10 December 2018; Accepted 9 February 2019

Abstract. In this paper, we present a safety and efficient speech biological hashing algorithm based on discrete wavelet transform (DWT) and Mersenne Twister algorithm for content-based speech authentication and retrieval, and effectively solved poor trade-off between robustness and discrimination, weak compactness, low security and operation efficiency in the process of constructing speech perceptual hashing. Firstly, the speech signal is conducted with DWT after pre-processing conducts to produce low frequency wavelet coefficients, which are used to generate the wavelet Merlin matrix. Then, the pseudo random matrix is generated by the Mersenne Twister algorithm and conducted with fast Fourier transform (FFT) to produce pseudo random Fourier matrix. Finally, the wavelet Merlin matrix and pseudo random Fourier matrix are conducted by iterations and thresholds to produce the binary perceptual hashing sequence. Experimental results show that when compared with the existing algorithm, the proposed algorithm has a good robustness and discrimination for speech content preserving operation, such as re-sampling, volume adjustment, echo addition and MP3 compression, etc., and has good compactness and high security.

Keywords: biological hashing, discrete wavelet transform (DWT), Mersenne Twister algorithm, pseudo random matrix, robustness, speech perceptual hashing

1 Introduction

How to accurately and quickly retrieve the required content from massive speech data, and realize the authenticity and integrity authentication of speech data has always been a hot topic in the field of multimedia research [1], especially the use of speech perceptual hashing technology [2-3] to extract efficient perceptual features is a necessary condition for accurate retrieval and integrity authentication. The biological hashing function is a special kind of perceptual hashing function. In addition to satisfying the discrimination and robustness, it also has the one-wayness with trapdoors [4].

At present, speech feature extraction and processing algorithms based on speech perceptual hashing mainly include logarithmic cepstral coefficients, linear prediction coefficients (LPC), linear spectrum frequencies (LSF), etc. For example, Zhao et al. [5] proposed a speech perceptual hashing algorithm that piecewise aggregate approximation (PAA) was used for compressing data size and multifractal extract perceptual hashing, but it was not good at robustness and discrimination of the speech perceptual hashing function. Awais et al. [6] proposed a feature extraction of the speech signal using Mel-Frequency Cepstral Coefficients (MFCC) feature vectors, the algorithm has good robustness and recognition accuracy, but its security is poor. He et al. [7] proposed a speech hashing algorithm based on syllable-level. Experimental results show that the syllable-level perceptual hashing of the proposed scheme has good discrimination, and perceptual robustness to common speech, but the security is not under consideration Kim et al. [8] proposed an audio fingerprint extraction algorithm based on modulated complex lapped transform (MCLT) and adaptive threshold. It has good robustness for content preserving operations and its time consumption is low, but the security is not under consideration. Ghouti et al. [9] proposed a robust and perceptual fingerprinting solution that serves both video content identification and authentication, the algorithm has good robustness, discrimination and sensitivity. Li et al. [10] proposed a

[*] Corresponding Author

speech perceptual hashing algorithm based on the correlation coefficient of MFCC and a pseudo random sequence. The algorithm has good robustness and security. However, its discrimination is weak. Wang et al. [11] proposed using the zero-crossing rate to extract the perceptual hashing sequence, but the robustness and discrimination are poor, besides, there was no better balance between robustness and compactness. Li et al. [12] proposed an audio perceptual hashing algorithm based on non-negative matrix factorization (NMF) and modified discrete cosine transform (MDCT) coefficients. It has highly robustness for content preserving operations, and its discrimination is good, but it needs more time to generate hashing sequences. Chen et al. [13] proposed a speech hashing function based on NMF and LPC, and the linear prediction analysis is applied to obtain LPCs, and NMF is performed on the LPCs to capture speech's feature. However, the algorithm do not consider the security. Zhang et al. [14] proposed an efficient perceptual hashing algorithm based on improved spectral entropy for speech authentication, combining with the linear prediction-minimum mean squared error (LP-MMSE), the algorithm has good robustness and discrimination, but its security is not considered. Zhang et al [15] proposed an efficient speech perceptual hashing algorithm based on DWT and measurement matrix. It has good robustness, discrimination and security, but the efficiency is not high. Viswanathan et al. [16] proposed a multi-granularity geometrically robust video hashing method. The algorithm is robust against temporal de-synchronization and geometrical transformation. Zannettou et al. [17] proposed an origins of memes by means of fringe web communities. The algorithm detects and measure the propagation of memes across multiple Web communities, using a processing pipeline based on perceptual hashing and clustering techniques. Lacharme et al. [18] proposed a biological template protection algorithm. By reconstructing fingerprints, the original and counterfeit fingerprint can be accurately identified by the biological hashing algorithm. It is proved that biological hashing has a good protection and recognition performance for biometric templates.

By analyzing the above literature, the existing methods do not have good trade-off between robustness and discrimination in feature extraction from speech signals, low operational efficiency, and lack of security considerations for hashing sequences. In addition, the biological hashing has better robustness and discrimination, and good protection for biometric templates as a feature extraction method. The biological hash algorithm is mainly applied to the feature extraction of images, in order to take advantage of the biological hash algorithm. It can be well applied to the construction of speech perceptual hashing [19].

In this paper, we present a safety and efficient speech biological hashing algorithm. The main contributions of our method are summarized as follows:

(1) We propose the features of the biological hashing function, which can have good protection for biometric templates as a feature extraction method, while traditional hashing algorithm lack of security considerations for hashing sequences.

(2) We apply DWT and Mersenne Twister algorithm to construct biological hashing. The experimental results show that the proposed algorithm has very good discrimination, compactness and one-wayness.

(3) When the generated hashing sequence is destroyed, a new biological hashing sequence can be generated by the new key, and thus has good revocability, while traditional hashing algorithm lose the revocability.

(4) We evaluate the proposed method on 1280 databases, and experimental results demonstrate the effectiveness and robustness of the proposed method.

The rest of this paper is organized as follows. Section 2 shows the problem statement and preliminaries. Section 3 gives the proposed algorithm. Section 4 presents experimental results and analysis. Finally, we conclude our paper in Section 5.

2 Related Theory

2.1 Biological Hashing Algorithm

The biological hashing [4, 18] is a special perceptual hashing function. The basic idea is to define an orthogonal random transformation function using a specific key. Using this function is to transform the biometrics and further binarize to obtain the hashing sequence. During the registration phase, the system calculates the hashing value of the user and stores it in the library. The hashing value of the user to be authenticated is obtained in the same way, and the Hamming distance between the two hash values of the authenticated user and the registered user is calculated. The identity authentication is completed according to the relationship between the Hamming distance and the threshold. The biological hashing algorithm principle is shown in Fig. 1.

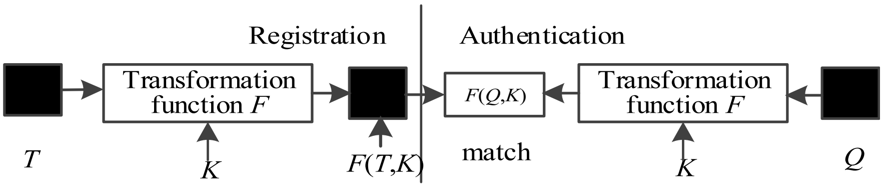


Fig. 1. Biological hashing algorithm principle

In Fig. 1, *T* represents a biometric template, *K* is a key, *F* (*T*, *K*) represents a transformed template, and *Q* represents a biometric to be authenticated.

Firstly, the biological hashing function of this paper is to preprocess the speech data, then using the DWT transform obtain the speech wavelet Merlin transform feature, and then according to the information stored in the user token produce a set of pseudo random numbers. The feature vector obtained by transforming the speech is iteratively inner product with the set of pseudo random numbers, and then threshold to obtain a specific set of binary vectors.

The biological hashing has better robustness and discrimination, and good protection for biometric templates as a feature extraction method. The biological hash algorithm is mainly applied to the feature extraction of images, in order to take advantage of the biological hash algorithm, the biological hashing is adopted for speech feature extraction in this work.

2.2 Discrete Wavelet Transform

Using wavelet transform construct biometric vector to generate retrieval digest. The discrete wavelet transform aims to discretize the scale and translation of the basic wavelet. In the application, it is necessary to discretize the scale factor *a* and the displacement factor *b*, as shown in Eq. (1):

$$a = a_0^m, b = nb_0 a_0^m \ , \tag{1}$$

where, *m* and *n* are integers, a_0 is a constant greater than 1, b_0 is a constant greater than 0, the choice of *a* and *b* is related to the specific form of the wavelet.

The discrete wavelet transform function is expressed as follows:

$$\varphi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} \varphi(\frac{t - nb_0 a_0{}^m}{a_0{}^m}) = \frac{1}{a_0{}^m} \varphi(a_0{}^{-m} t - nb_0) \ . \tag{2}$$

The corresponding discrete wavelet transform is represented as follows:

$$w_f(m,n) = < f, \varphi_{m,n}(t) > = \int_{-\infty}^{+\infty} f(t) \varphi_{m,n}(t) dt \ . \tag{3}$$

When a_0=2 and b_0=1, the DWT transform is called as binary discrete wavelet transform.

The significance of wavelet decomposition lies in the ability to decompose signals at different scales, and the choice of different scales can be determined according to different goals. For many signals, the low frequency component is quite important. They often contain the characteristics of signals, while high-frequency components give details or differences of signals. If the speech signal removes the high-frequency component, it may sound different from the previous one, but it can still know the content; if you remove enough low-frequency components, you hear some meaningless sound. Approximation and detail are often used in wavelet analysis, the approximation represents the high-level signal. That is, it is low-frequency information. Fig. 2 shows how a discrete signal can be transformed by discrete wavelet using a hierarchical architecture.

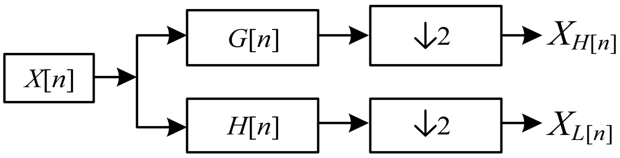


Fig. 2. Discrete signal discrete wavelet transform schematic

In Fig. 2, $X[n]$ is a discrete input signal of length N. $G[n]$ is a low-pass filter that filters the high-frequency part of the input signal and outputs the low-frequency part. $H[n]$ is a high-pass filter, in contrast to the low-pass filter, which filters out low-frequency components and outputs high-frequency components. ↓2 is a down sampling filter. If $X[n]$ is taken as the input, then $X_{L[n]}$ or $X_{H[n]}$ is output, where $X_{H[n]}$ is the high frequency part of $X[n]$ which obtained speech signal through the low-pass filter and the down-sampling filter; $X_{L[\mathrm{n}]}$ is the low-frequency part of the speech signal obtained by inputting low-pass filter and down-sampling filter as input, and the low-frequency part can be used to construct the wavelet Merlin matrix.

2.3 Mersenne Twister

The Mersenne Twister (MT) algorithm [20] is one of the pseudo random number generators used to generate pseudo random numbers based on matrix linear regeneration on finite binary fields. It can quickly produce high-quality pseudo random number, which corrects many of the shortcomings of the old pseudo random number generation algorithm. Its recursive formula is:

$$x_{k+m} := x_{k+m} \oplus \left(x_k^{\,u} \mid x_{k+1}^{\,l}\right) A,\ k = 0,1 \ldots n \ . \tag{4}$$

where, ⊕ represents exclusive OR (XOR) operation, | represents a logical OR operation, x^u and x^l are respectively the high and low bits of the vector $\boldsymbol{x}$, n and m are constant integers, and n>m>0, (x_0, x_1, x_2, ..., $x_{n\text{-}1}$) is the initialization seed, x_k is a row vector of binary representation, and the parameter matrix $\boldsymbol{A}$ is a w×w dimensional row vector consisting of 0 and 1 elements.

When the parameter matrix $A = \begin{pmatrix} 0 & l_{w-1} \\ a_{w-1} & a_{w-2} \cdots a_0 \end{pmatrix}$, the computational complexity can be minimized. Where $l_{w\text{-}1}$ represents the unit matrix of w-1×w-1 dimension, and matrix $\boldsymbol{A}$ is set to:

$$x \cdot \mathrm{A} = \begin{cases} x \gg 1 & ; x_0 = 0 \\ x \gg 1 \oplus a & ; x_0 = 1 \end{cases} , \tag{5}$$

where, x:=x_k^u|x_{k+1}^l, k=0, 1, ..., n, x>>1 means shifting one byte to the right.

After the result of the above process, multiplying a matrix $\boldsymbol{t}$ obtain the final pseudo random matrix $\boldsymbol{y}$ which is the required random matrix, and math formula is as shown in Eq. (6):

$$\begin{aligned} y &:= x \oplus (x >> u) \\ y &:= y \oplus ((y << s)\ \mathrm{AND}\, b) \\ y &:= y \oplus ((y << t)\ \mathrm{AND}\, c) \\ y &:= y \oplus (y >> t) \end{aligned} . \tag{6}$$

where u denotes additional Mersenne Twister algorithm tempering bit shifts, s and t denote shift register tempering bit shifts, b and c denote shift register tempering bitmasks, and >> denotes the shift register shift to the right, << denotes the shift register shift to the left.

We apply DWT and Mersenne Twister algorithm to construct biological hashing. The Mersenne Twister algorithm can quickly produce high-quality pseudo random number, which corrects many of the shortcomings of the old pseudo random number generation algorithm The Mersenne Twister algorithm is designed as the secret key to enhance the security of the proposed algorithm.

3 The Proposed Algorithm

Fig. 3 shows the processing flow chart of the speech biological hashing generation algorithm.

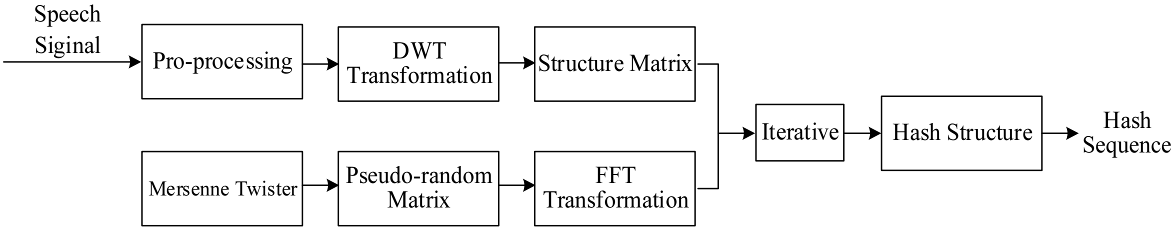


Fig. 3. The flow chart of proposed algorithm

Firstly, 3-level wavelet decomposition is performed on speech signal after the pre-processing part, the low-frequency coefficients of the DWT transformed speech is extracted, and using low-frequency coefficients construct speech feature vector; then the pseudo random matrix generated by the Mersenne Twister algorithm and the FFT produce pseudo random Fourier matrix. Finally, the speech feature vector and pseudo random Fourier matrix are conducted with iterations, and the result of the iteration is threshold to generate the binary hashing sequence which is used to complete the process of hash construction.

The biological hashing function is used to extract the speech perceptual features. The specific processing is as follows:

Step 1: Pre-processing. The input speech signal $s(t)$ is pre-emphasized to obtain the signal $s'(t)$, so that the high-frequency useful part of the signal is improved to facilitate subsequent feature extraction. The speech signal to be tested $s(t)$ has a sampling frequency of 16 kHz, the number of channels is mono, and the sampling accuracy is 16 bits.

Step 2: DWT transform. 3-level wavelet decomposition is performed on speech signal $s'(t)$, and low-frequency coefficients are obtained, which are denoted as $L(z) = \{L_{\mathrm{i}} \mid i = 1, 2, \ldots, N\}$, and N is the length of low-frequency coefficients. The low-frequency coefficient $L(z)$ extracted by the wavelet transform is constructed, and the wavelet Merlin matrix $\boldsymbol{X}(n, m)$ is generated as described in Section 1.2.

Step 3: Generate a Mersenne Twister pseudo random matrix.

(1) Set the initial value:

$$\begin{aligned} & u \leftarrow \underbrace{1\cdots1}_{w-r}\underbrace{0\cdots0}_{r} \\ & II \leftarrow \underbrace{0\cdots0}_{w-r}\underbrace{1\cdots1}_{r} \qquad , \\ & a \leftarrow a_{w-1}a_{w-2}\cdots a_1a_0 \end{aligned} \tag{7}$$

where r denotes the number of bits of the lower bitmask, $0 < r < w$-1, u denotes the high byte w-r bit mask, II denotes the low byte r bit mask, the initial value of w is 0x80000000UL, the initial value of r is 0x7fffffffUL, and a is the initial matrix $\boldsymbol{A}$=9908B0DF16.

(2) Initialize an n-dimensional matrix:

$$\begin{aligned} & i \leftarrow 0 \\ & x[0], x[1], \cdots x[n-1] \end{aligned} . \tag{8}$$

(3) Get the matrix $\boldsymbol{y}$ value:

$$y \leftarrow (x[i] \text{ AND } u) \text{ XOR } (x[i+1] \bmod n) \text{ AND } II \cdot \tag{9}$$

Calculate the value of $x_k^u \mid x_{k+1}^l$ and store it in matrix $\boldsymbol{y}$. Where II represents the low byte r bit mask.

(4) Matrix **y** multiplied by the initial matrix $\boldsymbol{A}$

$$x[i] \leftarrow (x[i+m] \bmod n) \text{ XOR } (y >> 1) \times A , \tag{10}$$

where n=624, m=397

(5) Output matrix $\boldsymbol{y}$ value

$$\begin{aligned} & y \leftarrow x[i] \\ & y \leftarrow y \text{ XOR } (y >> u) \\ & y \leftarrow y \text{ XOR } (y << s) \text{ AND } b\,, \\ & y \leftarrow y \text{ XOR } (y << t) \text{ AND } c \\ & y \leftarrow y \text{ XOR } (y >> l) \end{aligned} \qquad \textbf{(11)}$$

where u=11, s=1, t=15, b=9D2C568016, c= EFC6000016, l=18.

(6) i cycle assignment: i←(i+1) mod 624.

Returning to **Step 3**, the cycle is executed. Finally, the pseudo random matrix $\boldsymbol{y}$(624, 397) is obtained. The number less than 0.013 in the matrix $\boldsymbol{y}$ is selected, and the selected number is constructed to obtain the matrix $\boldsymbol{B}(n, m)$.

Step 4: FFT transformation. Performing FFT transformation on the pseudo random matrix $\boldsymbol{B}$ generated in **Step 3** to obtain a pseudo random Fourier matrix $\boldsymbol{Y}(n, m)$.

Step 5: Hash construction. Using the matrix $\boldsymbol{X}$ obtained in **Step 2** and the matrix $\boldsymbol{Y}$ generated in **Step 4** is carried out iteratively to generate the matrix $\boldsymbol{P}(n, n)$, and then the N-dimensional matrix $\boldsymbol{P}$ produce the matrix $\boldsymbol{H}(1, N)$ and calculate the average value of the matrix $\boldsymbol{H}$ as avg.

Using the matrix $\boldsymbol{H}$ is conducted with hash structure to generate a hashing sequence $h=\{h(i) \mid i=1, 2,\dots, M\}$. The binary hash structure method is as follows:

The average value avg of the matrix $\boldsymbol{H}$ is subtracted from all the data of the parameter matrix $\boldsymbol{H}$. If it is greater than 0, the data of the row becomes 1; otherwise it is 0. The hash constructor is:

$$h(i) = \begin{cases} 1, & \mathit{if}\ \ H(i) - avg > 0 \\ 0, & \quad else \end{cases} \quad i\text{=1,2...}M\ , \qquad \textbf{(12)}$$

where, M is the length of the perceptual hashing sequence.

4 Experimental Results and Analysis

The speech data used in the experiment is the speech in the Texas Instruments and Massachusetts Institute of Technology (TIMIT) and Text to Speech (TTS) speech library, which consists of different content of Chinese and English, men and women, and the sampling frequency is 16 kHz. The sampling accuracy is 16 bit, the number of channels is mono, and speech clip is 4 s long, all in wav format, 640 speech clips of English and Chinese speech files, totaling 1280 speech clips.

The experimental hardware platform is: Intel (T) Core (TM) i5-2450M CPU, 2.50GHz, 8GB of memory. The software environment is: Windows 7, MATLAB R2016a.

4.1 Discrimination Analysis

Discrimination is used in the evaluation algorithm to read the reliability of different speech content for different or the same person. The bit error rate (BER) of the hash value of different speech content basically obeys the normal distribution, and the BER is the ratio of the number of error bytes to the total number of bytes. Calculated as follows:

$$BER = \frac{\sum_{i=1}^{M}((h_1(i) \oplus h_2(i)}{M}\,, \qquad \textbf{(13)}$$

where h_1 and h_2 are speech hashing sequences and M represents the length of the perceptual hashing sequence.

Take 1280 speech clips (different speakers, different content) as test speech and extract the hashing sequence. 818560 BER data is obtained by comparing the two perceptual hashing values of 1280 speech clips. The normal distribution of BER for different content speech is shown in Fig. 4.

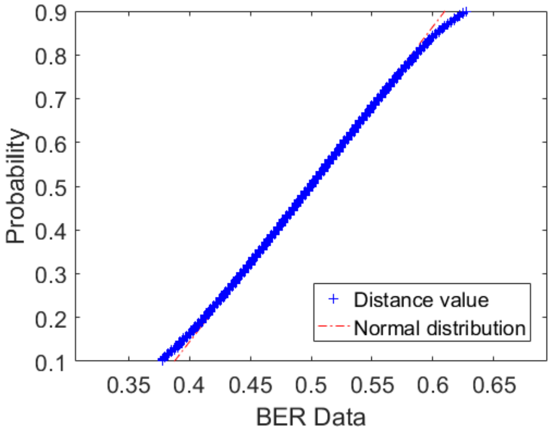


Fig. 4. BER distribution of different content speeches

It can be seen from Fig. 4 that the probability distribution of the BER value of different speeches almost overlaps with the probability curve of the standard normal distribution, so the hash distance value obtained by the proposed algorithm approximates a normal distribution.

According to the DeMoivre-Laplace central limit theorem, the Hamming distance approximates the normal distribution $\mu = p, \sigma = \sqrt{p(1-p)/M}$, where M is the length of the perceptual hashing sequence, μ is the BER mean, σ is the BER standard deviation, p represents the probability of "0" and "1", and in the ideal situation p=0.5. The length of the hashing sequence of the speech clip of the proposed algorithm is M=400. The mean of the idealized normal distribution parameter is μ=0.5, the standard deviation σ=0.0250, the mean of the actual test value in the experiment is μ_0=0.4993, and the standard deviation σ_0=0.0281. The test value is close to the theoretical value, indicating that the proposed algorithm has good randomness and discrimination (anti-collision).

In addition, discrimination and robustness are the most important characteristics of the speech perceptual hashing sequence, which can be measured by the two criteria of false accept rate (FAR) and false reject rate (FRR). The lower the FAR value, the better the discrimination, the lower the FRR value, and the better the robustness. The formula for calculating the FAR can be defined as Eq. (14).

$$R_{FAR}(\tau) = \int_{-\infty}^{\tau} f(x/\mu,\sigma) = \frac{1}{2\pi\sigma}\int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx\ . \tag{14}$$

The FRR can be defined as:

$$R_{FRR}(\tau) = 1 - \int_{-\infty}^{\tau} f(x/\mu,\sigma) = 1 - \frac{1}{2\pi\sigma}\int_{-\infty}^{\tau} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx\ , \tag{15}$$

where, R_{FAR} and R_{FRR} represent FAR and FRR respectively. τ is perceptual threshold, μ represents the BER mean, σ is called BER variance, x is false acceptance rate.

In order to verify the correctness of the proposed algorithm experiment, this paper uses Eq. (14) to further calculate the FAR value generated under different thresholds. Table 1 shows the error rate of 1280 speech clips under different thresholds in the proposed algorithm and the algorithm in Ref. [11-14]. FAR value comparison results are shown in Table 1.

Table 1. The FAR value of different algorithms comparison

Threshold τ	0.10	0.20	0.25	0.30	0.35
Ref. [11]	1.196×10^{-20}	5.518×10^{-16}	5.755×10^{-12}	1.367×10^{-08}	7.481×10^{-06}
Ref. [12]	3.114×10^{-35}	1.557×10-^{20}	9.493×10^{-15}	5.314×10^{-10}	2.785×10^{-6}
Ref. [13]	1.486×10^{-22}	1.742×10^{-13}	6.777×10^{-10}	6.264×10^{-7}	1.398×10^{-4}
Ref. [14]	7.746×10^{-32}	1.333×10^{-18}	2.153×10^{-13}	4.052×10^{-09}	9.056×10^{-06}
Ref. [15]	3.654×10^{-42}	1.405×10^{-24}	1.215×10^{-17}	6.166×10^{-12}	1.874×10^{-7}
Proposed	3.973×10^{-46}	8.602×10^{-27}	3.596×10^{-19}	6.584×10^{-13}	5.387×10^{-08}

As shown in Table1, the thresholds τ of proposed algorithm and in comparative Ref. [11-15] have good discrimination between 0.1 and 0.35.When the threshold τ=0.35, there are only 5.38 speech clips will be falsely accepted in 10^8 speech clips, which indicates that the proposed algorithm is resistant to collisions. The ability is very strong, and the error rate is much lower than the BER in Ref. [11-15]. It can be seen that the proposed algorithm has a good discrimination.

4.2 Robustness Analysis

In order to test the robustness of the proposed algorithm, selecting 1280 speech clips with wav speech format are used to perform 12 kinds of content preserving operations as shown in Table 2, and the BER mean and maximum values of various content preserving operations were calculated.

Table 2. Content preserving operations

Operating means and method		BER	
		Mean	Max
Volume I: Volume up 50%	V. I	0.0245	0.0850
Volume II: Volume down 50%	V. II	5.0861×10^{-5}	0.0050
Noise addition: *SNR*=30 dB narrowband Gaussian noise	E.A	0.0065	0.0250
Noise reduction: Noise reduction 75%	N.R	0.0866	0.1950
Re-quantization I: Quantizing the audio clip to 8bit, and then back to 16bit	R.Q I	0.0059	0.0700
Re-quantization II: Quantizing the audio clip to 32bit, and then back to 16bit	R.Q II	1.9562×10^{-6}	0.0025
Re-sampling I: The speech is conducted down-sampling to 8 kHz, and then back to 16 kHz	R.S I	0.0017	0.0275
Re-sampling II: The speech is conducted up-sampling to 32 kHz and then back to 16 kHz	R.S II	0.0079	0.0625
MP3 compression I: Compressing and decompressing the audio clip with MP3 at 48 kbps	M. I	0.0112	0.0350
MP3 compression II: Compressing and decompressing the audio clip with MP3 at 128 kbps	M. II	0.0018	0.0225
MP3 compression III: Compressing and decompressing the audio clip with MP3 at 32 kbps	M. III	0.0218	0.0550
MP3 compression IV: Compressing and decompressing the audio clip with MP3 at 192 kbps	M. IV	0.0018	0.0225

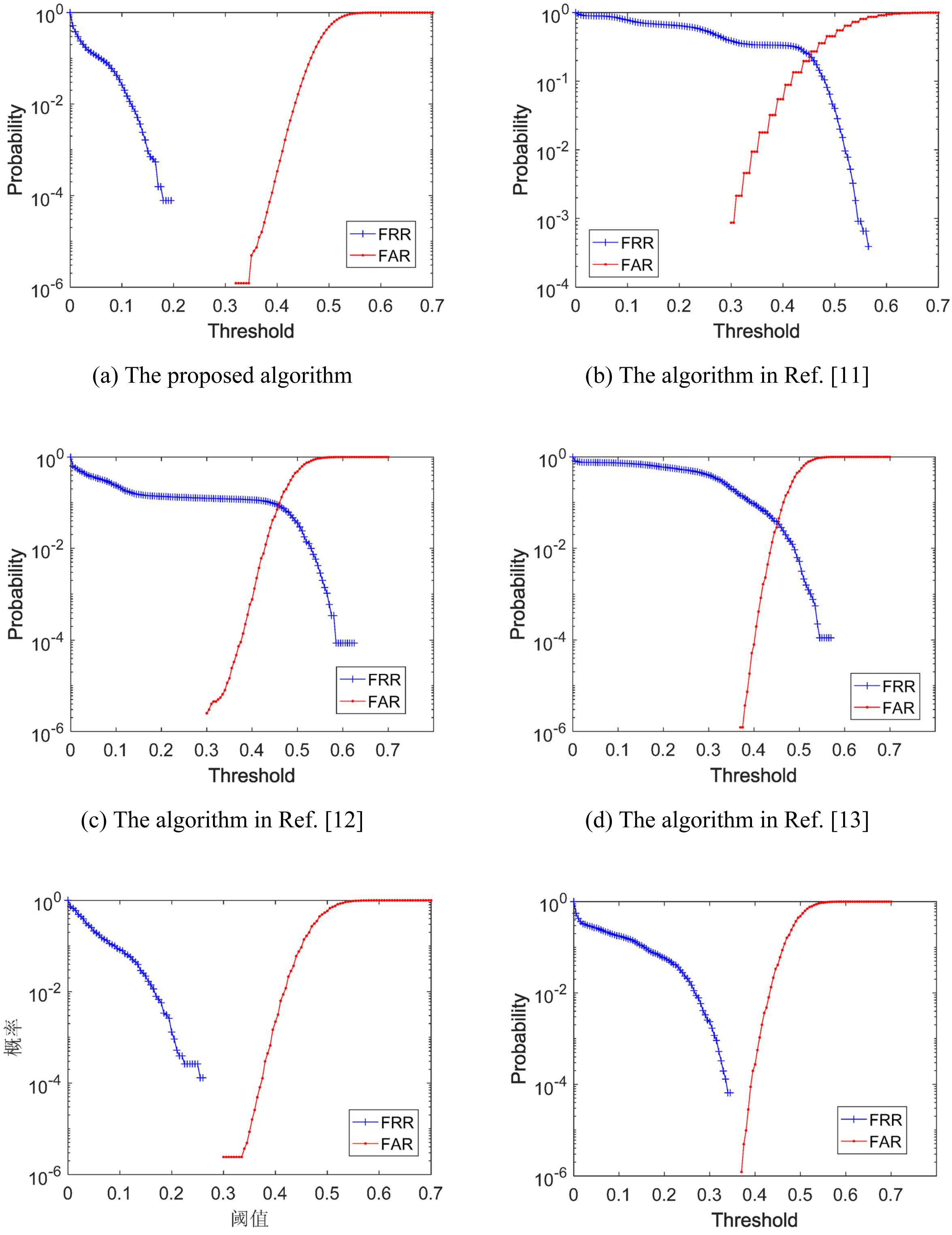
As shown in Table 2, the maximum value of the speech BER mean with the same perceptual content is 0.0866, and the maximum BER is 0.1950, which indicates that the proposed algorithm has better robustness.

The BER mean values between the original speech and processed speech are obtained according to their hash sequences. The BER mean values of the proposed algorithm and those of algorithms in [11-15] are compared in Table 3. It can be seen that the average BER values of the proposed algorithm are less than those of the algorithms in Ref. [11-15]. Therefore it denotes that the proposed algorithm has better robustness than the algorithms in Ref. [11-15].

Table 3. The average BER comparison results

Operating means	Proposed	Ref. [11]	Ref. [12]	Ref. [13]	Ref. [14]	Ref. [15]
V. I	0.0245	0.0054	0.0630	0.1761	0.0592	0.0264
V. II	5.1×10^{-5}	0.0151	2.3×10^{-4}	0.1469	5.6×10^{-5}	5.3×10^{-5}
E.A	0.0065	0.1071	0.1700	0.2132	0.1280	0.1427
N.R	0.0866	0.1778	0.1137	0.3444	0.1193	0.0879
R.Q I	0.0059	0.2239	0.0296	0.3335	0.0325	0.0156
R.Q II	2.0×10^{-6}	0	8.6×10^{-6}	4.3×10^{-6}	8.8×10^{-6}	0
R.S I	0.0017	0.0110	0.0217	0.1567	0.0125	0.0110
R.S II	0.0079	0.0874	0.0180	0.3766	0.0887	0.0136
M. I	0.0112	0.0475	0.4851	0.4835	0.2931	0.0142
M. II	0.0018	0.0239	0.4842	0.4817	0.2855	0.2189
M. III	0.0218	0.0702	0.5812	0.4873	0.3033	0.2583
M. IV	0.0018	0.0243	0.4872	0.4817	0.2856	0.1148

In order to further verify the robustness of the proposed algorithm, the FAR-FRR curve is drawn according to the 12 content preserving operations in Table 2, and the FAR-FRR curve is plotted and compared with the algorithm in Ref. [11-15], as shown in Fig. 5.



(a) The proposed algorithm

(b) The algorithm in Ref. [11]

(c) The algorithm in Ref. [12]

(d) The algorithm in Ref. [13]

(e) The algorithm in Ref. [14]

(f) The algorithm in Ref. [15]

Fig. 5. FAR-FRR curves

As shown in Fig. 5(a) to Fig. 5(f), the FAR-FRR curve of the proposed algorithm is not cross, which still can distinguish the same processed speech and the different speech well. While the curves of the algorithm in Ref. [11-13] are all crossed, which cannot be Content preserving operations are distinguished from different speech content. The FAR-FRR curve of the proposed algorithm and in the algorithm in Ref. [14-15] is not cross and have a large threshold interval, which indicate that the proposed algorithm has strong robustness and discrimination. When the decision threshold τ =0.25, the FRR value of the proposed algorithm is 0, the corresponding FRR of in the algorithm in Ref. [14] is 1×10^{-4}, and the corresponding FRR of in the algorithm in Ref. [15] is 1×10^{-3}, which indicates that the robustness of the proposed algorithm is better than in the algorithm in Ref. [14-15]. When the threshold τ is chosen in 0.20-0.30, the FAR and FRR values of the proposed algorithm are both zero. Therefore, comparing with in the algorithm in Ref. [11-15], especially in the cases of re-sampling, and MP3 compression etc, when the threshold is small, the FRR value is still small enough. It demonstrates the proposed algorithm has very good robustness for content preserving operations.

4.3 Security Analysis

In this paper, the pseudo random matrix generated by Mersenne Twister can be used as a key to encrypt the speech perceptual hashing sequence. In the process of perceptual hashing sequence generation, the proposed algorithm combines DWT transform and Mersenne Twister pseudo random matrix to extract the speech perceptual features and construct the biological hashing sequence. The proposed algorithm uses a pseudo random matrix generated by Mersenne Twister to generate 624 32-bit random numbers by cycle, which has strong randomness. In addition, the speech perceptual feature extraction using the random matrix generated by Mersenne Twister is equivalent to a single encryption process, thus ensuring the security of the proposed algorithm.

4.4 Revocability Analysis

If the perceptual feature extracted by the existing perceptual hash feature extraction algorithm is destroyed, it cannot be recovered and it will result in a certain loss. Therefore, biological hashing has a good protection for biometric templates in feature extraction. The key in the pseudo random generation function in the algorithm is different, and the generated perceptual hashing sequence (also referred to as a biometric template) is also different. Therefore, when the generated biometric template is stolen, the biometric template is invalidated by changing the key in the pseudo random generation algorithm, and a new feature template is generated through the new key. Therefore, revocability can be easily provided by changing the key in the pseudo random generation algorithm. However, in the proposed algorithm, the pseudo random generation algorithm is reversible. Once the attacker obtains the key and the transformed template, the original speech feature can be recovered and the user's private information can be obtained. Therefore, the security of the algorithm also depends on the security of the key.

4.6 One-wayness Analysis

The biological hashing has One-wayness with trapdoors. Only when the matching speech information is input again, can the original key be calculated from the biological hashing digest. However, for the attacker who does not match the speech information cannot be calculated key. Therefore, the algorithm shows secure one-wayness. When authenticating the user's identity by speech, firstly, the submitted query speech clips are pre-processed, and then the speech feature sets f' is obtained, only when the elements of the submitted query speech feature sets f' and the original speech feature sets have enough matches, it can be recovering the pseudo random number by recovery polynomial.

4.7 Efficiency Analysis

In order to evaluate the computational efficiency of the proposed algorithm, 1280 speech clips of 4s length were selected for testing. Calculating the feature extraction and matching average running time of the proposed algorithm were compared with in the algorithm in Ref. [11-15], and speech clips is 4 s long. As shown in Table 4.

Table 4. Comparison of average running time of algorithms

Algorithm	Platform working frequency/Hz	Average running time /s	Hash length
Proposed	2.50	0.0387	400
Ref. [11]	2.50	0.0288	1024
Ref. [12]	2.50	0.1304	360
Ref. [13]	2.27	0.1603	360
Ref. [14]	2.30	0.0388	266
Ref. [15]	3.20	0.0848	360

As shown in Table 4, it can be seen from Table 4 that the running efficiency of the proposed algorithm is significantly higher than that of in the algorithm in Ref. [12-15]. Because of the Mersenne Twister algorithm in the proposed algorithm is used to obtain speech feature values after DWT transformation, the schedule is simple and the data is reduced obviously, As in the algorithm in Ref. [12-13] uses non-negative matrix factorization, resulting in lower operating efficiency, in the algorithm in Ref. [15] is slower to combine with the linear prediction-minimum mean squared error, while in the algorithm in Ref. [15] is slower to construct matrix using measurement matrix, but the proposed algorithm is lower than the literature [11]], this is because the use of the construction matrix method to generate the hashing sequence results in low efficiency, while efficiency in Ref. [11] using the zero-crossing rate to extract the perceptual hashing sequence is higher than the proposed algorithm, but the length of the hashing sequence (compactness) is higher than the proposed algorithm in this paper.

5 Conclusions

For the practical application of speech content authentication or mass speech retrieval, in order to achieve efficient speech feature extraction and improve the security of hashing sequence, a safety and efficient speech biological hashing algorithm was proposed. The proposed algorithm extracts speech perceptual features by applying DWT and Mersenne Twister algorithm to construct biological hashing, and generates a binary biological hashing sequence representing the unique features of the speech. The experimental results show that the proposed algorithm has very good discrimination, compactness, one-wayness and operational efficiency, has excellent robustness on content preserving operations, especially in the cases of re-sampling, echo addition, and MP3 compression. In addition, the feature extraction using the pseudo random matrix generated by the Mersenne Twister algorithm is equivalent to construct one-time encryption process for hashing sequence extraction, so it has good security. When the generated hashing sequence is destroyed, a new biological hashing sequence can be generated by the new key, and thus has good revocability.

In this paper, although the proposed speech biological hashing algorithm has superior performance by applying DWT and Mersenne Twister algorithm, the robustness on content preserving of low pass filtering needs to be improved. In the future, we plan to improve the robustness on content preserving of low pass filtering. Besides, the proposed algorithm extracts speech perceptual features has only 1280 speech database, extracting speech perceptual features from tremendous speech database also needs to be further studied.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61862041, 61363078), the Research Project in Universities of Education Department of Gansu Province (2017B-16, 2018A-187). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] D. Josset, J. Pelon, N. Pascal, On the use of CALIPSO land surface returns to retrieve aerosol and cloud optical depths. IEEE Transactions on Geoscience and Remote Sensing 56(6)(2018) 3256-3264.

[2] J. Li, T. Wu, Perceptual audio hashing using RT and DCT in wavelet domain, in: Proc. of IEEE International Conference on Computational Intelligence and Security, 2015.

[3] H. Chen, Y. Wo, G. Han, Multi-granularity geometrically robust video hashing for tampering detection, Multimedia Tools and Applications 77(5)(2018) 5303-5321.

[4] R. Singh, D. Rai, R. Prasad, Similarity Detection in Biological Sequences using Parameterized Matching and Q-gram, in: Proc. IEEE Technology and Computational Sciences on 2018 Recent Advances on Engineering, 2018.

[5] H. Zhao, S. He, A retrieval algorithm for encrypted speech based on perceptual hashing, in: Proc. of IEEE International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2016.

[6] A. Awais, S. Kun, Y. Yu, Speaker recognition using mel frequency cepstral coefficient and locality sensitive hashing, in: Proc. of IEEE 2018 International Conference on Artificial Intelligence and Big Data, 2018.

[7] S. He, H. Zhao, A retrieval algorithm of encrypted speech based on syllable-level perceptual hashing, Computer Science & Information Systems 14(3)(2017). DOI:10.2298/CSIS170112024H.

[8] H.G. Kim, H.S. Cho, J.Y. Kim, Robust audio fingerprinting using peak-pair-based hash of non-repeating foreground audio in a real environment, Cluster Computing 19(1)(2016) 315-323.

[9] L. Ghouti, A new perceptual video fingerprinting system, Multimedia Tools and Applications 77(6)(2018) 6713-6751.

[10] J.F. Li, T. Wu, H.X. Wang, Perceptual hashing based on correlation coefficient of MFCC for speech authentication, Journal of BUPT 38(2)(2015) 89-93.

[11] H. Wang, L. Zhou, W. Zhang, Watermarking-based perceptual hashing search over encrypted speech, in: Proc. International Workshop on Digital Watermarking, 2013.

[12] J. Li, H. Wang, Y. Jing, Audio perceptual hashing based on NMF and MDCT coefficients, Chinese Journal of Electronics 24(3)(2015) 579-588.

[13] N. Chen, W.G. Wan, Robust speech Hash function, ETRI Journal 32(2)(2010) 345-347.

[14] Q. Zhang, W. Hu, Y. Huang, An efficient perceptual hashing based on improved spectral entropy for speech authentication, Multimedia Tools and Applications 77(2)(2018) 1555-1581.

[15] Q. Zhang, S. Qiao, Y. Huang, A high-performance speech perceptual hashing authentication algorithm based on discrete wavelet transform and measurement matrix, Multimedia Tools and Applications 77(16)(2018) 21653-21669.

[16] N. Viswanathan, K. Kokkinakis, B. Williams, Listeners experience linguistic masking release in noise-vocoded speech-in-speech recognition, Journal of Speech, Language, and Hearing Research 61(2)(2018) 428-435.

[17] S. Zannettou, T. Caulfield, J. Blackburn, On the origins of memes by means of fringe web communities, in: Proc. 18th ACM Internet Measurement Conference, 2018

[18] P. Lacharme, Revisiting the accuracy of the biohashing algorithm on fingerprints, IET biometrics 2(3)(2013) 130-133.

[19] Y. Zheng, Y. Cao, C.H. Chang, Facial biohashing based user-device physical unclonable function for bring your own device security, in: Proc. of International Conference on Consumer Electronics, 2018.

[20] J. Zhu, L. Huai, R. Cui, Research and Application of hybrid random selection genetic algorithm, in: Proc. IEEE International Symposium on Computational Intelligence and Design, 2017.