An efficient speech perceptual hashing authentication algorithm based on DWT and symmetric ternary string

Zhang Qiuyu*, Xing Pengfei, Huang Yibo, Dong Ruihong and Yang Zhongping

School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China

Email: zhangqylz@163.com
Email: xingpengfei0202@126.com
Email: huangyibo1982@163.com
Email: dongrh@lut.cn
Email: zpyang90@126.com

*Corresponding author

Abstract: According to the situation that speech perceptual hashing methods are not appropriated for real-time speech content authentication in mobile computing environment, a novel DWT-based perceptual hashing algorithm, which uses a combination of time-domain and frequency-domain features, was proposed to protect the speech data in the cloud. Firstly, by discrete wavelet transform (DWT), a new signal in frequency-domain is generated from the original speech signal after pre-processing and intensity-loudness transform (ILT). Secondly, coefficients of low frequency wavelet decomposition are partitioned into equal-sized and non-overlapping blocks, and logarithmic short-time energy of each block is computed to obtain speech signal’s features in frequency-domain. Finally, combined with spectral flux features (SFF) of speech signal in time-domain, a ternary perceptual hashing sequence is created. Experiment results illustrate that ternary form is better to stand for hash digest than binary form, the proposed algorithm has a good robustness against content preserving operations, discrimination, good compaction and high efficiency, and detects the tamper localisation as well.

Keywords: speech perceptual authentication; perceptual hashing; discrete wavelet transform; DWT; symmetric ternary string; tamper localisation.


Biographical notes: Zhang Qiuyu is a researcher/PhD supervisor, and has graduated from Gansu University of Technology in 1986, and then worked at School of Computer and Communication in Lanzhou University of Technology. He is the Vice Dean of Gansu Manufacturing Information Engineering Research Center, a CCF senior member, and 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.

Xing Pengfei received his BSc in Communication Systems and Communication Theory from Henan University of Science and Technology, Henan, China in 2008. His research interests include audio signal processing and application, multimedia authentication techniques.

Huang Yibo received his BSc and MSc in Communication Systems and Communication Theory from Lanzhou University of Technology, Lanzhou, China in 2007 and 2011, respectively, and is currently pursing his PhD in the Department of College of Electrical and Information Engineering at Lanzhou University of Technology. His research interests include information security, steganalysis speech signal processing.

Dong Ruihong is a vice researcher, and he worked at School of Computer and Communication in Lanzhou University of Technology. His research interests include network and information security, information hiding and steganalysis, computer network.

Yang Zhongping received his BSc in Communication Systems and Communication Theory from Lanzhou University of Technology, Lanzhou, China in 2008. His research interests include: audio signal processing and application, multimedia authentication techniques.

This paper is a revised and expanded version of a paper entitled ‘An efficient time-frequency domain speech perceptual hashing authentication algorithm based on discrete wavelet transform’ presented at 3PGCIC-2014: 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Guangzhou, China, November 8–10, 2014.

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, result 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 for malicious intent which include illegal eavesdropping and tampering. 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 etc. 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 (Greveler et al., 2012; Shibuya et al., 2013).

Unlike cryptographic hashing, perceptual hashing involves the one-way mapping of a multimedia digital representation to a perceptual hashing digest, that is, a compact
An efficient speech perceptual hashing authentication algorithm

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 and broadband audio signal in cloud storage, so it is widely used in the field of information security (Hadmi et al., 2013; Liu et al., 2012).

At present, many speech fingerprinting extraction algorithms have been proposed for speech content authentication and identification, and mostly make optimise scheme based on human auditory system. Feature extraction methods mainly contained logarithmic cepstrum coefficient (Ozer et al., 2005), linear spectrum frequencies (Nouri et al., 2012), MFCCs (Gu et al., 2010; Panagiotou and Mitianoudis, 2013; Gaikwad et al., 2011) and LPCCs (Lotia and Khan, 2013). Some other methods are also widely adopted, such as Hilbert-Huang transform (Zhang et al., 2014), temporal modulation normalisation (Lu et al., 2011) and bark-bands energy (Mathieu and Geoffroy, 2011), etc. Jiao et al. (2009) proposes a perceptual hashing algorithm based on linear spectrum frequencies, and the algorithm has robustness and randomness but not good compactness. Chen et al. (2013) treats the cochleagram of speech data as an image, from which speeded up robust features are extracted as essential features. The results illustrate that the method exhibits good performance but a poor efficiency, which is not appropriated for real-time speech authentication. Huang et al. (2014) tries to improve the algorithm based on linear forecast analysis, called E+LPC algorithm, experiments of which show that the algorithm has a high efficiency but less robustness.

In order to satisfy the requirements of real-time speech content authentication in the speech communication terminal designed with limited resources or cloud computing, a speech perceptual hashing authentication algorithm based on discrete wavelet transform (DWT) was presented in this paper. The algorithm extracts the features of logarithmic short-time energy (LSE) in the frequency-domain and spectral flux in the time-domain to generate the ternary perceptual hashing digest. Finally, authentication can be implemented by perceptual hashing match.

2 Discrete wavelet transform

DWT uses the idea of multi-resolution, overcomes the shortcomings of fast Fourier transform and short-time Fourier transform and gives an accurate representation of speech signal’s local details. Its time-frequency window is not fixed, which is appropriated for non-stationary signals analysis including speech signal (Sharma and Pyara, 2012; Ali et al., 2014).

2.1 Wavelet transformation definitions

For given a wavelet function \( \psi(t) \) which satisfies certain conditions, the continuous wavelet transform of time series \( S(t) \in L^2(R) \) is computed as follows.

\[
W_S(a,b) = |a|^{-1/2} \int_{-\infty}^{+\infty} S(t) \overline{\psi\left(\frac{t-b}{a}\right)} dt
\] (1)

where \( a \) represents a time dilation, \( b \) a time translation, and the bar stands for complex conjugate. Generally, time series is discrete signal, and discrete wavelet can be obtained
by the discretisation of parameters $a$ and $b$. In particular, we assume that $a$ and $b$ have the form $a = 2^m$ and $b = n2^m$, and dyadic wavelet series can be calculated out. Thus, discrete time sequence $S(t)$ can be formalised as follows.

$$S(t) = \sum_{m,n \in \mathbb{Z}} W_{m,n} \psi_{m,n}(t)$$  \hspace{1cm} (2)

where $W_{m,n} = \langle S, \psi_{m,n} \rangle$, $m, n \in \mathbb{Z}$, equation (2) is called as the DWT of signal $S(t)$.

### 2.2 Wavelet decomposition

For a given subspace sequence $\{W_j \mid j \in \mathbb{Z}\}$, we assume that $W_j$ is the orthogonal component of $V_j$ in the multi-resolution.

$$V_{j+1} = V_j \oplus W_j, W_j \perp V_j$$  \hspace{1cm} (3)

The scaling function $\varphi(t)$ and wavelet function $\psi(t)$ satisfy the two-scale, as in equations (4) and (5):

$$\varphi(t) = \sum_{n \in \mathbb{Z}} h(n) \varphi(2t - n), h(n) \in l^2(\mathbb{Z})$$  \hspace{1cm} (4)

$$\psi(t) = \sum_{n \in \mathbb{Z}} (-1)^n g(n) \varphi(2t - n), g(n) \in l^2(\mathbb{Z})$$  \hspace{1cm} (5)

where $g(n)$ is the low pass filters group, $h(n)$ is high-pass filter group. Let’s assume that $L^n_h$ and $H^n_g$ correspond to projection coefficients of the subspace $V_n$ and $W_n$, respectively. Speech signal $S(t)$ conducts wavelet decomposition, as follows:

$$\begin{cases} 
L^n_{h,m} = \sum_{m,n,k \in \mathbb{Z}} h(k - 2m)L^{n-1,m}_h \\
H^n_{g,m} = \sum_{m,n,k \in \mathbb{Z}} g(k - 2m)H^{n-1,m}_g 
\end{cases}$$  \hspace{1cm} (6)

Discrete wavelet decomposition is shown in Figure 1.

**Figure 1** Discrete wavelet decomposition schematic
3 Proposed speech perceptual hashing authentication system

3.1 An overview of perceptual hashing authentication system

Recall that perceptual hashing can be seen as a short summary of a speech object. Therefore a perceptual hashing function \( PH \) should map an audio object \( S \), consisting of a large number of bits, to a perceptual hashing of only a limited number of bits. In some applications, the integrity of speech 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. A perceptual hashing authentication system provides the ability to identify short, unlabelled speech clips in a fast and reliable way.

For given a speech clip \( S \), let \( S_1 \) denotes a modified recording of this clip which is perceptually the same as \( S \). Let \( S_2 \) to be a perceptually different audio clip. \( PH(\cdot) \) represents a hash function and takes as an input the excerpt speech signal. Therefore, an attacker cannot forge the speech signature, i.e., perceptual hashing digest. Our aim is to achieve the following probabilities.

\[
Pr[PH(S) = PH(S_1)] \approx 1
\]

\[
Pr[PH(S) = PH(S_2)] \approx 0
\]

In content-based speech authentication system, application to certify content integrity is done in two-steps:

Step 1 Feature extractions. The speech file is divided into segments and from every speech clip perceptively relevant speech features are extracted. This is done by means of perceptual hashing. The ultimate goal is to obtain the feature parameters of perceptual significance which can uniquely represent a speech clip to assure that the clip stored on the cloud has not been modified or is not too distorted (see Figure 2).

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 (see Figure 3).

Figure 2 Feature extractions (see online version for colours)
3.2 Proposed speech perceptual hashing authentication algorithm

Human auditory system has better anti-interference than speech processing technology, and the combination of acoustics physiology computational auditory model with the front-end of speech authentication system can improve the anti-interference ability of perceptual feature sequences. The relationships between loudness and intensity of speech signal are nonlinear, and loudness response curve is often used to mimic human auditory processing principle (Hartmann, 2013). Considered of this, the proposed algorithm firstly performs the intensity-loudness transform (ILT) to the speech signal after pre-processing. The nonlinear relationships between loudness $L$ and intensity $I$ are shown as follows.

$$L = I^n$$ (9)

It is well-known that feature information is mainly concentrated in the low-frequency parts of speech signal. Decomposition low-frequency coefficients $L_h$ is extracted from speech signal after a pre-emphasis, ILT and DWT. It is generally known that short-time average energy can well reflect amplitude variation of speech signal, and at high signal noise ratio (SNR) which is not only an efficient feature to distinguish between speech and noise signal, but LSE is consistent with the human auditory system. So we extract block features of decomposition low-frequency coefficients $L_h$ by calculating out block LSE (Jalil et al., 2013).

LSE is shown, as follows.

$$g(k) = 10 \log_{10} \sum_{m=1}^{N-1} s^2(m)$$ (10)

where $k$ stands for the current frame, $N$ is the length of Hamming window function.

DWT has low time resolution in low frequency, which can be prone to cause feature information loss in time-domain and even misunderstanding. Due to the loss of high frequency information, feature information cannot well reflect full information of speech signal. In order to make up the shortfall, we further extract time-domain features, i.e., spectral flux features (SFF). SFF reflect speech signal’s envelop information, which will be badly affected by the change of wave form. In order to ensure stability of SFF, speech signal conducts a band-pass filtering (BPF) process in advance. The algorithm process is shown as in Figure 4.
Figure 4  Flow chart of proposed algorithm (see online version for colours)

Feature parameters of speech signal are extracted, and the hash modelling phase contains nine steps.

Step 1  Pre-processing. The input speech signal \( S(t) \) firstly needs to conduct a pre-emphasis process to enhance the role of high frequency and its length is \( K \).

Step 2  ILT. Speech signal after pre-emphasis process, denoted by \( S'(t) \), is converted into signal \( L(t) \) by ILT, where \( \alpha = 0.33 \).

Step 3  DWT. The low-frequency wavelet decomposition coefficients \( L_h = \{L_h^i \mid i = 1, 2, ..., n\} \) are obtained by DWT of signal \( L(t) \), as in Figure 1.

Step 4  Block. The low-frequency coefficients \( L_h \) are split into \( N \) equal and non-overlapping block to create coefficients matrix \( T \) as in equation (11), and \( M \) is the size of block.

\[
T = \begin{bmatrix}
L_h^1 & L_h^2 & \cdots & L_h^M \\
L_h^{M+1} & L_h^{M+2} & \cdots & L_h^{2M} \\
\vdots & \vdots & \ddots & \vdots \\
L_h^{(N-1)\times M+1} & L_h^{(N-1)\times M+2} & \cdots & L_h^{N\times M}
\end{bmatrix}
\]  

Step 5  Extract LSE features. Compute LSE of each line in matrix \( T \) to get LSE feature sequences, as in equation (10), denoted as \( H_1 = \{g(k) \mid k = 1, 2, ..., N\} \).

\[
H_1 = \begin{bmatrix}
g(1) \\
g(2) \\
\vdots \\
g(N)
\end{bmatrix}, \quad g(k) = 10\log_{10} \sum_{m=1}^{M} T(k, m)
\]
Step 6  BPF. Signal $S'(t)$ conducts a band-pass filtering process to get signal $B = \{ B(t) | t = 1, 2, \ldots, K \}$, which is a 60 to 4000 Hz pass-band filter in this paper.

Step 7  Extract SFF features. The signal $B$ is split into $N$ equal and non-overlapping blocks, and SFF for each block is computed to get SFF feature sequences, denoted as $H_2 = \{ SFF(k) | k = 1, 2, \ldots, N \}$.

Step 8  Hash structure. The hash bit, denoted as $ph_j(i) = \{ ph_j(i) | i = 1, 2, \ldots, N, j = 1, 2 \}$, is decided by the sign of feature sequences $H_j = \{ H_j(i) | i = 1, 2, \ldots, N, j = 1, 2 \}$, as in equation (13). The integrated hash feature sequences of speech signal $S(t)$ is $ph(m)$, and $ph = [ph_1, ph_2]$, where $m = 1, 2, \ldots, 2N$.

$$\begin{array}{ll}
ph_j(i) = & \begin{cases} 
1 & \text{if } H(i)^2 - H(i-1) \times H(i+1) > 0 \\
0 & \text{else if } H(i) - H(i-1) > 0 \\
-1 & \text{otherwise}
\end{cases} \\
\end{array}$$

(13)

Step 9  Hash digital distance and matching. With regard to two speech clips $s_1$ and $s_2$, the hashing digital distance $DH(\cdot, \cdot)$ is computed as follows.

$$DH(PH(s_1), PH(s_2)) = \sum_m |ph_{s_1}(m) - ph_{s_2}(m)|$$

(14)

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

$P_0$: Two audio clips $s_1$ and $s_2$ are from the same clip if

$$DH(PH(s_1), PH(s_2)) \leq \tau$$

(15)

$P_1$: Two audio clips $s_1$ and $s_2$ are from different clip if

$$DH(PH(s_1), PH(s_2)) > \tau$$

(16)

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

- **Robustness**: If two audio clips $s_1$ and $s_2$ are the same or similar, are compared then, it is desired that $DH(PH(s_1), PH(s_2)) \leq \tau$ with high probability.

- **Discrimination**: If two different audio clips $s_1$ and $s_2$ are compared then, it is desired that $DH(PH(s_1), PH(s_2)) > \tau$ 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 $\tau$ 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 are from TIMIT and TTS speech library, which composed of different speech recorded both in Chinese and English by men and women. Every speech clip is converted to a general WAV format with the same length four seconds, 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 1,280 speech clips consist of 640 English speech clips and 640 Chinese speech clips. All experiments have been carried out under two different environments, shown as follow:

1 experimental hardware platform: Intel Celeron (R) E3300, 2G, 2.5 GHz, software environment is the MATLAB R2012b under Windows 7 operating system
2 experimental hardware platform: Inter (R) Core (TM) i3-2120, 4G, 3.30 GHz, software environment is the MATLAB R2012b under Windows 7 operating system.

4.2 Bit error rate

In binary system, bit error rate (BER) has been widely accepted and often used as hashing digital distance measurement and it is the basic measurement of perceptual hashing algorithm’s performance. BER is pointed out that error bits percentage in the total number of bits, namely, the normalised hamming distance. For ternary sequence, BER can still be used as basic measurement to evaluate perceptual hashing algorithm’s performance. The normalised hamming distance calculated by the following formula.

\[
BER = \frac{\sum_{m=1}^{2N} |ph_{s1}(m) - ph_{s2}(m)|}{2N}
\]  

where \(ph_{s1}\) and \(ph_{s1}\) correspond to the perceptual hashing values generated by speech clip \(s1\) and \(s2\).

4.3 Robustness test and analysis

4.3.1 Different bases system robustness

At present, all proposed perceptual hashing authentication algorithms are based on binary system, and their perceptual hashing sequences are in the form of binary bits. Yet, hash binary sequences either ‘0’ or ‘1’ can produce a high probability of mutation against content preserving operation, and BER is lack of stability. Especially for a single speech clip, the probability of false accept rate (FAR) and false reject rate (FRR) is higher. In the paper, we try to build different bases system and then compare binary system with ternary system. The speeches in the library against content preserving operations are as shown in Table 1.
Table 1  Content preserving operations

<table>
<thead>
<tr>
<th>Operating means</th>
<th>Level</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decrease volume</td>
<td>50%</td>
<td>V↓</td>
</tr>
<tr>
<td>Increases volume</td>
<td>50%</td>
<td>V↑</td>
</tr>
<tr>
<td>Resampling1</td>
<td>8 kHz→16 kHz</td>
<td>R. 8→16</td>
</tr>
<tr>
<td>Resampling2</td>
<td>32 kHz→16 kHz</td>
<td>R. 32→16</td>
</tr>
<tr>
<td>Echo addition</td>
<td>60%, 300 ms</td>
<td>E.A</td>
</tr>
<tr>
<td>Narrowband noise</td>
<td>50 db</td>
<td>G.N</td>
</tr>
<tr>
<td>Low-pass Filter1</td>
<td>3.4 kHz Butterworth filter</td>
<td>B.W</td>
</tr>
<tr>
<td>Low-pass Filter2</td>
<td>3.4 kHz FIR filter</td>
<td>F.I.R</td>
</tr>
<tr>
<td>MP3 Coding1</td>
<td>32 kbps</td>
<td>M.32</td>
</tr>
<tr>
<td>MP3 Coding2</td>
<td>192 kbps</td>
<td>M.192</td>
</tr>
</tbody>
</table>

Compute the BER of 1,280 different speech clips of the binary and ternary system against content preserving operations respectively, shown in Table 1. For the binary hash sequences, the occurrence of ‘0’ and ‘1’ are theoretically equal to probability, and the average hamming distance of different content speech is $\frac{N}{2}$. For the ternary hash sequences, the occurrence of ‘–1’, ‘0’ and ‘1’ are also theoretically equal to probability, and the average hamming distance of different content speech is $\frac{8N}{9}$. To make comparisons with binary and ternary system in the same scope, BER in ternary system multiplies by the scale factor $\theta$ to quantise it in the range of 0-1, where $\theta = \frac{N/2}{8N/9}$. In this paper, the markings of BER data sources are shown in Table 2.

Table 2  BER data source

<table>
<thead>
<tr>
<th>Label</th>
<th>Ternary system (O)</th>
<th>Ternary system (Q)</th>
<th>Binary system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data source</td>
<td>Original ternary</td>
<td>Quantised ternary</td>
<td>Binary</td>
</tr>
</tbody>
</table>

Compute the average BER ($ABER$), range ($RBER$) and standard deviation ($SBER$) of BER in different system according to Table 2 respectively and make comparisons with ternary system (Q) and binary system as shown as Table 3. Compute the average BER ($ABER$), range ($RBER$) and standard deviation ($SBER$) of BER in different system respectively (see Table 2), and compare ternary system (Q) with binary system, shown in Tables 3 and 4. $ABER$, $RBER$ and $SBER$ in ternary system (Q) are almost less than those in binary system (see Tables 3 and 4). It shows that the BER data of perceptual hashing ternary sequences has higher concentration structure and is more stable. The reason mainly lies in hash binary sequences either ‘0’ or ‘1’ can produce a high probability of mutation for content preserving operation, and BER is lack of stability. Due to existing three sequences: ‘1’, ‘0’ and ‘–1’, hash ternary sequence slows down the mutation and reduces the error probability of FAR and FRR against content preserving operations.

BER data of 818,560 are obtained to make comparisons between any two different perceptual hashing sequences of 1,280 speech clips in different system respectively (see Table 2). Combining with the BER data against content preserving operations, FRR and FAR curve of ternary system (Q) and binary system can be drawn, shown in Figures 5 and 6.
An efficient speech perceptual hashing authentication algorithm

### Table 3  
BER data in ternary system (Q)

<table>
<thead>
<tr>
<th>Operating means</th>
<th>Ternary system (Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>ABER</td>
</tr>
<tr>
<td>V. ↓</td>
<td>0.0057</td>
</tr>
<tr>
<td>V. ↑</td>
<td>0.0192</td>
</tr>
<tr>
<td>F.I.R</td>
<td>0.0987</td>
</tr>
<tr>
<td>B.W</td>
<td>0.0844</td>
</tr>
<tr>
<td>R. 8→16</td>
<td>0.0047</td>
</tr>
<tr>
<td>R. 32→16</td>
<td>0.0346</td>
</tr>
<tr>
<td>E.A</td>
<td>0.1489</td>
</tr>
<tr>
<td>G.N</td>
<td>0.0647</td>
</tr>
<tr>
<td>M.32</td>
<td>0.1163</td>
</tr>
<tr>
<td>M.192</td>
<td>0.0194</td>
</tr>
</tbody>
</table>

### Table 4  
BER data in binary system

<table>
<thead>
<tr>
<th>Operating means</th>
<th>Binary system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>ABER</td>
</tr>
<tr>
<td>V. ↓</td>
<td>0.0075</td>
</tr>
<tr>
<td>V. ↑</td>
<td>0.0253</td>
</tr>
<tr>
<td>F.I.R</td>
<td>0.1343</td>
</tr>
<tr>
<td>B.W</td>
<td>0.1290</td>
</tr>
<tr>
<td>R. 8→16</td>
<td>0.0081</td>
</tr>
<tr>
<td>R. 32→16</td>
<td>0.0445</td>
</tr>
<tr>
<td>E.A</td>
<td>0.2239</td>
</tr>
<tr>
<td>G.N</td>
<td>0.0920</td>
</tr>
<tr>
<td>M.32</td>
<td>0.1187</td>
</tr>
<tr>
<td>M.192</td>
<td>0.0185</td>
</tr>
</tbody>
</table>

### Figure 5  
FAR – FRR curve in ternary system (Q) (see online version for colours)
The $FAR - FRR$ curve has no cross (see Figures 5 and 6). This shows that the proposed algorithm in different system has both good distinction and robustness, which can accurately identify content preserving operation and malicious operation. Compared Figure 5 with Figure 6 the range of matching threshold in ternary system (Q), which can be set is greater than the range in binary system. Coupled with Tables 3 and 4, it is not hard to find out that ternary digital representation of perceptual hashing digest is better than the binary system. So we adopt the ternary system.

### 4.3.2 Different algorithm robustness

The average $BER$ of ternary systems (O) against content preserving operations are as shown in Table 5.

<table>
<thead>
<tr>
<th>Operating means</th>
<th>Ternary systems (O)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>ABER</td>
</tr>
<tr>
<td>V. ↓</td>
<td>0.0102</td>
</tr>
<tr>
<td>V. ↑</td>
<td>0.0341</td>
</tr>
<tr>
<td>F.I.R</td>
<td>0.1754</td>
</tr>
<tr>
<td>B.W</td>
<td>0.1500</td>
</tr>
<tr>
<td>R. 8→16</td>
<td>0.0083</td>
</tr>
<tr>
<td>R. 32→16</td>
<td>0.0614</td>
</tr>
<tr>
<td>E.A</td>
<td>0.3026</td>
</tr>
<tr>
<td>G.N</td>
<td>0.1150</td>
</tr>
<tr>
<td>M.32</td>
<td>0.2068</td>
</tr>
<tr>
<td>M.192</td>
<td>0.0344</td>
</tr>
</tbody>
</table>
An efficient speech perceptual hashing authentication algorithm

Decrease volume, increases volume and resampling does not change vocal tract model, the energy characteristic and SFF of each frame are less affected and average BER are almost zero (see Table 5), so robustness of the proposed algorithm is best to decrease volume, increases volume and resampling operations. For the low-pass filter with different type, the algorithm still holds a low average BER. Moreover, the average BER of any two speech clips which has the same perceptual content is below 0.35, so the proposed algorithm has a good robustness against content preserving operations, especially the volume adjustment and resampling operations.

Comparison of the average BER of E+LPC algorithm (Huang et al., 2014) with the proposed algorithm in ternary system (Q), shown in Table 6.

Table 6 Average BER of proposed algorithm and E+LPC algorithm

<table>
<thead>
<tr>
<th>Operating means</th>
<th>Ternary system (Q)</th>
<th>Huang et al. (2014)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>ABER</td>
<td>ABER</td>
</tr>
<tr>
<td>V. ↓</td>
<td>0.0057</td>
<td>0.1267</td>
</tr>
<tr>
<td>V. ↑</td>
<td>0.0192</td>
<td>0.2294</td>
</tr>
<tr>
<td>R. 8→16</td>
<td>0.0047</td>
<td>0.0995</td>
</tr>
<tr>
<td>E.A</td>
<td>0.1489</td>
<td>0.3138</td>
</tr>
<tr>
<td>G.N</td>
<td>0.0647</td>
<td>0.3185</td>
</tr>
</tbody>
</table>

FRR and FAR curve of ternary system (O) and E+LPC algorithm is drawn in Figures 7 and 8.

ABER of the proposed algorithm is far lower than the ABER of E+LPC algorithm against the above several kinds of attacks (see Table 6). Compared Figure 7 with Figure 8, FAR – FRR curve of E+LPC algorithm is cross in the picture, but not cross for the proposed algorithm, so the algorithm robustness in the paper against content preserving operations is better than E+LPC algorithm.

Figure 7 FAR – FRR curve in ternary system (O) (see online version for colours)
4.4 Discrimination test and analysis

The BER of different content speech clips basically obey normal distribution. To make a statistical analysis on BER data of the 1280 speech clips in ternary system (O), and eventually the normal distribution of those BER is shown in Figure 9.

Figure 9 BER normal distribution diagram (see online version for colours)
4.4.1 False accept rate

Due to the randomness intrinsic to speech data, the occurrence of ‘–1’, ‘0’ and ‘1’ are theoretically equal to probability. The probability is defined as $p_0$, $p_1$, and $p_2$. Perceptual hashing ternary sequences obey trinomial distribution (Nagaraja, 1992; Jung and Tremayne, 2006), and the probability model is shown in Figure 10.

![Figure 10 ternary sequence probability models](image)

According to the central limit theorem of De Moivre-Laplace, hamming distance approximately obey normal distribution. When adopt error rate as distance measure, the bit error rate approximately obey the normal distribution ($\mu = \frac{N}{3}, \sigma = \sqrt{\frac{1-(1-p_0)^2}{N}}$), where $N$ is the length of perceptual hashing sequences. Probability distribution parameters mean, in theory, should be $\mu = 0.8889$ and the standard deviation $\sigma = 0.0417$. In the actual experiment, the length of perceptual hashing sequence of every speech clip is $N = 160 \times 2$, the mean $\mu_0 = 0.7986$ and standard deviation $\sigma_0 = 0.0445$, and they are closed to the parameter values theoretically calculated. Therefore, hash sequences in the paper possess randomness and collision resistance.

To validate the experiment results further, we calculate $\text{FAR}$ of the proposed algorithm. $\text{FAR}$ is the probability of different multimedia object judged as the same content and accepted by the system, which has the computational method as follows.

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

The smaller decision threshold $\tau$ and the smaller $\text{FAR}$, as in Table 7. When decision threshold $\tau$ is equal to 0.50, as shown in Figure 7, the algorithm can completely distinguish content preserving operations and different speech data, by this time there is approximately one speech clip misjudged among the $10^{11}$ speech clips. The algorithm can maintain a good recognition capability.

4.4.2 Entropy rate

$\text{FAR}$ is highly affected by the size of hash sequence and the same algorithm can possess a different $\text{FAR}$ due to the different size of hash sequences. Hence, it is one-sided and unfair to compare the algorithms performance by only using $\text{FAR}$. However, an entropy rate ($\text{ER}$) 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.

\[
ER = -p \log_2 p - (1 - p) \log_2 (1 - p)
\]

(19)

where \( p = \frac{1}{2} \left( \frac{\sqrt{1 + \sigma^2} - \sqrt{1 - \sigma^2}}{\sqrt{1 + \sigma^2} + \sqrt{1 - \sigma^2}} \right) + 1 \).

\( ER \) of the proposed algorithm is larger than the other two algorithms (Jiao et al., 2009; Huang et al., 2014), as shown in Table 8. It shows that the proposed algorithm has a better recognition performance.

**Table 7** \( FAR \) of the proposed algorithm

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( FAR )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.30</td>
<td>1.9378e–29</td>
</tr>
<tr>
<td>0.35</td>
<td>3.3555e–24</td>
</tr>
<tr>
<td>0.40</td>
<td>1.6636e–19</td>
</tr>
<tr>
<td>0.45</td>
<td>2.3684e–15</td>
</tr>
<tr>
<td>0.50</td>
<td>9.7237e–12</td>
</tr>
</tbody>
</table>

**Table 8** \( ER \) of the different algorithms

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( ER )</td>
<td>0.9527</td>
<td>0.9367</td>
<td>0.8992</td>
</tr>
</tbody>
</table>

4.5 **Tamper detection and location**

Malicious operating modifies local speech data in general, \( BER \) is low and so we cannot judge whether the original speech clip is tampered only by its \( BER \). On the basis of experiences, deviation caused by content preserving operation is always uniform distribution along the time axes, but deviation caused by malicious operating comes about a large distortion in certain local area. The waveform comparison of speech signal against content preserving operation and tamper attack is shown in Figure 11.

4.5.1 **Tamper detection**

To distinguish content preserving operation and malicious tamper attacks to a limited extent, we make a secondary certification for speech signal. The integrated hash feature sequence of speech signal \( S(t) \) is \( ph = [ph_1, ph_2] \) (see Section 3). We assume that the hash template sequence is \( ph' = [ph'_1, ph'_2] \). This paper defines two new metrics to identify malicious tampering attacks and localisation as follow.

**Definition 1:** For the given the block \( i \) and its hash value, define its distance distortion \( DD(i) \) by

\[
DD(i) = |ph'_1(i) - ph_1(i)| + |ph'_2(i) - ph_2(i)|
\]

(20)
An efficient speech perceptual hashing authentication algorithm

Figure 11 The waveform of different speech signal (see online version for colours)

The hash sequences $ph_j$ are split into equal and overlapping groups, denoted as $G_j = \{ph_j(c)|c = 1, 2, \ldots, N - 7, j = 1, 2\}$, the size of group is 8, and group shift is 1. Thus, the total distance distortion of the new group $G(c) = [G_1(c); G_2(c)]$, denoted as $TDD(c)$, is computed as follows.

$$TDD(c) = \sum_{k=1}^{8} DD(c + k - 1)$$  (21)

Assume that $Z_{\text{max}}$ is maximum value, $Z_0$ the number of zero elements and $Z_1$ the number of those elements less than $2Z_{\text{max}}$ in the vector $TDD$.

**Definition 2:** For given three parameters $Z_{\text{max}}, Z_0, Z_1$ in the vector $TDD$, define the tampering metric function $TMF$ by

$$TMF = \frac{Z_{\text{max}} \times Z_0}{Z_1 - Z_0}$$  (22)

When $TMF$ is greater than $T_\tau$, we believe that the local speech signal corresponding to the group $G(c)$ is tempered and so secondary certification should be no pass, where $T_\tau$ is a tampering threshold.

Five thousand speech clips, which are consist of 1,000 speech clips against the tampering attack (10%), B.W, F.I.R, E.A and G.N respectively, randomly are extracted form speech database. When threshold $T_\tau$ is equal to 17, their secondary certification pass rates are shown in Table 9.

Certification pass rate against tampering attack is as low as 0.0735, and false reject rate is lower (see Table 9). So the algorithm holds a good recognition performance against tampering attack.

4.5.2 Tamper location

The local speech signal corresponding to the group $G(c)$ is tempered, and the original signal and its tempered location are shown in Figure 12.

Due to group shift which is set as 1, location accuracy in the algorithm is a block. And the smaller the size of the block is, the higher the location accuracy is.
Table 9 Certification pass rates

<table>
<thead>
<tr>
<th>Operating means</th>
<th>Certification pass rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tampering attack(10%)</td>
<td>0.0735</td>
</tr>
<tr>
<td>B.W</td>
<td>1.0000</td>
</tr>
<tr>
<td>F.I.R</td>
<td>1.0000</td>
</tr>
<tr>
<td>G.N</td>
<td>0.9359</td>
</tr>
<tr>
<td>E.A</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Figure 12 Tamper location (see online version for colours)

4.6 Efficiency analysis

In order to measure the complexity and computational efficiency of proposed algorithm, we randomly extract 100 speech clips, record the average run time, and then make comparisons with the E+LPC algorithm and those algorithms of Gu et al. (2010) and Chen et al. (2013), as in Table 10.

It is a fact that the computing efficiency of the proposed algorithm in the paper is better to E+LPC algorithm, and far better than the algorithms of Gu et al. (2010) and Chen et al. (2013) (see Table 10). So the complexity of the proposed algorithm is lower and more effective. Considering the higher efficiency of cloud computing, the algorithm will be further improved. What is more, the size of perceptual hashing sequences is just 1/6.4 of the size ($N = 64 \times 8 \times 4$) in Jiao’s et al. (2009) algorithm, therefore the proposed algorithm possesses the characteristics of strong summary and relatively small authentication data. In conclusion, the proposed algorithm could satisfy the requirements of real-time speech communication quality, and also apply in the speech communication terminal designed with limited resources in the mobile computing environment.

Table 10 The operation time of proposed algorithm

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating means</td>
<td>Average run time (s)</td>
<td>Average run time (s)</td>
<td>Average run time (s)</td>
<td>Average run time (s)</td>
</tr>
<tr>
<td>File size (s)</td>
<td>4</td>
<td>4</td>
<td>5.6</td>
<td>4</td>
</tr>
<tr>
<td>Basic frequency (GHz)</td>
<td>2.5</td>
<td>3.3</td>
<td>2.8</td>
<td>3.3</td>
</tr>
<tr>
<td>Feature extraction (s)</td>
<td>0.05418</td>
<td>0.02439</td>
<td>0.2700</td>
<td>0.7824</td>
</tr>
<tr>
<td>Match (s)</td>
<td>0.01415</td>
<td>0.002896</td>
<td>0.0006</td>
<td>0.1184</td>
</tr>
<tr>
<td>Total (s)</td>
<td>0.06833</td>
<td>0.02729</td>
<td>0.2706</td>
<td>0.9008</td>
</tr>
</tbody>
</table>
5 Conclusions

The paper proposes an efficient time-frequency domain speech perceptual hashing authentication algorithm based on DWT, which extracts the features of LSE in the frequency-domain and spectral flux in the time-domain to generate the ternary perceptual hashing digest. Experiment results show that ternary form is better to stand for hashing digest than binary form. The proposed algorithm has a good robustness against content preserving operations, good compaction and high efficiency. Hence, the proposed algorithm can satisfy the requirements of real-time speech authentication as well and can be used to ensure content security of the speech data in the cloud.

For future work, we plan to improve the efficiency of this algorithm with focus on accurate positioning and approximate recovery issues.

Acknowledgements

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


