Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **2**, March **2015**

An Efficient Speech Perceptual Hashing Authentication Algorithm Based on Wavelet Packet Decomposition

Qiu-Yu Zhang, Peng-Fei Xing, Yi-Bo Huang, Rui-Hong Dong and Zhong-Ping Yang

School of Computer and Communication
Lanzhou University of Technology
Gansu, Lanzhou, 730050, P. R. China
zhangqylz@163.com; xingpengfei0202@126.com

Received April, 2014; revised November, 2014

Abstract. *In this paper, we present a new speech perceptual hashing authentication algorithm based on wavelet packet decomposition (WPD) for speech content authentication. Firstly, a wavelet packet coefficients matrix of wavelet reconstruction is generated from an original speech signal after high frequency pre-processing followed by wavelet packet decomposition. Secondly, the wavelet packet coefficients matrix is partitioned into equal-sized sub-matrixes, and each sub-matrix is converted into new sub-matrixes by two-dimensional discrete cosine transform (DCT) resulting in good decorrelation and energy compaction. Finally, the feature parameter matrix is obtained by QR Decomposition (QRD) using a Givens Rotation (GR) with the new sub-matrixes. The experiment results illustrate that the proposed algorithm was very robust in content preserving operations, had a very low hash bit rate, and can meet the requirements of real-time speech authentication with high certification efficiency.*
Keywords: Speech content authentication, Perceptual hashing, Wavelet packet decomposition, Givens rotation, Robustness

1. **Introduction.** Because of the rapid development of cloud computing and mobile internet technology, various forms of digital audio, such as speech and music play an increasingly important role in the remote communication resulting in infinite dissemination and preservation. However, people can edit and modify digital audio products with the help of professional multimedia software (such as the audio editing software Cool Edit, MP3Cut). The openness of the wireless and network communication channel also provides opportunities which include illegal eavesdropping and tampering with malicious intent. Moreover, recorded and transmitted speech often contains important contents and sensitive information such as the commands in military communication networks, secret business communication, confidential information in confidential phone calls and the like. In order to guarantee reliable communication and the security of speech multimedia information in the cloud, it is necessary to verify the authenticity and integrity of digital speech content [1, 2].

Unlike cryptographic hashing, perceptual hashing involves the one-way mapping of a multimedia digital representation to a perceptual hashing digest, that is, a compact content-based digest that uniquely represents a speech clip. Perceptual hashing is sensitive to content change, robust in content preserving operations, and can better certify the

content integrity of the speech signal [3, 4]. For example, Jiao et al. [5] applied a randomization scheme controlled by a random seed in a hash generation method for random feature selection and post randomization to authenticate content integrity. However, this randomization method had little effect on robustness and the hash value was invariant in content preserving operations. Nouri et al. [6] used a robust hashing technique based on linear spectrum frequencies to model verbal territory to be authenticated. This result showed that the method was robust in content protective operations and could be used to perform very adequately in identification and verification. Chen et al. [7] treated the cochleagram of the audio as an image, from which accelerated robust features were extracted as essential features. The result indicated that the method performed very well but the algorithm was inefficient and did not perform well in real-time speech authentication. Huang et al. [8] proposed a perceptual hashing authentication algorithm based on linear forecast analysis (LPC) and, experiments showed that the algorithm had a high efficiency but was not adequately robust in content preserving operations.

In the paper, we describe a speech identification system which uses a novel summarization technique based on wavelet packet decomposition (WPD) that we suggest is an excellent way to authenticate non-stationary signals, such as speech signals. The system extracts speech feature parameters by QR decomposition (QRD) with a Givens Rotation applied to the wavelet packet coefficients matrix of the original speech signal resulting in strong robustness for content preserving operations even with signals contaminated by narrowband noise, amplitude boosting, low-pass filtering and the like. Compared with other methods, the proposed method has better overall performance especially in terms of robustness and signal discrimination.

The rest of this paper is organized as follows. Section 2 describes wavelet packet decomposition (WPD), discrete cosine transform (DCT) and QR decomposition (QRD) using a Givens Rotation. A detailed Speech Perceptual Hashing Authentication System is described in Section 3. Subsequently, Section 4 gives the experimental results as compared with other related methods. Finally, we conclude our paper in Section 5.

2. Related Theory Introduction.

In this section, the related theories of wavelet packet decomposition (WPD), discrete cosine transform (DCT) and QR decomposition (QRD) using Givens Rotation and their relationship to this paper are described briefly.

2.1. Wavelet Packet Decomposition (WPD).

WPD can analyze the features of a signal, adaptively select the appropriate frequency, and match with the signal spectrum to improve time-frequency resolution [9]. Therefore, WPD can well reflect the features and nature of the signal, and provide a new way to deal with non-stationary signals including speech signals.

Speech signal S conducts wavelet packet decomposition through recursion in Eq. (1).

$$\begin{aligned} U_{2m}(t) &= \sqrt{2}\sum_{n\in Z} h(n)u_m(2t-n) \\ U_{2m+1}(t) &= \sqrt{2}\sum_{n\in Z} g(n)u_m(2t-n) \end{aligned} \tag{1}$$

where $g(n)$ is a low pass filter group, $h(n)$ is a high-pass filter group, and $g(n) = (-1)^n h(1-n)$ indicate that two coefficients have an orthogonal relationship.

Fig. 1 shows N-level wavelet packet decomposition.

2.2. Discrete Cosine Transform (DCT).

After DCT, the frequency-domain speech signal results in better stability, good decorrelation capacity and energy compaction capability. In addition after a variety of attacks such as noise and resampling, the speech

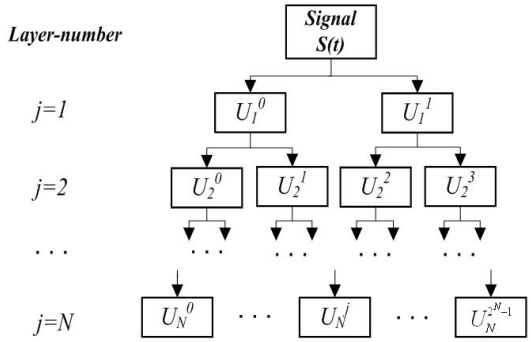


FIGURE 1. N-level wavelet packet decomposition schematic.

signal still maintains high reliability [10, 11]. DCT's orthogonal transformation is defined as follows.

$$\theta(i) = \sqrt{\frac{2}{N}} c(i) \sum_{n=0}^{N-1} s(n) \cos \frac{(2n+1)i\pi}{2N}, i = 0, 1, ..., (N-1) \tag{2}$$

Where $c(0) = \frac{1}{\sqrt{2}}$ and $c(i) = 1$.

With the FFT algorithm, DCT can achieve fast calculation, requiring only $2Nlog_2N$ multiplication / addition operations smaller than N^2, which is crucial in the specific application of the speech signal when N is relatively large.

2.3. **QR Decomposition Using Givens Rotation.** A Givens Rotation is an important kind of orthogonal transformation in linear algebra, and QRD with a Givens Rotation can reduce computational complexity and maintain numerical stability [12]. The mathematical computation of QRD can be represented as the follows.

$$A = QR \tag{3}$$

where A is a real $m \times n$ matrix with full column rank, Q is a real $m \times m$ orthogonal matrix, and R is a real upper triangular matrix of order n.

In considering a speech signal parameter matrix W then $W = QR$, it is clear that $R = Q^{-1}W$. R matrix is obtained by QRD using a Givens Rotation on the basis of the Eq. (3), so that it can reflect the rotation property of the speech signal parameter matrix W, give expression to the stability of the internal structure of speech signal and be used for feature extraction.

3. **Speech Perceptual Hashing Authentication Algorithm.** In this section, we give a detailed description of the content-based speech authentication system and the proposed hashing algorithm.

3.1. **Content-Based Speech Authentication System.** In some applications, the integrity of audio clips must be established before the signal can actually be used, i.e. one must assure that the clip stored on the cloud has not been modified or that it is not too distorted. Perceptual hashing is a compact content-based digest approach that uniquely represents a speech clip to certify content integrity. The general framework includes the following two-steps:

Step 1.: Feature extraction. Analysis and extraction of the invariant of the input speech signal against content preserving operations. The ultimate goal is to obtain the feature parameters of perceptual significance which can uniquely represent a speech clip. At present, many speech feature extraction algorithms have been proposed for speech

content authentication and identification, with features that include MFCCs [13, 14, 15], LPCCs [16, 17], a logarithmic cepstrum coefficient [18], linear spectrum frequencies (LSF) [19], and Hilbert transform spectrum estimation [20].

Step 2.: Match. The perceptual hashing sequences of the speech signal to be detected are compared with the sequences of the original speech signal. The result is used to identify the content integrity of multimedia speech information.

3.2. **Proposed Speech Perceptual Hashing Authentication Algorithm.** In sound processing, the input speech signal first needs to be pre-processed to facilitate subsequent processing, and in our method we do this by pre-emphasizing high frequency.

WPD provides a new way to reduce the speech signals overlapped degree [21] and the integrity and orthogonally of WPD assures that the speech information is completely retained, and the result of WPD in speech signal processing matches well with the signal spectrum which reflects the features and nature of the signal clearly. Given this, the proposed algorithm obtains the wavelet packet coefficients matrix of the input speech signal after pre-emphasis and wavelet packet decomposition. Next, the wavelet packet coefficients matrix is partitioned into non-overlapping and equal-sized blocks. The two-dimension DCT transformation (2D-DCT) is used to make each sub-block energy more concentrated and stable. Then we obtain feature parameter matrix of the speech signal by using QRD with Givens Rotation for each sub-block to eventually reflect the rotation property and give expression to the stability of the internal structure of speech signal. The feature parameter sequences are created to the standard deviation of each feature parameter matrix. Finally, perceptual hashing sequences are generated by quantizing the feature parameter sequences and authentication can be implemented by perceptual hashing match.

The algorithm process is shown in Fig. 2.

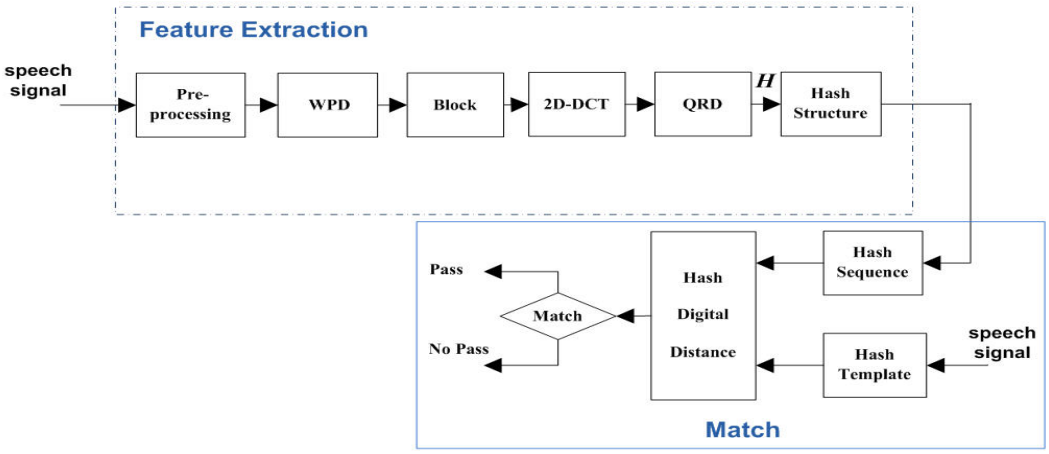


FIGURE 2. The flow chart of proposed algorithm.

Feature parameters of speech signal are extracted, and the hash modeling phase contains seven steps.

Step 1.: Pre-processing

The input speech signal first needs to undergo a pre-emphasis process to enhance the role of high frequency and the speech is digitized and converted to a general format, which is of the form of 16 bits PCM, mono and sampled at 16 kHz.

Step 2.: Wavelet packet decomposition(WPD)

According to the Eq. (1), the wavelet packet coefficients matrix $A_{M\times n}$ of wavelet reconstruction is obtained by WPD of the speech signal S after pre-emphasis, as follows.

$$A_{M\times n}=\begin{bmatrix} S_1^{u_0} & S_1^{u_1} & \dots & S_1^{u_n} \\ S_2^{u_0} & S_2^{u_1} & \dots & S_2^{u_n} \\ \vdots & \vdots & \ddots & \vdots \\ S_M^{u_0} & S_M^{u_1} & \dots & S_M^{u_n} \end{bmatrix} \tag{4}$$

where $S=\{S_j|j=1,2,...,M\}$, M is the length of speech signal S, and n is data width of WPD. In consideration of the information capacity finiteness of a single speech segment, WPD layer-number in the paper is set to 3, and $n=2^3$.

Step 3.: Matrix blocking

The matrix $A_{M\times n}$ is split into N equal and non-overlapping block around rows to generate sub-matrix group $W_N=\{W_i^{m\times n}|i=1,2,...,N,m=M/N\}$, the process of which is as follows.

$$W_N=\begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_N \end{bmatrix}, W_i=\begin{bmatrix} S^{u_0}_{(i-1)\times N+1} & S^{u_1}_{(i-1)\times N+1} & \dots & S^{u_n}_{(i-1)\times N+1} \\ S^{u_0}_{(i-1)\times N+2} & S^{u_1}_{(i-1)\times N+2} & \dots & S^{u_n}_{(i-1)\times N+2} \\ \vdots & \vdots & \ddots & \vdots \\ S^{u_0}_{i\times N} & S^{u_1}_{i\times N} & \dots & S^{u_n}_{i\times N} \end{bmatrix} \tag{5}$$

In this paper, the size m of the block matrix is in accordance with frame length, and the frame length is 32 *ms*.

Step 4.: 2D-DCT

Each sub-matrix W_i is transformed by a two-dimensional DCT in the Eq. (2), and we can obtain the new sub-matrix group $D_N=\{D_i^{m\times n}|i=1,2,...,N,m=M/N\}$, the process of which is as follows.

$$D_N=\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{bmatrix}, D_i=\begin{bmatrix} \theta^i_{1,1} & \theta^i_{1,2} & \dots & \theta^i_{1,n} \\ \theta^i_{2,1} & \theta^i_{2,2} & \dots & \theta^i_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta^i_{m,1} & \theta^i_{m,2} & \dots & \theta^i_{m,n} \end{bmatrix} \tag{6}$$

Step 5.: QRD using Givens Rotation

Each new sub-matrix D_i is transformed by QRD using Givens Rotation to generate the matrix group $R_N=\{R_i^{m\times n}|i=1,2,...,N,m=M/N\}$ in the Eq.(3), which is computed as follows.

$$R_N=\begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_N \end{bmatrix}, R_i=Q_i^{-1}D_i \tag{7}$$

where Q_i^{-1} is the inverse of matrix Q_i , a real $m\times m$ orthogonal matrix for matrix D_i in the Eq.(3), and the standard deviation of each matrix R_i is computed to obtain the feature parameter sequence $H(N,1)$ as follows.

$$H=\begin{bmatrix} std_1 \\ std_2 \\ \vdots \\ std_N \end{bmatrix}, std_i=\sqrt{\frac{1}{m\times n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{n}(R_i(j_1,j_2)-\mu_i)^2} \quad j_1,j_2\in N^+ \tag{8}$$

where $\mu_i=\frac{1}{m\times n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{n}R_i(j_1,j_2)$.

Step 6.: Hash structure

The hash bit, denoted as $ph(i) = \{ph(i)|i = 1, 2, ..., N\}$, is decided by the sign of feature sequence $H = \{H(i)|i = 1, 2, ..., N\}$ as follows.

$$ph\,(i) = \begin{cases} 1 & if H(i) > H(i-1) \quad i = 1, 2, ..., N \\ 0 & else \end{cases} \tag{9}$$

Step 7.: Hash digital distance and match

With regard to two speech clips α and β, the normalized hamming distance $D(:,:)$, also known as bit error rate (BER), is computed as follows.

$$D(PH(\alpha), PH(\beta)) = \frac{\sum_{i}^{N} |ph_\alpha(i) \oplus ph_\beta(i)|}{N} \tag{10}$$

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

L_0: Two audio clips α and β are from the same clip if

$$D(PH(\alpha), PH(\beta)) \leq \tau \tag{11}$$

L_1: Two audio clips α and β are from a different clip if

$$D(PH(\alpha), PH(\beta)) > \tau \tag{12}$$

For a certain threshold τ, a robust perceptual hashing algorithm should satisfy the following two properties:

Discrimination: If two different audio clips α and β are compared then, it is desired that $D(PH(\alpha), PH(\beta)) > \tau$ with high probability.

Robustness: If two audio clips α and β, which are the same or similar, are compared then, it is desired that $D(PH(\alpha), PH(\beta)) \leq \tau$ with high probability.

By similar means the same speech clips are preserved by content preserving operations, with little hash vector change. By setting the matching threshold τ in advance, the digital distance of two perceptual hashing sequences are compared to implement the audio object classification and identify the content integrity of speech multimedia information.

4. Experimental Results and Analysis.

4.1. Experimental Environment.

The speech data used in the experiment is from TIMIT and TTS speech library composed of different speech recorded by the Chinese men and women and English men and women. Every speech segment is converted to a general WAV format with the same length 4 s, which is of the form of 16 bits PCM, mono and sampled at 16 kHz. The speech library in this paper is a total of 1280 speech clips consist of 640 English speech clips and 640 Chinese speech clips.

Experimental hardware platform: Intel Celeron (R) (R) E3300, 2G, 2.5 GHz, software environment is the MATLAB R2012b under Windows XP operating system.

4.2. Discrimination Analysis.

Discrimination reflects the algorithm performance and often is proved by computing and comparing the $BERs$ and FAR of different algorithms.

A. Bit error rate (BER). BER has been widely accepted and often used to hash a digital distance measure in the binary form, and is the basic measure of a perceptual hashing algorithm's performance. BER points out the error bits percentage in the total number of bits, the normalized hamming distance calculated by the following formula.

$$BER = \frac{\sum_{i=1}^{N} \left(|ph_\alpha(i) \oplus ph_\beta(i)| \right)}{N} \tag{13}$$

The $BERs$ of different content speech clips are usually normally distributed to make a statistical analysis on $BERs$ of the 1280 speech clips, resulting in 818560 $BERs$, which eventually becomes normally distributed as shown in Fig. 3.

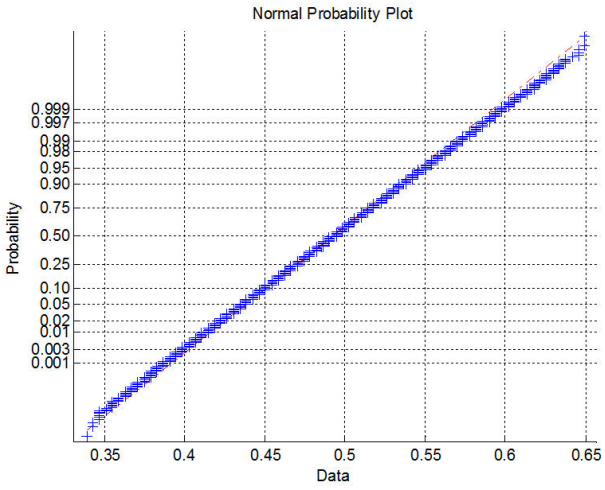


FIGURE 3. BER normal distribution diagram.

According to the central limit theorem of De Moivre-Laplace, hamming distance approximately obeys normal distribution. When adopting bit error rate as the distance measure, the $BERs$ approximately obey a normal distribution $(\mu = p, \sigma = \sqrt{p(p-1)/N})$ [22], where N is the size of perceptual hashing sequences and its value equals 256 and in the paper, μ and σ are the expected value and the standard deviation. Normal distribution parameter values in theory and in the actual experiment are as shown in Table 1.

TABLE 1. Normal distribution parameters

Type	Theoretical values		Experimental results	
Parameters	μ	σ	μ_0	σ_0
Values	0.5	0.0316	0.4927	0.0341

As can be seen from Table 1, the normal distribution parameter values in the experimental results approximate the theoretically calculated parameter values. Therefore, the hash sequence possesses randomness and collision resistance.

B. False accept rate (FAR). FAR, the probability of different multimedia object judged as the same content and accepted by the system, can be obtained from the $BERs$ of different content speech clips. FAR is computed as follows.

$$FAR(\tau) = \int_{-\infty}^{\tau} f(\alpha\,|\mu, \sigma) d\alpha = \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\alpha-\mu)^2}{2\sigma^2}} d\alpha \tag{14}$$

Accordingly, the false reject rate (FRR) is denoted as follows.

$$FRR(\tau) = 1 - \int_{-\infty}^{\tau} f(\alpha\,|\mu, \sigma) d\alpha = 1 - \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\alpha-\mu)^2}{2\sigma^2}} d\alpha \tag{15}$$

TABLE 2. FAR of the proposed algorithm

τ	FAR
0.20	4.8389e-18
0.25	5.7421e-13
0.30	8.2513e-09

As can be seen from Table 2, when the matching threshold τ is equal to 0.30, there are approximately eight speech clips misjudged among the 10^9 speech clips demonstrating that the algorithm maintains acceptable recognition capability.

TABLE 3. FAR of the different algorithms

Parameters	FAR			
	Proposed algorithm	Jiao et al.'s algorithm [5]	Huang et al.'s algorithm [8]	Jiao et al.'s algorithm [19]
N=1024, τ=0.30	5.5891e-33	9.9034e-28	-	-
N=512, τ=0.30	2.9252e-17	-	6.6981e-15	5.2604e-17

As can be seen from Table 3, when the matching threshold τ is set to 0.30, the FAR in the proposed algorithm is lower than the $FARs$ in Jiao et al.'s algorithm [5], Huang et al.'s algorithm [8] and Jiao et al.'s algorithm [19].

C. Entropy rate (ER). FAR is highly affected by the size of the hash sequence and the same algorithm can possess a different FAR due to the different values of N (see Table 3). Hence, it is one-sided and biased to compare the algorithms performance by only using FAR and BER. However, an ER with clear upper and lower limit values constitutes a unit of information and is not affected by the size of hash sequencing, reflecting the algorithm performance and is used in joint evaluation index discrimination and compaction. ER is computed as follows.

$$ER = -p \log_2 p - (1-p) \log_2 (1-p) \tag{16}$$

where p is the transition probability between two hash sequence signs and is computed as follows.

$$p = \frac{1}{2}\left(\sqrt{\frac{\sigma_0^2 - \sigma^2}{\sigma_0^2 + \sigma^2}} + 1\right) \tag{17}$$

TABLE 4. ER of the different algorithms

τ	Proposed algorithm	Jiao et al.'s algorithm [5]	Nouri et al.'s algorithm [6]	Huang et al.'s algorithm [8]	Jiao et al.'s algorithm [19]
ER	0.9445	0.9158	0.8266	0.8992	0.9367

As can be seen from Table 4, ER of the proposed algorithm is larger than the Jiao et al.'s algorithm [5], Nouri et al.'s algorithm [6], Huang et al.'s algorithm [8] and Jiao et al.'s algorithm [19] demonstrating that the proposed algorithm has a better recognition performance.

4.3. **Robustness Analysis.** The speech in the library uses the following content preserving operations:

(1) Decrease volume: volume decreased by 50%;

(2) Increases volume: volume increases by 50%;

(3) Resampling 8-16: sampling frequency reduced to 8 kHz, and up to 16 kHz;

(4) Resampling 32-16: sampling frequency up to 32 kHz, and reduced 16 kHz;

(5) Echo addition: stack attenuation was 60%, the time delay for 300 *ms*, the echo of the initial intensity of 20% and 10% respectively;

(6) Narrowband noise: with the center frequency distribution in 0∼4 kHz narrowband Gaussian noise;

(7) Butterworth filter: using a twelve order Butterworth low-pass filter with cut-off frequency of 3.4 kHz;

(8) FIR filter: using a twelve order FIR low-pass filter with cut-off frequency of 3.4 kHz;

Table 5. The proposed algorithm's average $BERs$

Operating means	Average BER
Decrease volume	7.8436e-04
Increases volume	0.0080
Resampling 8-16	0.0037
Resampling 32-16	0.0485
Echo addition	0.1140
Narrowband noise	0.0624
Butterworth filter	0.0876
FIR filter	0.0934

As we can see from Table 5, decrease volume, increase volume and resampling do not change the vocal tract model, the energy ratio and feature of each frame is less affected and their average $BERs$ are almost zero. Consequently the proposed algorithm's robustness is best with the decrease volume, increase volume and resampling operations. For different types of the low-pass filter operations, the algorithm still holds a low average BER because the internal structure of a speech signal is much less vulnerable against various kinds of low-pass filter operations. Moreover, for this algorithm, the average BER of any two speech clips which have the same perceptual content is below 0.2. In addition the proposed algorithm in this paper exhibits high robustness for a variety of content preserving operations, especially the volume adjustment and resampling operations.

As can be seen in Fig.4, the proposed algorithm's average $BERs$ are far lower than the average $BERs$ of Huang et al.'s algorithm [8] algorithm when subjected to the above several kinds of attacks.

This paper totally get 818560 $BERs$ data by conducted pairwise compare between perceptual hash values from 1280 different speech clips. Combining the $BERs$ in Table 5, with the data in Jiao et al.'s algorithm [5] and Huang et al.'s algorithm [8], and drawing the FRR (false reject rate) and FAR curve, yielded good results as shown in Fig. 5(a), Fig. 5(b), Fig. 5(c) and Fig. 5(d).

As can be seen from the Fig. 5(a), the $FAR - FRR$ curve has no cross in the picture, indicating that the proposed algorithm has both good distinction and robustness and can accurately identify content preserving operations and malicious operations. When comparing Fig. 5(a) with Fig. 5(b), Fig. 5(c) and Fig. 5(d), the $FAR - FRR$ curve in Jiao et al.'s algorithm [5] approximately cross, the $FAR - FRR$ curve in Huang et al.'s algorithm [8] and the LPC algorithm are crossed in the figure, but not crossed for the

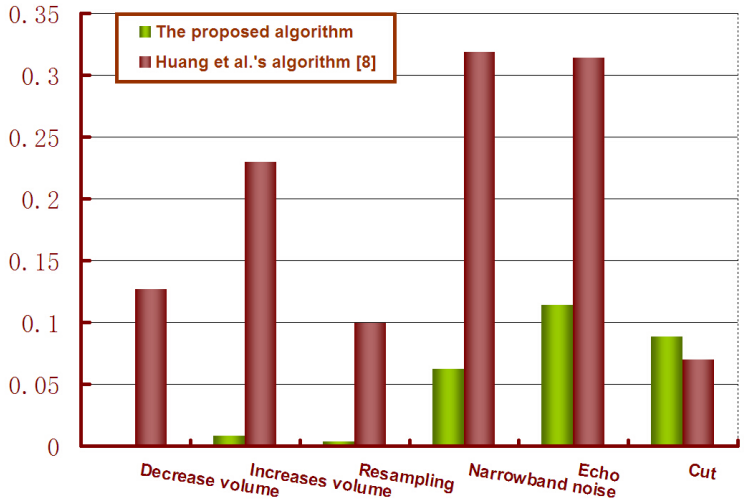
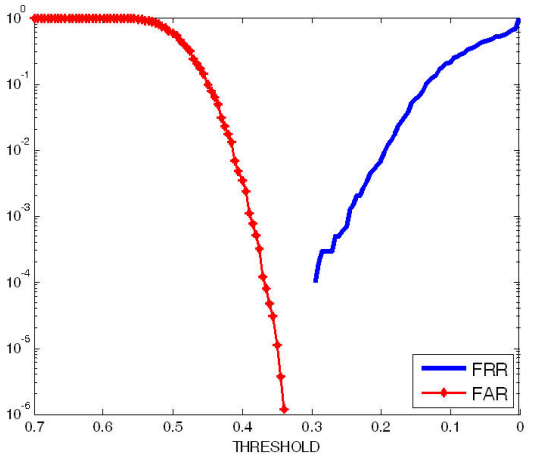
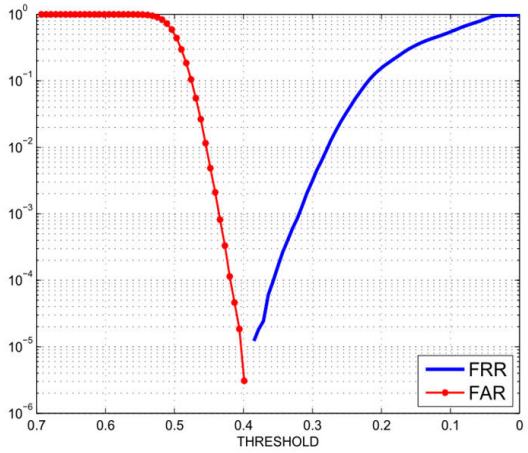


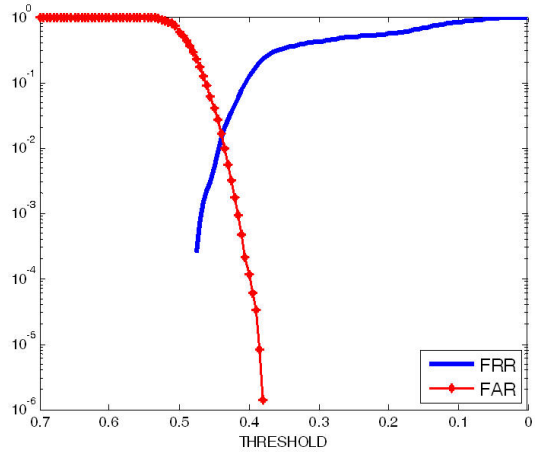
FIGURE 4. The average $BERs$ of proposed algorithm and Huang et al.'s algorithm [8].



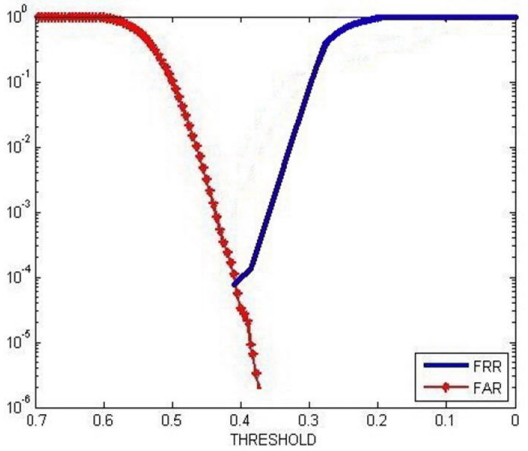
(a)$FAR - FRR$ curve of proposed algorithm.



(b)$FAR - FRR$ curve in Jiao et al.'s algorithm [5].



(c)$FAR - FRR$ curve of LPC algorithm.



(d)$FAR - FRR$ curve in Huang et al.'s algorithm [8].

FIGURE 5. $FAR - FRR$ curve of different algorithms.

proposed algorithm. Coupled with Table 5, it is apparent that the robustness of algorithm in the paper against content preserving operation is better.

4.4. **Efficiency Analysis.** In order to measure the algorithm complexity and computational efficiency, the paper randomly extracts 100 speech clips, record the run time of the proposed algorithm, and then makes comparisons with the Chen et al.'s algorithm [7].

Table 6. Operation Time

Algorithm	Proposed algorithm	Chen et al.'s algorithm [7]
Parameters	Run time (s)	
File length (s)	4	4
Platform dominant frequency	2.5GHz	3.3GHz
Feature extraction	36.39	78.24
Match	0.42	11.84
Total	36.81	90.08

Table 6, shows that the algorithm's computing efficiency in this paper is better than the Chen et al.'s algorithm [7] and the proposed algorithm complexity is less and, more effective and more elegant.

Table 7. The size of hash sequence

Algorithm	Proposed algorithm	Jiao et al.'s algorithm [5]	Huang et al.'s algorithm [8]	Jiao et al.'s algorithm [19]
Size	$62.5/s$	$314.3/s$	$125/s$	$512/s$

As can be seen in Table 7, the average size of perceptual hashing sequences per second speech clip is 62.5, far smaller than the size in the Jiao et al.'s algorithm [5], Huang et al.'s algorithm [8] and Jiao et al.'s algorithm [19], and therefore the proposed algorithm possesses the features of good compaction and uses relatively smaller authentication data. In conclusion, the algorithm in the paper satisfies the requirements of real-time speech communication quality, and can be positively applied to limited resource speech communication terminal design in the mobile computing environment.

5. **Conclusions.** The paper proposes a new speech perceptual hashing authentication algorithm based on WPD. The algorithm extracts the features of the reflecting internal structure of speech signal, which is demonstrated that is an excellent speech signal feature. Finally, experiment results show that the algorithm has strong robustness in content preserving operations, good compaction and discrimination features, and therefore satisfies the requirements of real-time speech authentication with high certification efficiency.

Further research is planned is improve the algorithms efficiency with focus on accurate positioning and approximate recovery issues.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No. 61363078), the Natural Science Foundation of Gansu Province of China (No. 1212RJZA006, No. 1310RJYA004). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

[1] U. Greveler, B. Justus, and D. Loehr, Multimedia Content Identification Through Smart Meter Power Usage Profiles, *Proc. of 5th International Conference on Computers, Privacy and Data Protection*, Springer, 2012.

[2] T. Shibuya, M. Abe, and M. Nishiguchi, Audio fingerprinting robust against reverberation and noise based on quantification of sinusoidality, *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, 2013.

[3] A. Hadmi, W. Puech, B. Ait Es Said, and A. Ait Ouahmanb, A robust and secure perceptual hashing system based on a quantization step analysis, *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 929-948, 2013.

[4] Z. Q. Liu, Q. Li, J. R. Liu, and X. M. Niu, A novel evaluation method for perceptual hash algorithm in authentication mode, *Proc. of Fourth International Conference on Digital Image Processing (ICDIP 2012)*, 2012.

[5] Y. H. Jiao, M. Y. Li, Q. Li, and X. M. Niu, Key-dependent compressed domain audio hashing, *Proc. Of Eighth International Conference on Intelligent Systems Design and Applications*, vol.3, pp. 29-32, 2008.

[6] M. Nouri, N. Farhangian, Z. Zeinolabedini, and M. Safarinia, Conceptual authentication speech hashing base upon hypotrochoid graph, *Proc. of IEEE Sixth International Symposium on Telecommunications (IST)*, pp. 1136-1141, 2012.

[7] N. Chen, H. D. Xiao, and J. Zhu, Robust audio hashing scheme based on cochleagram and cross recurrence analysis, *Electronics Letters*, vol. 49, no. 1, pp. 7-8, 2013.

[8] Y. B. Huang, Q. Y. Zhang, and Z. T. Yuan, Perceptual Speech Hashing Authentication Algorithm Based on Linear Prediction Analysis, *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 12, no. 4, pp. 3214-3223, 2014.

[9] R. Sharma, and V. P. Pyara, A Robust Denoising Algorithm for Sounds of Musical Instruments Using Wavelet Packet Transform, *Circuits and Systems*, vol. 4, no. 7, pp. 459-465, 2013.

[10] C. Yin, and S. J. Yuan, A Novel Algorithm for Embedding Watermarks into Audio Signal Based on DCT, *Proc. of the International Conference on Information Engineering and Applications (IEA 2012)*, vol. 220, pp. 683-688, 2013.

[11] V. R. Balaji, and S. Subramanian, A Novel Speech Enhancement Approach Based on Modified DCT and Improved Pitch Synchronous Analysis, *American Journal of Applied Sciences*, vol. 11, no. 1, pp. 24-37, 2013.

[12] F. Merchant, A. Chattopadhyay, G. Garga, S. K. Nandy, R. Narayan, and N. Gopalan, Efficient QR Decomposition Using Low Complexity Column-wise Givens Rotation, *Proc. of IEEE VLSI Design and 2014 13th International Conference on Embedded Systems*, pp. 258-263, 2014.

[13] V. Panagiotou, and N. Mitianoudis, PCA summarization for audio song identification using Gaussian Mixture models, *Proc. of IEEE 18th International Conference on Digital Signal Processing*, pp. 1-6, 2013.

[14] S. Gaikwad, B. Gawali, P. Yannawar, and S. Mehrotra, Feature extraction using fusion MFCC for continuous marathi speech recognition, *Proc. of India Conference (INDICON), 2011 Annual IEEE*, pp. 1-5, 2011.

[15] N. Chen, H. D. Xiao, and W. G. Wan, Audio hash function based on non-negative matrix factorisation of mel-frequency cepstral coefficients, *Information Security, IET*, vol. 5, no. 1, pp. 19-25, 2011.

[16] P. Lotia, and D. M. R. Khan, Significance of Complementary Spectral Features for Speaker Recognition, *IJRCCT*, vol. 2, no. 8, pp. 579-588, 2013.

[17] O. Chia Ai, M. Hariharan, and S. Yaacob, Classification of speech dysfluencies with MFCC and LPCC features, *Expert Systems with Applications*, vol. 39, no. 2, pp. 2157-2165, 2012.

[18] H. Őzer, B. Sankur, and N. Memon, Perceptual audio hashing functions, *EURASIP Journal on Applied Signal Processing*, vol. 2005, no.12, pp.1780-1793, 2005.

[19] Y. H. Jiao, L. P. Ji, and X. M. Niu, Robust speech hashing for content authentication, *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 818-821, 2009.

[20] H. Zhao, H. Liu, K. Zhao, and Y. Yang, Robust Speech Feature Extraction Using the Hilbert Transform Spectrum Estimation Method, *International Journal of Digital Content Technology and its Applications*, vol. 5, no. 12, pp. 85-95, 2011.

[21] J. L. Liu, F. Q. Yu, and Y. Chen, Speech Separation Based on Improved Fast ICA with Kurtosis Maximization of Wavelet Packet Coefficients, *New Perspectives in Information Systems and Technologies*, vol. 1, pp. 43-50, 2014.

[22] J. Haitsma, and T. Kalker, A highly robust audio fingerprinting system, *Proc. of ISMIR. Vol. 2002*, pp. 107-115, 2002.