


# A high-performance speech perceptual hashing authentication algorithm based on discrete wavelet transform and measurement matrix

Qiu-yu Zhang<sup>1</sup>  · Si-bin Qiao<sup>1</sup> · Yi-bo Huang<sup>2</sup> ·  
Tao Zhang<sup>1</sup>

Received: 24 January 2017 / Revised: 10 December 2017 / Accepted: 3 January 2018 /  
Published online: 13 January 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

**Abstract** Aiming at the problems of existing speech authentication algorithms, such as poor robustness and discrimination, security vulnerability, low efficiency, poor ability of tamper detection and localization, a high-performance speech perceptual hashing authentication algorithm based on Discrete Wavelet Transform (DWT) and measurement matrix is proposed in this paper. Firstly, the speech signal is conducted with DWT by applying preprocessing, and the low-frequency wavelet coefficients are regarded as the perceptual feature value. Then the measurement matrix controlled by chaos map is applied to reduce the dimension of feature value. Finally, the feature value is used to generate the perceptual hash sequence by the process of hashing structure. The measurement matrix is designed as the secret key to enhance the security of the proposed algorithm. The experimental results demonstrates the proposed algorithm has high efficiency in perceptual robustness, discrimination, time consumption and security, as well as having a high accuracy of tampering detection and localization.

---

✉ Qiu-yu Zhang  
zhangqylz@163.com

Si-bin Qiao  
qiaosibin@163.com

Yi-bo Huang  
Huangyibo1982@163.com

Tao Zhang  
1090408922@qq.com

<sup>1</sup> School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

<sup>2</sup> College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China

**Keywords** Speech authentication · Perceptual hashing · Discrete wavelet transform (DWT) · Measurement matrix · Tampering detection

## 1 Introduction

It is self-evident that the fundamental function of speech in human social activities is important. For example, mobile voice communication is the most important and pervasive way for telecommunication, telephone recording is an effective and powerful evidence presented at court and news broadcasting is a popular medium for people to get essential various news events. However, tampered or replaced critical speech clips in these cases will result in a non-negligible impact on people, organizations and even countries. Especially, if the content of news broadcasting was replaced, wrong information will be transmitted to the public, which may have negative influence on society. Therefore, in order to guarantee the reliable communication and content security of speech information, it is necessary to authenticate the speech content and speech integrity [1, 18]. Traditional cryptographic hash function, with the extreme sensitivity to speech content change, including noise and re-sampling, is not suitable for speech content authentication anymore. Nevertheless, speech perceptual hash function protects multimedia information by verifying the integrity of multimedia contents, which makes multimedia information service safer and more reliable, also, introduces into speech retrieval and speech content integrity certification [19].

The speech feature extraction methods based on perceptual hashing mainly include Logarithmic cepstral coefficients [15], Linear Spectrum Frequency (LSF) [14], Mel-Frequency Cepstral Coefficients (MFCC) [7, 16], Linear Prediction Coefficients (LPC) [12], Hilbert transform [20], temporal modulation [13] and bark-bands energy [17], etc. Nouri et al. [14] proposed a speech authentication algorithm based on linear spectrum frequencies (LSF) and DWT. Huang et al. proposed a speech perceptual hashing algorithm based on MFCC combined with Linear Prediction Cepstral Coefficient (LPCC). The algorithm has good robustness and accurate tamper localization, but it is not good at distinguishing different speech. Jiao et al. [8] proposed a speech perceptual hashing algorithm which regarded the LSP parameterization as the perceptual feature, and its hash structure depends on a secret key. Although the algorithm has good security and collision resistance, strong robustness of content preserving operations and good ability of detecting and locating malicious attack, the positional accuracy of it remains to be improved. Chen et al. [2] proposed a speech hashing algorithm based on LPCC and vector quantization. Kim et al. [9] proposed an audio fingerprint extraction algorithm based on Modulated Complex Lapped Transform (MCLT) and adaptive thresholding. It is robust for content preserving operations and its time consumption is low, but the security is not under consideration. Chen et al. [4] proposed a novel audio perceptual hashing algorithm which uses the Zernike matrix amplitude of audio signals and virtual watermark detection to generate the perceptual hash sequence. The algorithm has good robustness and discrimination no matter what the test object is, being it either music or speech. While, its biggest disadvantage is that its running time is more than ten times of other algorithms. Li et al. [10] proposed a 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. Chen et al. [6] proposed a perceptual hashing algorithm based on cochleagram and cross-recurrence analysis, the Non-negative Matrix Factorization (NMF) is applied to reduce dimension. The algorithm has good

robustness, but its efficiency is low. An audio perceptual hashing algorithm based on NMF and Modified Discrete Cosine Transform (MDCT) coefficients are proposed by Li et al. [11]. It is highly robust to content preserving operations, and its discrimination is good, but it needs more time to generate hash sequences. Chen et al. [3] proposed a speech hash function based on NMF and LPC, and the linear prediction analysis is applied to obtain LPCs, and the NMF is performed on the LPCs to capture speech's feature. However, the algorithm do not consider the security, and the robustness against various types of content preserving operations is not up to the expected standard. Chen et al. [5] proposed a perceptual hashing algorithm based on DWT and NMF, the algorithm has good robustness and discrimination, but its time consumption is unsatisfied.

Aiming at the problems mentioned above, to improve authentication efficiency, achieve a balance between robustness and discrimination, guarantee the security of the authentication algorithm, realize high accuracy of tampering localization and meet the requirement of low complexity, we present a high-performance speech perceptual hashing authentication algorithm based on DWT and measurement matrix. The algorithm has better robustness and discrimination, higher authentication efficiency based on speech content, good in security. Furthermore, it can achieve a small range of tamper detection and localization. To assure the security of the proposed algorithm, the measurement matrix is designed as a secret key to encrypt the perceptual feature value, and the key consumption is reduced by combining the measurement matrix with the logistic chaotic sequence.

The remaining part of this paper is organized as follows. Section 2 describes the discrete wavelet transform. Section 3 introduces the details of the proposed algorithm. Section 4 gives the experimental results and performance analysis as compared with other related methods. Finally, we conclude our paper in Section 5.

## 2 Discrete wavelet transform

The discrete wavelet transform (DWT) is performed on discrete data sets to produce discrete output. The time-frequency window can be adaptively transformed with the signal, and the DWT can accurately express the local details of the speech signal.

For a given arbitrary function  $f(t)$  in the space of energy limited  $L^2(R)$ , and the expansion that processed under the wavelet basis function  $\varphi_{a,b}(t)$  is called continuous wavelet transform when  $a$  and  $b$  are continuous values. The function is represented as follows:

$$W_f(a, b) = \int_{-\infty}^{+\infty} f(t) \overline{\varphi_{a,b}(t)} dt \quad (1)$$

where  $\overline{\varphi_{a,b}(t)}$  is the conjugate of  $\varphi_{a,b}(t)$ ,  $W_f(a, b)$  is called wavelet transform coefficients,  $t$  is function variable,  $a$  and  $b$  are scale and translation parameters respectively.

The wavelet basis function  $\varphi_{a,b}(t)$  and wavelet transform coefficients  $W_f(a, b)$  are processed with the discrete way against scale parameter  $a$  and translation parameter  $b$ . Generally, the discrete formulas of scale parameter  $a$ , the translation parameter  $b$  are  $a = a_0^j$  and  $b = ka_0^j b_0$ , and the discrete wavelet basis function is represented as follows:

$$\phi_{j,k} = a_0^{-1/2} \phi(a_0^{-j} - kb_0) \quad (2)$$

where  $j, k$  belong to integer set, the value of  $a_0$  and  $b_0$  depend on wavelet basis function  $\varphi_{a,b}(t)$  and  $a_0 > 0$ ,  $b_0 > 0$ .

Therefore, the DWT transform of signal  $f(t)$  could be represented as follows:

$$W_f(j, k) = \int_{-\infty}^{+\infty} f(t) \overline{\phi_{j,k}}(t) dt \quad (3)$$

For the most of speech and images, the wavelet transform has good time–frequency localization property, and low-frequency components contain the feature of the signal. The high-frequency components contain the details of the signal. When processing the speech signal  $s(t)$  using DWT, the speech signal  $s(t)$  is discretized firstly, and get the discrete signal  $s(z)$ . The DWT schematic of the speech signal is shown in Fig. 1.

In Fig. 1,  $H_0(z)$  is the low-pass filter factor, and  $H_1(z)$  is the high-pass filter factor.  $s(z)$  is processed by down-sampling after filtered by low pass filter, and the coarse signal  $L(z)$  is obtained whose scale and resolution are halved, which is called low-frequency components.  $s(z)$  is processed by down-sampling after filtered by the high-pass filter, and the detail signal  $H(z)$  is obtained whose scale and resolution are halved, which is called high-frequency components.

### 3 The proposed algorithm

The processing flow of the proposed algorithm based on DWT and measurement matrix is shown in Fig. 2. First, 3-level wavelet decomposition is performed on speech signal after the pre-processing part, and then the low-frequency coefficients are processed by measurement matrix to reduce the dimension. The hash vector is generated by using the feature value which is extracted in the previous step. Finally, the speech signal is authenticated through matching the corresponding hash sequences.

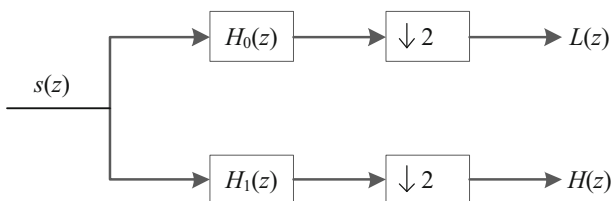
The measurement matrix is also used to encrypt data. If the entire measurement matrix is regarded as a secret key, the key consumption will be large, in order to reduce the key consumption while encrypting, the key-controlled measurement matrix [21] is introduced to reduce the key consumption.

The processing steps are as follows:

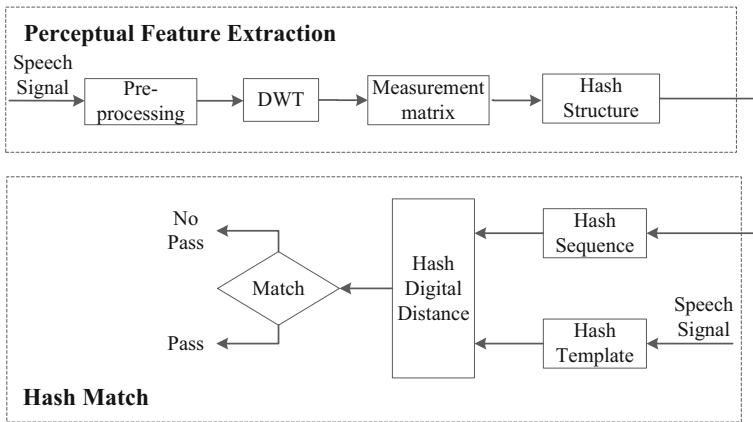
**Step 1:** Pre-processing. The input speech signal  $s(t)$  needs to conduct a pre-emphasis process to enhance the part of high frequency. The speech signal processed by pre-emphasis is denoted as  $s'(t)$ .

**Step 2:** DWT. 3-level wavelet decomposition is performed on speech signal  $s'(t)$ , and low-frequency coefficients are obtained, which are denoted as  $\mathbf{L} = \{L_i \mid i = 1, 2, \dots, N\}$ , and  $N$  is the length of low-frequency coefficients.

**Step 3:** Design the measurement matrix. A  $M \times N$  Bernoulli matrix  $\mathbf{B}$  is designed according to Eq. 4, whose each element obeys Bernoulli distribution independently,



**Fig. 1** The DWT schematic of speech signal



**Fig. 2** Schematic diagram of the proposed speech perceptual hashing authentication algorithm

and the elements  $\sqrt{\frac{3}{M}}$ , 0, and  $-\sqrt{\frac{3}{M}}$  appear with the probability of 1/6, 2/3 and 1/6 respectively.

$$\mathbf{B} = \begin{cases} +\sqrt{\frac{3}{M}} & P = \frac{1}{6} \\ 0 & P = \frac{2}{3} \\ -\sqrt{\frac{3}{M}} & P = \frac{1}{6} \end{cases} = \sqrt{\frac{3}{M}} \begin{cases} +1 & P = \frac{1}{6} \\ 0 & P = \frac{2}{3} \\ -1 & P = \frac{1}{6} \end{cases} \quad (4)$$

where  $P$  denotes the probability of the elements occurred.

When  $M=4$ ,  $N=6$ , we will obtain a measurement matrix  $\mathbf{B}$  as follows:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & -0.8660 & 0 \\ 0 & 0 & 0.8660 & -0.8660 & 0 & -0.8660 \\ 0 & 0 & 0 & 0 & 0 & 0.8660 \\ -0.8660 & 0 & 0 & -0.8660 & 0 & 0.8660 \end{bmatrix}$$

Generating a sequence with length  $M$  by logistic map with initial condition  $r = 0.11$  should be generated,  $\mathbf{a} = [a_1, a_2, \dots, a_M]$ , sorting the nature sequence  $n = [1, 2, \dots, M]$  with the index sequence  $\mathbf{a}$ , and noting the sorted sequence as  $\mathbf{a}^* = [a^*_1, a^*_2, \dots, a^*_M]$ , where  $a^*_i \in [1, M], i \in [1, M]$ , and then choosing the row vectors,  $\mathbf{B}(a^*_1, :)$ ,  $\mathbf{B}(a^*_2, :)$ , ...,  $\mathbf{B}(a^*_M, :)$  to group into the measurement matrix  $\Phi$ .

$$\Phi = \begin{bmatrix} B(a^*_1, :) \\ B(a^*_2, :) \\ \vdots \\ B(a^*_M, :) \end{bmatrix} \quad (5)$$

**Step 4:** Feature matrix extraction. The speech feature values  $\mathbf{H} = \{H(i) \mid i = 1, 2, \dots, M\}$  are extracted by low frequency coefficients  $\mathbf{L} = \{L_i \mid i = 1, 2, \dots, N\}$  and measurement matrix  $\Phi$ . The detail is shown as follows:

$$\mathbf{H} = \Phi \times \mathbf{L} \tag{6}$$

**Step 5:** Hash structure. The hash sequence  $\mathbf{ph} = \{ph(i) \mid i = 1, 2, \dots, M\}$  is decided by the feature value vector  $\mathbf{H}$ . In the process of hash structure, the median value of  $\mathbf{H}$  is added to vector  $\mathbf{H}$  and get a new vector  $\mathbf{H} = \{H(i) \mid i = 1, 2, \dots, M + 1\}$ ,  $H(1)$  is the median value of the original  $\mathbf{H}$ . The hash sequence structure method is shown as follows:

$$\mathbf{ph}(i) = \begin{cases} 1, & \text{if } H(i + 1) > H(i) \\ 0, & \text{else} \end{cases} \quad i = 1, 2, \dots, M \tag{7}$$

where  $M$  is the dimension of row space of measurement matrix, its size is equal to the length of the hash sequence.

**Step 6:** Hash digital distance and matching. For two different speech clips  $s_1$  and  $s_2$ , and the Hamming distance method is utilized to measure the distance between any two hash vectors. The distance is measured by Bit Error Rate (BER), denoted as  $x$ :

$$x = D(PH(s_1), PH(s_2)) = \frac{1}{M} \sum_{j=1}^M |ph_{s_1}(j) - ph_{s_2}(j)| \tag{8}$$

where  $ph_{s_1}$  and  $ph_{s_2}$  represent the hash sequence value of  $s_1$  and  $s_2$  respectively.  $\mathbf{PH}(\cdot)$  represents speech hash function, which can make speech signal transform into the hash sequence.

The problem of hash matching can be formulated as the hypothesis testing using the hash function  $\mathbf{PH}(\cdot)$  and the distance measure  $\mathbf{D}(\cdot, \cdot)$ .

$P_0$ : if the perceptual content of the two speech clips  $s_1$  and  $s_2$  are the same:

$$D(PH(s_1), PH(s_2)) \leq \tau \tag{9}$$

$P_1$ : if the perceptual content of the two speech clips  $s_1$  and  $s_2$  are not the same:

$$D(PH(s_1), PH(s_2)) > \tau \tag{10}$$

where  $\tau$  represents the perceptual authentication threshold,  $\mathbf{PH}(\cdot)$  is called perceptual hashing function. By setting the size of matching threshold  $\tau$ , and calculating the digital distance between perceptual hashing sequences of the speech clips  $s_1$  and  $s_2$ , we can judge whether they are the same. If the digital distance  $\mathbf{D}(\cdot, \cdot) \leq \tau$ , then their perceptual content is treated as the same, and the authentication is passed, otherwise it could not be passed.

In order to evaluate the performance of the authentication algorithm, the False Accept Rate (FAR) and False Reject Rate (FRR) are defined as follows

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

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

where  $R_{FAR}$  and  $R_{FRR}$  represent FAR and FRR respectively,  $\mu$  and  $\sigma$  are the expected values and the standard deviation of  $x$ ,  $\tau$  represents predetermined threshold. Generally speaking, FAR and FRR are utilized to evaluate the robustness and discrimination of the authentication algorithm. The lower FAR denotes the better discrimination, and the lower FRR denotes the better robustness.

## 4 Experimental results and analysis

The speech data used in the experiment is from Texas Instruments and Massachusetts Institute of Technology (TIMIT) and Text to Speech (TTS) speech library which composed of different speech recorded by the Chinese men and women and English men and women. The library is composed of 1280 speech clips that 16 bits signed, 16 kHz, mono, and 4 s long. Experimental platform: Pentium® Dual-Core CPU E6700 @3.20GHz, 2G. Windows 7 SP1, MATLAB R2013a. The parameters are set as measurement matrix of  $M \times N$ , where  $M = 360$ ,  $N = 8002$ .

### 4.1 Robustness verification and analysis

In order to evaluate the robustness of the proposed algorithm, twelve common operations in Table 1 are used to simulate the interference to the signal, including noise, volume, echo and filter. On every speech in the speech data, twelve kinds of content preserving operations are performed one by one. After that, the feature values of the processed speech are extracted, and hash sequences are generated.

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 [3, 5, 11] are compared in Table 2.

It can be seen that the average BER values of the proposed algorithm are less than those of the algorithms in [3, 5, 11], and the maximum value is 0.2189. Therefore it denotes that the proposed algorithm has better robustness than the algorithms in [3, 5, 11].

For each speech clip in the speech data which composed of 1280 speech clips, the BER between its hash vector and that of each of the remaining 1279 speech clips is calculated, so a total amount of 818,560 BER values are obtained. According to the eight content preserving operations in Group I, the corresponding FAR values and FRR values are obtained, and the FAR-FRR curves of the proposed algorithm and the algorithms in [3, 5, 11] are shown in Fig. 3.

In Fig. 3, the FAR-FRR curves of the proposed algorithm and the algorithm in [11] are not cross, and the other two are cross. The threshold can be chosen in no crossing filed, if so, the

**Table 1** Content-preserving operation

| Group | Operating means      | Level  |
|-------|----------------------|--|
| I     | Volume adjustment I  | Volume up 50%  |
|       | Volume adjustment II | Volume down 50%  |
|       | Echo addition        | Adding an echo signal with a delay of 300 ms and a decay to 25%                          |
|       | Noise addition       | $SNR = 30$ dB narrowband Gaussian noise, center frequency distribution in $0 \sim 4$ kHz |
|       | Noise reduction      | Noise reduction 75%  |
|       | Re-quantization I    | Quantizing the audio clip to 8bits/sample and then back to 16bit /sample                 |
|       | Re-quantization II   | Quantizing the audio clip to 32bits/sample and then back to 16bit /sample                |
| II    | Re-sampling I        | The speech is conducted down-sampling to 8 kHz and then back to 16 kHz                   |
|       | Re-sampling II       | The speech is conducted up-sampling to 32 kHz and then back to 16 kHz                    |
|       | FIR filter           | 12 order FIR low-pass filter with cutoff frequency of 3.4 kHz                            |
|       | MP3 compression I    | Compressing and decompressing the audio clip with MP3 at 48 kbps                         |
|       | MP3 compression II   | Compressing and decompressing the audio clip with MP3 at 128 kbps                        |

FAR value and FRR value can be very small simultaneously. Therefore, the proposed algorithm and the algorithm in [11] can distinguish the same processed speech and different speech well. On the contrary, in Fig. 3(c) and Fig. 3(d), no matter what the threshold is, the FAR and FRR cannot be very small simultaneously, so the algorithms in [3, 5] cannot be able to distinguish the same processed speech and different speech well. Compared with the Ref. [11], the obvious advantage of the proposed algorithm is that the threshold could be chosen in a larger range 0.255–0.370, and the threshold of the algorithm [11] can only choose in 0.280–0.315, which illustrates that the threshold could be adjusted more flexibly.

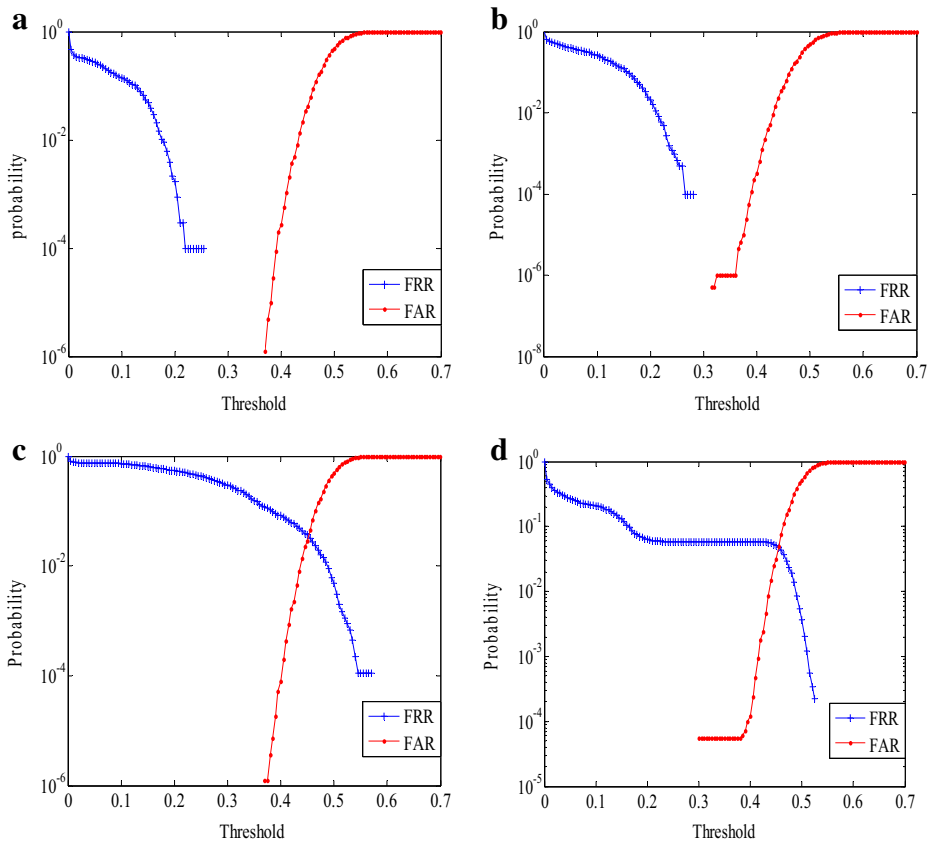
According to the all of the content-preserving operations in Table 1, all FAR values and FRR values are obtained, and the FAR-FRR curves of the proposed algorithm and the algorithms in [3, 5, 11] are shown in Fig. 4.

As shown in Fig. 4, when the other four kinds of content preserving operations in the Group II are considered, the FAR-FRR curve of the proposed algorithm is still not cross which still can distinguish the same processed speech and the different speech well. All the FAR-FRR curves of the algorithms in [3, 5, 11] are cross, and no matter what is the threshold, the FAR value and FRR value cannot be very small simultaneously, so they cannot distinguish the same processed speech and the different speech well. Especially when the four content preserving

**Table 2** The average BER comparison results

| Algorithm            | Proposed algorithm   | Algorithm [11]       | Algorithm [3]        | Algorithm [5]        |
|----------------------|----------------------|----------------------|----------------------|----------------------|
| Operating means      | Average BER          |                      |                      |                      |
| Volume Adjustment I  | 0.0264               | 0.0630               | 0.1761               | 0.0171               |
| Volume Adjustment II | $4.6 \times 10^{-5}$ | $2.3 \times 10^{-4}$ | 0.1469               | $5.9 \times 10^{-4}$ |
| Echo Addition        | 0.1427               | 0.1700               | 0.2132               | 0.1492               |
| Noise addition       | 0.0063               | 0.0346               | 0.3883               | 0.0686               |
| Noise reduction      | 0.0779               | 0.1137               | 0.3444               | 0.2492               |
| Re-quantization I    | 0.0056               | 0.0296               | 0.3335               | 0.2065               |
| Re-quantization II   | 0                    | $8.6 \times 10^{-6}$ | $4.3 \times 10^{-6}$ | 0                    |
| Re-sampling I        | 0.0010               | 0.0217               | 0.1567               | 0.0034               |
| Re-sampling II       | 0.0104               | 0.1280               | 0.3766               | 0.0364               |
| FIR Filter           | 0.0136               | 0.1821               | 0.3668               | 0.2080               |
| MP3 Compression I    | 0.0042               | 0.4851               | 0.4835               | 0.4520               |
| MP3 Compression II   | 0.2189               | 0.4842               | 0.4817               | 0.4517               |





**Fig. 3** Comparison of the FAR-FRR curves between the proposed algorithm and the algorithms in [3, 5, 11] according to content preserving operations in Group I: (a) the proposed algorithm (b) algorithm in [11] (c) algorithm in [3] (d) algorithm in [5]

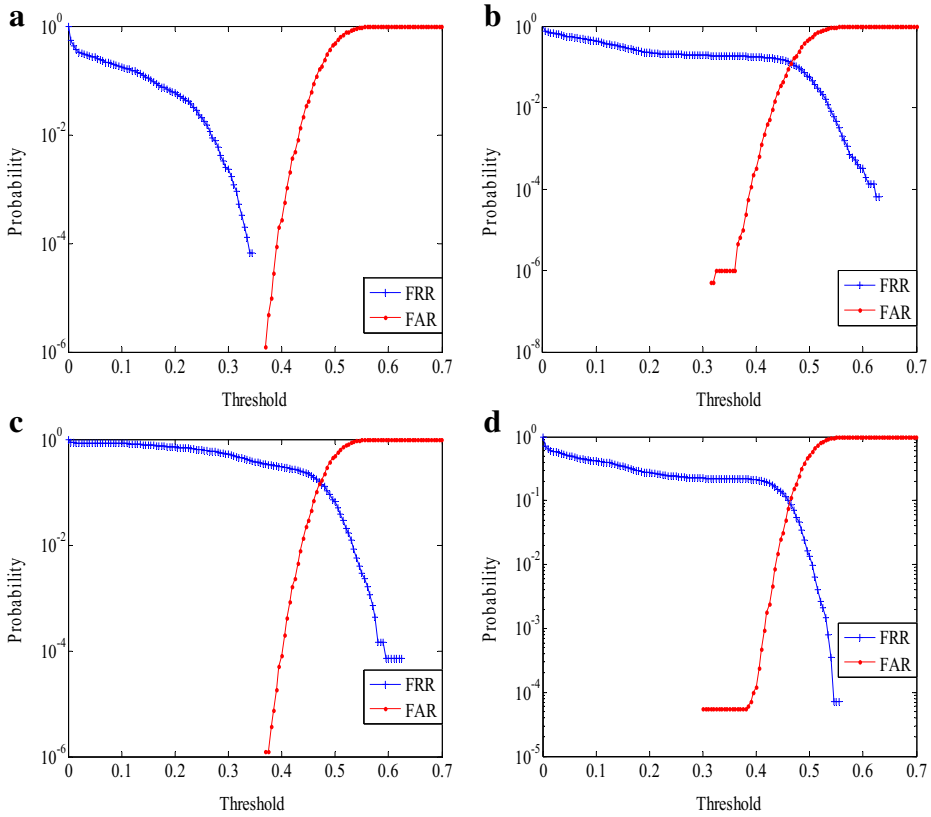
operations in Group II are added, the FRR value is bigger, and it denotes that the robustness of algorithm in [11] has become worse.

Comparing with other three algorithms on the robustness to content preserving operations in Table 1, especially in the cases of re-sampling, low pass filtering, and MP3 compression, when the threshold is small, the FRR value is still small enough. It demonstrates the proposed algorithm has huge advantages on the robustness to different content preserving operations, and its overall performance is superior to other three algorithms. When the threshold is chosen in 0.345–0.370, the FAR value and FRR value are small enough simultaneously.

### 4.2 Discrimination analysis

The BER values of the different perceptual hashing sequence basically obey the normal distribution. According to the BER values obtained in Section 4.1, the normal probability plot distribution of the BER is shown in Fig. 5.

According to the De Moivre-Laplace central limit theorem, the Hamming distance is similar to a normal distribution ( $\mu = p, \sigma = \sqrt{p(1-p)} / M$ ,  $M$  is the number of bits in a hash



**Fig. 4** Comparison of the FAR-FRR curves between the proposed algorithm and the algorithms in [3, 5, 11] according to content preserving operations in Table 1: (a) the proposed algorithm (b) algorithm in [11] (c) algorithm in [3] (d) algorithm in [5]

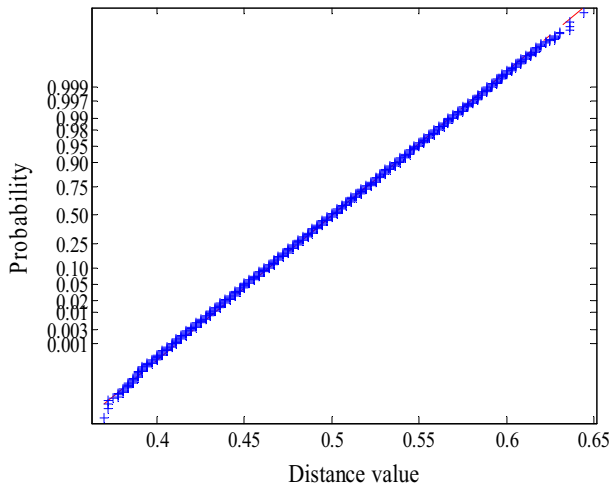
sequence). In this paper, the Hamming distance is used to denote BER,  $\mu$  is the mean value of BER,  $\sigma$  denotes the standard deviation of BER,  $p$  represents the probability of “0” and “1”, and in the ideal situation  $p = 0.5$ . For  $M = 360$  in the proposed algorithm, the mean and standard deviation of the normal distribution are expected to be 0.5 and 0.0264 respectively. Table 3 describes the mean and standard deviation of the normal distribution of theoretical values and experimental values.

It can be seen in Table 3 that the normal distribution parameters obtained by the proposed algorithm are very close to the theoretical ones. Therefore, the hash sequence generated by the proposed algorithm has better randomness and collision resistance.

In order to assess the discrimination of the proposed algorithm in the different threshold, the FAR value is obtained by Eq. 11. The comparison of the FAR curve according to experiment and theoretical value is shown in Fig. 6, which can be seen that the FAR value of the proposed algorithm is close to the theoretical value.

Table 4 describes the comparison of the FAR value of the proposed algorithm and the algorithms in [3, 5, 10, 11].

As shown in Table 4, it shows the FAR values of different algorithms in different thresholds. When the threshold  $\tau = 0.35$ , the FAR value of the proposed algorithm is smallest,



**Fig. 5** The normal probability plot distribution of the BER

there are only 1.87 speech clips will be falsely accepted in  $10^7$  speech clips which is lower than the FAR in [3, 5, 10, 11]. Therefore, we can conclude that the proposed algorithm achieves better discrimination.

### 4.3 Security analysis

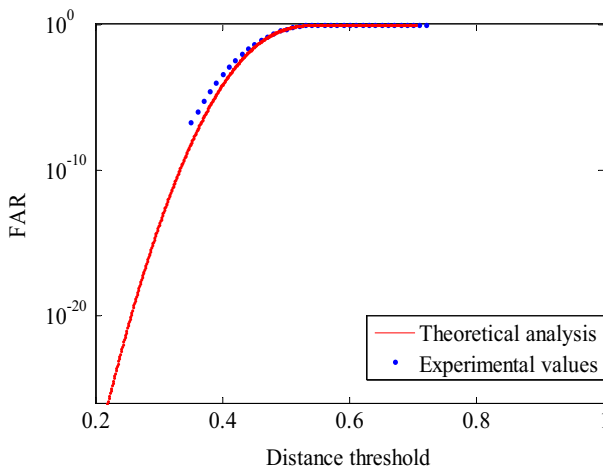
A key-controlled measurement matrix is introduced in the proposed algorithm to enhance the security. The measurement matrix based on logistic chaos map is designed as the secret key to encrypt speech feature value [21], due to the logistic chaotic sequence is combined with the traditional measurement matrix, the key consumption is reduced. In this paper, speech feature values are extracted by applying the measurement matrix, the measurement matrix is random and satisfying Bernoulli distribution independently, and it could be regarded as one-time encryption process if the speech features are extracted by applying measurement matrix. Therefore, the security of the proposed algorithm is greatly improved by using measurement matrix to extract speech feature.

### 4.4 Efficiency analysis

The computational efficiency of the algorithm is a very important evaluation criterion in the speech content authentication system. To evaluate the complexity and computational efficiency of the proposed algorithm, there are 400 speech clips were chosen randomly and the average running time is recorded. Table 5 describes the comparison results of the proposed algorithm and the algorithms in [3, 5, 6, 11].

**Table 3** The parameters of normal distribution

| Theoretical values |          | Experimental values |          |
|--------------------|----------|---------------------|----------|
| $\mu$              | $\sigma$ | $\mu$               | $\sigma$ |
| 0.5                | 0.0264   | 0.4999              | 0.0295   |



**Fig. 6** FAR curve of the proposed algorithm

As shown in Table 5, if the 4 s long speech clips were used to test, the efficiency of the algorithm proposed in this paper is 10.6 times of that in [6], it is 2.7 times of that in [11], it is 1.9 times of that in [3], and it is 6.3 times of that in [5]. Because of the measurement matrix is used to obtain speech feature values, the schedule is simple and the data is reduced obviously, so the efficiency is improved. The calculation is large in [3, 5, 6, 11] because of the NMF is used. What is more, the size of perceptual hash sequence of the proposed algorithm is 360 which is much shorter than the size ( $M=64 \times 8 \times 4$ ) in [8], and it is equal to the size in [3, 5, 11], it illustrates that the proposed algorithm has good ability of compactness and it is beneficial for fast speech authentication. So the proposed algorithm can meet the requirements of real-time authentication, and it has a big advantage to retrieve speech data on the cloud.

#### 4.5 Tampering detection and localization

The small-scale malicious attack generally happened by cutting and pasting part of the speech clip, the tampering range is small, and BER is low. If 10% of a speech clip is modified, the BER is 0.1278. According to Table 2, it cannot distinguish the content preserving operations and malicious attack.

The errors generated by malicious attack usually cause a great impact, and the BER which produced by content preserving operations is distributed evenly. To distinguish content preserving operations and malicious attacks, the local detection is needed. The BER value

**Table 4** Comparison of FAR values of different algorithms in different predetermined threshold

| Threshold | Proposed algorithm      | Algorithm [10]          | Algorithm [11]          | Algorithm [3]           | Algorithm [5]           |
|-----------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| 0.10      | $3.654 \times 10^{-42}$ | $2.939 \times 10^{-12}$ | $3.114 \times 10^{-35}$ | $1.486 \times 10^{-22}$ | $3.031 \times 10^{-38}$ |
| 0.20      | $1.405 \times 10^{-24}$ | $1.144 \times 10^{-5}$  | $1.557 \times 10^{-20}$ | $1.742 \times 10^{-13}$ | $2.689 \times 10^{-22}$ |
| 0.25      | $1.215 \times 10^{-17}$ | $2.715 \times 10^{-4}$  | $9.493 \times 10^{-15}$ | $6.777 \times 10^{-10}$ | $5.174 \times 10^{-16}$ |
| 0.30      | $6.166 \times 10^{-12}$ | $1.682 \times 10^{-3}$  | $5.314 \times 10^{-10}$ | $6.264 \times 10^{-7}$  | $7.542 \times 10^{-11}$ |
| 0.35      | $1.874 \times 10^{-7}$  | $9.99 \times 10^{-3}$   | $2.785 \times 10^{-6}$  | $1.398 \times 10^{-4}$  | $8.490 \times 10^{-7}$  |

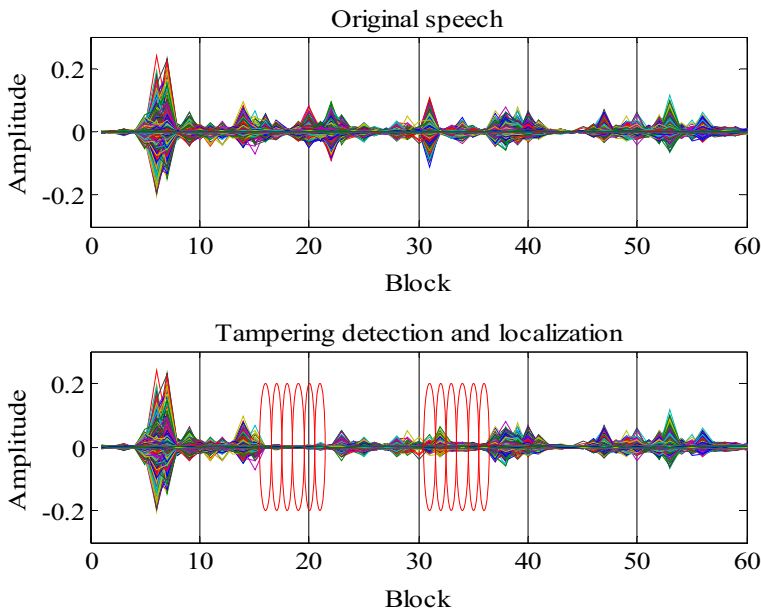
**Table 5** Comparison of operating efficiency of algorithms (average running time)

| Algorithm          | Platform working frequency | Average running time |
|--------------------|----------------------------|----------------------|
| Algorithm [6]      | 3.30 GHz                   | 0.9008 s             |
| Algorithm [11]     | 3.20 GHz                   | 0.2270 s             |
| Algorithm [3]      | 2.27 GHz                   | 0.1603 s             |
| Algorithm [5]      | 3.20 GHz                   | 0.5323 s             |
| Proposed algorithm | 3.20 GHz                   | 0.0848 s             |

of local speech clips is used to judge whether it is caused by content preserving operation or malicious attacks.

The average English speaking rate is 125 words per minute, and Chinese character speaking rate is 250 words per minute, every English word needs 480 ms, and a Chinese character needs 240 ms. In order to improve the precision of tampering detection, the 4 s speech clips are divided into 60 small blocks to generate the hash sequence, and the malicious attack is judged by that if there are 3 or more hash values are different in continuous six hash values. The tampering detection takes the small block as the minimum detection unit, if multiple positions are attacked, the related block will also be detected, so the proposed algorithm can be able to detect multiple positions of the signal.

Tampering detection and localization are shown in Fig. 7, the area of the red, elliptic curve are tamper areas. In Fig. 7, there are two positions of the speech signal is attacked, and the attacked positions are detected successfully by the proposed algorithm. The tampering detection takes the block as the minimum detection unit, and a 4 s long speech clip is divided into 60 small blocks which is regarded as the granularity, so the granularity is  $G = 4 \text{ s}/60 \approx 67\text{ms}$  which

**Fig. 7** Tampering detection and localization

**Table 6** Efficient of tampering detection and localization

| Type of attack | Range | Accuracy |
|----------------|-------|----------|
| Replacement    | 10%   | 0.9196   |

could meet the requirement of tampering detection and localization for one word or one Chinese character.

Generally, the malicious attack types include replacement and deletion, the speech content is randomly replaced and deleted by an attacker, so the integrity of speech content is damaged. In order to assess the efficient of tampering detection and localization, a speech database which includes 1020 speech clips is applied, and 10% of all speech clips are replaced randomly. The result is as follows:

As shown in Table 6, the accuracy of tampering detection and localization is 91.96%, it demonstrates that the part of the tampered speech could be detected and localized successfully by the algorithm.

## 5 Conclusions and future work

In this paper, we proposed a high-performance speech perceptual hashing authentication algorithm based on DWT and measurement matrix. The algorithm greatly solves the existing problems of current speech authentication algorithms and it achieves good robustness, discrimination, high computational efficiency, strong ability of tampering and localization as well as safety. Based on experiment results, the proposed algorithm has excellent robustness on content preserving operations, especially in the cases of re-sampling, low pass filtering, and MP3 compression. Besides, the efficiency of authentication and security are greatly enhanced by applying measurement matrix based on logistic chaos map and the granularity is reduced to one word.

Further research is planned to combine the proposed perceptual hashing algorithm with speech retrieval technology to implement efficient speech retrieval and authentication from tremendous speech contents.

**Acknowledgments** This work is supported by the National Natural Science Foundation of China (61363078), the Natural Science Foundation of Gansu Province of China (1606RJYA274), the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University (2014D13). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

## References

1. Adibi S (2014) A low overhead scaled equalized harmonic-based voice authentication system. *Telematics Inform* 31(1):137–152. <https://doi.org/10.1016/j.tele.2013.02.004>

2. Chen N, Wan WG (2009) Speech hashing algorithm based on short-time stability. In: International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, p 426–434. [https://doi.org/10.1007/978-3-642-04277-5\\_43](https://doi.org/10.1007/978-3-642-04277-5_43)
3. Chen N, Wan W (2010) Robust speech hash function. *ETRI J* 32(2):345–347. <https://doi.org/10.4218/etrij.10.0209.0309>
4. Chen N, Xiao HD (2013) Perceptual audio hashing algorithm based on Zernike moment and maximum-likelihood watermark detection. *Digital Signal Process* 23(4):1216–1227. <https://doi.org/10.1016/j.dsp.2013.01.012>
5. Chen N, Wan W, Xiao HD (2010) Robust audio hashing based on discrete-wavelet-transform and non-negative matrix factorization. *IET Commun* 4(14):1722–1731. <https://doi.org/10.1049/iet-com.2009.0749>
6. Chen N, Xiao HD, Zhu J, Lin JJ, Wang Y, Yuan WH (2013) Robust audio hashing scheme based on cochleagram and cross-recurrence analysis. *Electron Lett* 49(1):7–8. <https://doi.org/10.1049/el.2012.3812>
7. Huang Y, Zhang Q, Yuan Z, Yang Z (2015) The hash algorithm of speech perception based on the integration of adaptive MFCC and LPCC. *J Huazhong Univ Sci Tech (Natural Science Edition)* 43(2): 124–128. <https://doi.org/10.13245/j.hust.150226>
8. Jiao Y, Ji L, Niu X (2009) Robust speech hashing for content authentication. *IEEE Signal Process Lett* 16(9):818–821. <https://doi.org/10.1109/LSP.2009.2025827>
9. Kim HG, Cho HS, Kim JY (2016) Robust audio fingerprinting using a peak-pair-based hash of non-repeating foreground audio in a real environment. *Clust Comput* 19(1):315–323. <https://doi.org/10.1007/s10586-015-0523-z>
10. Li J, Wu T, Wang H (2015) Perceptual hashing based on the correlation coefficient of MFCC for speech authentication. *J Beijing Univ Posts Telecommun* 38(2):89–93. <https://doi.org/10.13190/j.jbupt.2015.02.016>
11. Li J, Wang H, Jing Y (2015) Audio perceptual hashing based on NMF and MDCT coefficients. *Chin J Electron* 24(3):579–588. <https://doi.org/10.1049/cje.2015.07.024>
12. Lotia P, Khan DM (2013) Significance of complementary spectral features for speaker recognition. *IJRCCCT* 2(8):579–588
13. Lu X, Matsuda S, Unoki M, Nakamura S (2011) Temporal modulation normalization for robust speech feature extraction and recognition. *Multimedia Tools Applications* 52(1):187–199. <https://doi.org/10.1007/s11042-010-0465-7>
14. Nouri M, Farhangian N, Zeinolabedini Z, Safarina M (2012) Conceptual authentication speech hashing base upon hypotrochoid graph. In: Telecommunications (IST), 2012 Sixth International Symposium on. IEEE 1136–1141. <https://doi.org/10.1109/ISTEL.2012.6483157>
15. Özer H, Sankur B, Memon N, Anarim E (2005) Perceptual audio hashing functions. *EURASIP J Adv Signal Process* 12:1780–1793. <https://doi.org/10.1155/ASP.2005.178>
16. Panagiotou V, Mitianoudis N (2013) PCA summarization for audio song identification using Gaussian mixture models. In: Digital Signal Processing (DSP), 2013 18th International Conference on. IEEE 1–6. <https://doi.org/10.1109/ICDSP.2013.6622803>
17. Ramona M, Peeters G (2011) Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 I.E. International Conference on. IEEE, p 477–480. <https://doi.org/10.1109/ICASSP.2011.5946444>
18. Wang ZR, Li W, Zhu BL, Li XQ (2012) Audio authentication based on music content analysis. *J Comput Res Dev* 49(1):158–166
19. Zhao H, He S (2016) A retrieval algorithm for encrypted speech based on perceptual hashing. In: Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on. IEEE 1840–1845. <https://doi.org/10.1109/FSKD.2016.7603458>
20. Zhao H, Liu H, Zhao K, Yang Y (2011) Robust speech feature extraction using the Hilbert transform spectrum estimation method. *Int J Digital Content Technol Appl* 5(12):85–95
21. Zhou N, Zhang A, Zheng F, Gong L (2014) Novel image compression–encryption hybrid algorithm based on key-controlled measurement matrix in compressive sensing. *Opt Laser Technol* 62:152–160. <https://doi.org/10.1016/j.optlastec.2014.02.015>



**Qiu-yu Zhang**, Researcher/Ph.D. 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.



**Si-bin Qiao**, received the BS degrees in communication engineering from Lanzhou University of Technology, Gansu, China, in 2014. His research interests include audio signal processing and application, multimedia authentication techniques.





**Yi-bo Huang** , received Ph. D. degree from Lanzhou University of Technology, Lanzhou, China, in 2015, and now working as a lecturer in the College of Physics and Electronic Engineering in Northwest Normal University. He main research interests include Multimedia information processing, Information security, Speech recognition.



**Tao Zhang** , received the BS degrees in communication engineering from Lanzhou University of Technology, Gansu, China, in 2015. His research interests include audio signal processing and application, multimedia authentication techniques.