Speech Perceptual Hashing Authentication Algorithm Based on Spectral Subtraction and Energy to Entropy Ratio

Qiu-Yu Zhang[1], Wen-Jin Hu[1], Si-Bin Qiao[1], and Yi-Bo Huang[2]
(Corresponding author: Qiu-Yu Zhang)

School of Computer and Communication, Lanzhou University of Technology[1]
No. 287, Lan-Gong-Ping Road, Lanzhou 730050, China
(Email: zhangqylz@163.com)
College of Physics and Electronic Engineering, Northwest Normal University[2]
No. 967, An-ning East Road, Lanzhou 730070, China
(Received Dec. 9, 2016; revised and accepted Mar. 1 & 12, 2017)

Abstract

In order to meet the requirements of robustness and discrimination of content preserving operations after conversion of speech communication format on the heterogeneous mobile terminal, and noise reduction and efficient authentication, a new efficient speech perceptual hashing authentication algorithm based on spectral subtraction and energy to entropy ratio was proposed. Firstly the proposed algorithm uses spectral subtraction method to denoise the speech signals which processed by applying pre-processing. Secondly, the energy to entropy value matrix of each frame is obtained by applying the method of energy to entropy ratio. Finally, the binary perceptual hash sequence is generated. Experiment results show that the proposed algorithm can denoise the speech effectively, and have good robustness and discrimination to content preserving operations, as well as having high efficiency and good ability to implement tamper detection.

Keywords: Energy to Entropy Ratio; Speech Noise Reduction; Speech Perceptual Hashing; Spectral Subtraction; Tamper Detection

1 Introduction

Currently, Android and iOS are the most popular mobile phone systems, code conversion is needed when there is a communication between two different systems, such as Android system and iOS system. Android's AMR (adaptive multi-rate) format should be converted to WAV format. So when one speech format is converted to another speech format, how to ensure the integrity and authenticity of the speech content? In addition, in the speech instant messaging, the speech is usually affected by coding and decoding, channel noise, delay, packet loss, and the impact of the retrieval speed. In order to achieve efficient speech authentication, how to solve the problem of the interaction between robustness, distinguish and authentication efficiency, so it is very important to study the speech perceptual hashing authentication and speech noise reduction technology [1, 18, 19].

At present, the speech noise reduction methods mainly include: noise cancellation method, spectral subtraction, Wiener filtering method, Kalman filtering method, adaptive filtering method and so on. The spectral subtraction is one of the most commonly used methods. The speech perceptual hashing feature value extraction and processing methods mainly include: logarithmic cepstral coefficients [15], linear frequency spectrum [14], Mel-frequency cepstral coefficients [7, 16], linear prediction coefficient [12], Hilbert transform [22], space-time modulation [13], bark-bands energy [17] and so on. Huang *et al*. [7] proposed a speech perceptual hashing algorithm based on Mel-frequency cepstral coefficients (MFCC) combined with LPCC. The algorithm has good robustness and tamper localization, but it is not good at distinguishing and keeping the content of different speeches, in addition, the signal noise ratio is too high.

Chen *et al*. [4] proposed a speech perceptual hashing algorithm based on LPC combined with non-negative matrix factorization (NMF). The algorithm has good ability of collision resistance, but it is not effective to distinguish the different speeches and content preserving operations. Jiao *et al*. [9] proposed a LSF speech perceptual hashing algorithm based on compressed domain. The algorithm has good robustness and discrimination at low bit rate, but the LSF algorithm is of high computational complexity which affects real-time communication. Zhang *et al*. [20] proposed an efficient speech perception hashing algorithm based on a linear predictive residual coefficient

(LPR) of LP analysis combined with G.729 coding. The algorithm has good robustness, discrimination and high efficiency, but its robustness is poor when the signal noise ratio is low. Jiao *et al*. [8] proposed a speech perception hashing algorithm for the LSP parameterization of speech, which uses the discrete cosine transformation to extract the final characteristic parameters. The algorithm has a good compactness, randomicity and collision resistance, but the extraction efficiency is not high.

Chen *et al*. [2] proposed a speech perception hashing algorithm, which conducts NMF operation on the matrix of the wavelet coefficients based on the wavelet transformation, and gets the hash value finally. Although the algorithm has good robustness in all kinds of content preserving operations, but its processing efficiency is low. Deng *et al*. [5] proposed a hashing algorithm which extracts perceptual feature value based on spectrum energy and divides the audio signal into 33 equal frequency subbands, and the energy of each sub-band is further processed by frequency time filter to get higher robustness to noise and channel distortion, each sub-band energy is represented by 2 bits to obtain the hash value after processing, but the performance is not good at low signal noise ratio (SNR). Huang *et al*. [6] proposed a speech perceptual hashing algorithm based on the improved LPC. The algorithm has good effect on the robustness and the sensitivity of the malicious attacks, and the authentication efficiency is high, but the effect is not very good in distinguishing and keeping the content of different speeches. Li *et al*. [10] proposed a speech perceptual hashing algorithm based on modified discrete cosine transform (MDCT) correlation coefficients combined with NMF. Although the algorithm has good robustness of content preserving operations, but the performance is poor in hashing extraction and matching authentication. Li *et al*. [11] proposed a speech perception hashing algorithm based on MFCC correlation coefficients combined with pseudo random sequences. The algorithm has good robustness, discrimination and security, but its collision resistance is poor and performance at the low signal noise ratio is not good. Chen *et al*. [3] proposed a speech perceptual hashing algorithm based on cochlea and cross recursion, which reduces dimensions by using NMF. The algorithm has good robustness, but the authentication efficiency is low.

In order to solve the problems above, we present an efficient perceptual hashing based on spectral subtraction and energy to entropy ratio for speech authentication after analyze the data that used spectral subtraction and without applying spectral subtraction. The proposed algorithm can solve the problem of the mutual influence between the robustness of content preserving operations, discrimination and authentication efficiency when the AMR format speech converted to WAV format. Firstly, preprocessing of the speech signal is performed after format conversion of the proposed algorithm. And then the spectral subtraction is used to denoise the speech signal. Secondly, the energy entropy ratio parameter matrix of each frame is calculated by using energy to entropy ratio, and the final binary perceptual hashing sequence is generated. Finally, the hashing matching is performed by calculating the hashing number, and the integrity of the speech content is realized perfectly.

The rest of this paper is organized as follows. Section 2 describes the basic theory of spectral subtraction for noise reduction and energy to entropy ratio. A detailed Speech Perceptual Hashing Authentication scheme 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 Problem Statement and Preliminaries

2.1 Spectral Subtraction for Noise Reduction

The spectral subtraction is the most commonly used speech noise reduction method [21]. Let $s(n)$ be the time series of the speech signal, N represent the frame length, and $s_i(m)$ describe the i-th frame for speech signal after windowing and framing. Any frame of speech signal after performed discrete Fourier transform (DFT) is defined as in Equation (1):

$$S_i(k) = \sum_{m=0}^{N-1} s_i(m) exp(j\frac{2\pi mk}{N}) \quad k = 0, 1, \cdots, N-1. \quad (1)$$

Then the amplitude and phase angle of each component of $S(k)$ are obtained. The amplitude can be expressed as $|S_i(k)|$, and phase angel formula can be written as:

$$S^i_{angle} = arctan\left[\frac{Im(S_i(k))}{Re(S_i(k))}\right]. \quad (2)$$

It is assumed that the length of time of no speech section which at the beginning of speech signal (noise clip) denoted as IS, and the corresponding frames are denoted as NIS. Then the average energy of the noise clip can be obtained:

$$D(k) = \frac{1}{NIS}\sum_{i=1}^{NIS} |S_i(k)|^2.$$

The calculation formula for spectral subtraction is shown as in Equation (3):

$$|\hat{S}_i(k)|^2 = \begin{cases} |\hat{S}_i(k)|^2 - a \times D(k) & |\hat{S}_i(k)|^2 \geq a \times D(k) \\ b \times D(k) & |\hat{S}_i(k)|^2 < a \times D(k) \end{cases} \quad (3)$$

where, a and b are two constants, a is defined as reduction factor and b is defined as gain compensation factor.

It can be inferred by Equation (3) that the amplitude is $|\hat{S}_i(k)|$ after performed by the method of spectral subtraction. Combining with Equation (2), the speech sequence $\hat{s}_i(m)$ that processed by the method of spectral subtraction can be obtained by the inverse fast Fourier

transform (IFFT). In this paper, we use the characteristic of the phase insensitive of the speech signal, and the phase angle information of original speech is directly used in the speech signal processed by the method of spectral subtraction.

2.2 Energy to Entropy Ratio

The core of the method of energy to entropy ratio is that the energy of speech section in the speech signal is upward bulge, and the spectral entropy value is less than the spectral entropy value of noise clip. The difference between the speech section and the noise section is more prominent by the method of the energy to entropy ratio. Supposing $s(n)$ is the time series of the speech signal, the i-th frame of speech signal denotes as $s_i(m)$ after processed by windowing and framing, and the length of frame denotes as N. Then energy of each frame is shown as follows.

$$E_i = \sum_{m=1}^{N} s_i^2(m). \tag{4}$$

On the basis of Equation (4), the calculation relationship of energy is improved as follows.

$$LE_i = log_{10}(1 + E_i/c).$$

where, c is a constant. Because of the parameter c, when the parameter c is set to larger value and the amplitude of the energy E_i of each frame fiercely fluctuated and it will be decreased in the LE_i. So the noise and unvoiced section will be distinguished well by a optional parameter c. Parameter c is set to 2 in this paper.

Supposing speech signal in the time domain waveform denoted as $s(n)$, and the i-th frame of the speech signal which processed by applying windowing and framing denotes as $s_i(m)$. And then FFT is performed on $s_i(m)$ and the normalized spectral probability density function of each frequency component is defined as $p_i(k) = Y_i(k)/\sum_{k=0}^{N/2} Y_i(k)$. $Y_i(k)$ denotes the energy spectrum of the k-th line frequency component, $p_i(k)$ represents the probability density of the k-th frequency component of the i-th frame, and N is the length of the FFT. The short-time spectral entropy of each analysis speech frame is shown in Equation (5):

$$H_i = -\sum_{k=0}^{N/2} p_i(k) log p_i(k). \tag{5}$$

Thus the energy to entropy ratio is denoted as $EEF_i = \sqrt{1 + |LE_i/H_i|}$.

3 The Proposed Scheme

The processing flow of the efficient perceptual hashing algorithm based on spectral subtraction and energy to entropy ratio for speech authentication is shown in Figure 1. The speech of Android's AMR format signal is converted to WAV format by the server platform of client, when the Android system communicated with iOS system. Firstly, the pre-processing is needed to the speech signal. Secondly, the method of spectral subtraction is performed in order to denoise the speech. And then the speech is processed by applying windowing and framing. Finally, the method of energy to entropy ratio is used to obtain energy to entropy value.

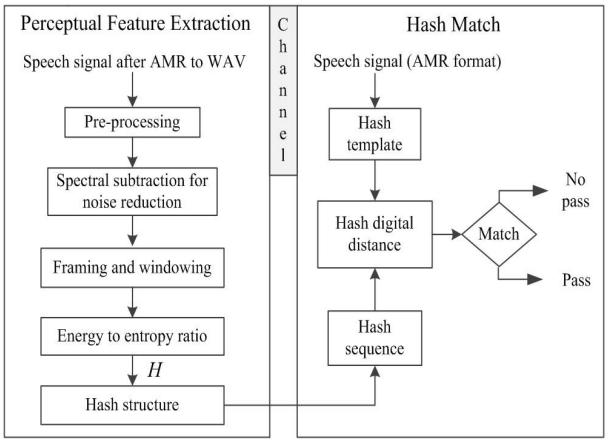

Figure 1: The flow chart of proposed algorithm

The hashing structure and matching of the speech signal are performed, and the processing steps are as follows:

Step 1: Pre-processing. The speech signal $s'(n)$ is obtained by pre-emphasis processing for the input signal $s(n)$. It is useful to improve the high frequency useful part of the signal and extract the subsequent feature. The sampling frequency of the speech signal $s(n)$ is 16 kHz, the number of channels is single channel, and the sampling precision is 16 bit.

Step 2: Spectral subtraction for noise reduction. The speech signal $s'(n)$ is processed by spectral subtraction, and then the speech signal $s''(n)$ is obtained. In the spectral subtraction experiment, the parameters are set as below: the length of frame is 30 ms, frame shift is 25 ms, NIS=8, a=3 and b=0.5. Different selection of the experimental parameters has significant impact on the results (especially noise). The above parameters are the optimal value after testing the experiment.

Step 3: Framing and windowing. The smoothed frame edge is added for speech signal $s''(n)$ by Hamming window. The length of frame is m. It is supposed that the speech $s''(n)$ is divided into n frame, and signal $A_i = \{A_i(k)|i = 1, 2, \cdots, n, k = 1, 2, \cdots, m\}$ is obtained.

Step 4: Energy to entropy ratio. Firstly, FFT is performed on each frame signal A_i, then the frequency domain signal $F_i = \{F_i(k)|i = 1, 2, \cdots, n, k = 1, 2, \cdots, m\}$ is obtained. Secondly, the energy value

of signal F_i is calculated through logarithmic energy algorithm, and then the spectral entropy value of signal F_i is calculated by spectral entropy algorithm. Finally, use the energy to entropy ratio to obtained the parameter matrix $\mathbf{G}(1,n)$, the parameter matrix $\mathbf{G}(1,n)$ is obtained by using the method of energy to entropy ratio, the middle value of matrix $\mathbf{G}(1,n)$ is extracted and it is added in the last new line of the matrix. The matrix $\mathbf{G}(1,n)$ is transformed into matrix $\mathbf{H}(1,n+1)$.

Step 5: Hashing construction. Binary hashing construction is performed by $\mathbf{H}$, the hashing sequence $\mathbf{h}$ is obtained, and the perceptual hashing sequence of speech signal $s(n)$ is $\mathbf{h}(1,n)$=[$\mathbf{h}$].

The binary hashing construction method is as follows.

Using the parameter matrix in the first row of data to subtract the next line of data, if the result is more than 0, the line data turn into 1, otherwise 0.

$$h(i) = \begin{cases} 1 & H(i) > H(i+1) \\ 0 & H(i) \leq H(i+1) \end{cases} \quad i = 1, 2, \cdots, n.$$

Step 6: Hash digital distance and matching. The bit error rate (BER) is defined as normalized hamming distance $D(:,:)$ of the perceptual hashing sequence that is derived from two speech clips $s1$ and $s2$, namely, the ratio of the error bit number to the total number of the perceptual hashing value. The calculation formula is shown as follows:

$$D = \frac{\sum_{i=1}^{N}(|h_{s1} - h_{s2}|)}{N} = \frac{\sum_{i=1}^{N}(h_{s1} \oplus h_{s2})}{N}.$$

where, D is the BER, h_{s1} and h_{s2} correspond to the perceptual hashing values generated by speech clip $s1$ and $s2$, and N is the length of the perceptual hashing values.

The probability of the appearance of "0" and "1" sequence is equal in theory, and the average normalized hamming distance is $0.5N$. We use the hypothesis test of the BER to describe the hashing matching.

P_0**:** Two speech clips $s1$ and $s2$ are the same clip if $D \leq \tau$.

P_1**:** Two speech clips $s1$ and $s2$ are different clip if $D > \tau$.

The hashing values of the same speech clips will take some changes if it be processed by content preserving operations. By setting the size of matching threshold τ, the perceptual hashing sequence mathematical distance of the speech clips $s1$ and $s2$ are compared. If the two mathematical distances $D \leq \tau$, and their perceptual content are treated as the same, the certification is passed, otherwise it doesn't pass the certification.

4 Experimental Results and Analysis

The speech data used in the experiment is the voice in the Texas Instruments and Massachusetts Institute of Technology (TIMIT) and the Text to Speech (TTS) speech library, which is composed of different contents recorded both in Chinese and English by men and women. Every speech clip is converted to WAV format by AMR format with the same length 4 s, which is of the form of 16 bits PCM, mono sampled at 16 kHz, the bit rate is 256 kbit/s, and the length of frame is 30 ms. The speech library in this paper is a total of 1,280 speech clips consisting of 600 English speech clips and 680 Chinese speech clips. The operating experimental hardware platform is Intel(R) Core(TM) i5-2410M CPU, 2.30 GHz, with computer memories of 4G. The operating software environment is MATLAB R2013a of Windows 7 system.

4.1 Robustness Test and Analysis

The content preserving operations are performed for the 1,280 speech clips, as shown in Table 1. The comparison results in various BER and running time between the proposed algorithm and the algorithm without applying spectral subtraction method are shown in Table 2.

As can be seen from Table 2, the proposed algorithm has good robustness and higher operating efficiency for increasing and decreasing of the volume, filtering, resampling and re-encoding than that without applying spectral subtraction algorithm. This is due to the above content preserving operations have little effect on energy and spectral entropy of speech section, at the same time, the algorithm is simple, so it has good robustness and efficiency. However, the noise has great influence on the method of spectral entropy, so the effect is not good on the speech added noise whether it is 20 dB or 30 dB. But the echo is relatively significant influence on the speech section energy, the mean is still high. We can analyze the data from Table 2, when applying spectral subtraction method, we can see that the mean values of all content preserving operation are decrease, but the running efficiency is improved by nearly one times. It has a good improvement on the volume adjustment, echo, resampling and Gaussian noise, this is because of the above operations have great influence on the speech amplitude and noise clip, so the effect is improved obviously by applying spectral subtraction method. Filtering and re-coding has little influence on no speech section which at the beginning of speech signal (noise clip) and the speech amplitude, so the effect of improvement is not remarkable. However, the spectral subtraction method increased the computational complexity and decreased the efficiency. The speech signal to noise ratio is obtained after the speeches processed by spectral subtraction method: the average SNR of 20 dB speech increased by 6.1993 dB and the average SNR of 30 dB speech increased by 6.2538 dB.

Table 1: Content preserving operations

Operating means	**Operation method**	**Abbreviation**
Volume Adjustment 1	Volume down 50%	V.↓
Volume Adjustment 2	Volume up 50%	V.↑
FIR Filter	12 order FIR low-pass filtering, Cutoff frequency of 3.4 kHz	F.I.R
Butterworth Filter	12 order Butterworth low-pass filtering, Cutoff frequency of 3.4 kHz	B.W
Resampling 1	Sampling frequency decreased to 8 kHz, and then increased to 16 kHz	R.8→16
Resampling 2	Sampling frequency increased to 32 kHz, and then dropped to 16 kHz	R.32→16
Echo Addition	Echo attenuation 25%, delay 300 ms	E.A
Narrowband Noise 1	SNR=30 dB narrowband Gaussian noise, center frequency distribution in 0 ∼ 4 kHz	G.N1
Narrowband Noise 2	SNR=20 dB narrowband Gaussian noise, center frequency distribution in 0 ∼ 4 kHz	G.N2
MP3 Compression 1	Re-encoded as MP3, and then decoding recovery, the rate is 32 kbit/s	M.32
MP3 Compression 2	Re-encoded as MP3, and then decoding recovery, the rate is 192 kbit/s	M.192

Table 2: The comparison results in various BER and running time

Algorithm	**Spectral subtraction algorithm**					**Without applying spectral subtraction algorithm**				
Operating means	Mean	Variance	Max	Time (s)	Average time (s)	Mean	Variance	Max	Time (s)	Average time (s)
V.↓	0.0007	0.0023	0.0149	117		0.0119	0.0112	0.0597	65	
V.↑	0.0175	0.0198	0.0896	121		0.0291	0.0270	0.1343	62	
F.I.R	0.0504	0.0196	0.1269	123		0.0529	0.0246	0.1493	64	
B.W	0.0369	0.0172	0.1194	126		0.0359	0.0207	0.1343	63	
R.8→16	0.0081	0.0084	0.0448	121		0.0119	0.0119	0.0672	65	
R.32→16	0.0004	0.0018	0.0149	116	121.2	0.0006	0.0022	0.0149	64	64.5
E.A	0.1042	0.0287	0.2090	122		0.1185	0.0308	0.2239	60	
G.N1	0.0770	0.0314	0.2164	128		0.0990	0.0518	0.3433	64	
G.N2	0.1363	0.0360	0.2836	124		0.1684	0.0568	0.3806	64	
M.32	0.0208	0.0145	0.0821	118		0.0249	0.0201	0.1119	70	
M.192	0.0027	0.0047	0.0299	117		0.0039	0.0061	0.0299	68	

The results of comparison between the proposed algorithm and the algorithm of Ref. [4], the average BER are shown in Table 3.

Table 3: Comparison of average BER

Operating means	**Proposed**	**Ref. [4]**
V.↓	0.0007	0.0726
V.↑	0.0175	0.1123
F.I.R	0.0504	0.3428
B.W	0.0369	0.3445
R.8→16	0.0081	0.1004
R.32→16	0.0004	0.0163
E.A	0.1042	0.1886
G.N1	0.0770	0.4615
M.32	0.0208	0.1682
M.192	0.0027	0.1009

As shown in Table 3, the average BER of the proposed algorithm underwent above attacks is lower than the algorithm of Ref. [4], which shows that our algorithm has good robustness on the content preserving operation, especially on volume controlling, resampling and re-coding. And it is also far superior to the algorithm in Ref. [4] about the 30 dB Gaussian noise and filtering.

This paper totally get 816,003 BER values by conducted pairwise comparison between perceptual hash values from 1,280 different speech clips, and the false accept rate (FAR) and false reject rate (FRR) is obtained via above attacks, and drawing the FAR-FRR curve, the results of comparison between without applying spectral subtraction method and the algorithm in Ref. [4] are shown in Figure 2.

The above FAR-FRR curve is without the content preserving operation of 20 dB Gaussian noise. As shown in Figure 2(a), the FAR-FRR curve obtained by the proposed algorithm is not cross, which means that the proposed algorithm has good distinction and robustness, and it can identify the content of the content preserving operation and the different speech content accurately. As shown in Figure 2(b), when did not apply spectral subtraction method, the FAR-FRR curve of the algorithm was cross, this is due to the poor effect in the Gaussian noise, and the problem of discrimination and robustness cannot be solved very well. As shown in Figure 2(c), the FAR-FRR curve obtained by the algorithm in Ref. [4] is cross, and the problem of discrimination and robustness cannot be solved very well. Combined with Table 2 and Table 3, we can conclude that the robustness on the content preserving operations of the proposed algorithm is better than the algorithm in Ref. [4] and the algorithm without applying spectral subtraction method. Moreover, the noise greatly reduced after applying the spectral subtraction method and the balance with discrimination, robustness

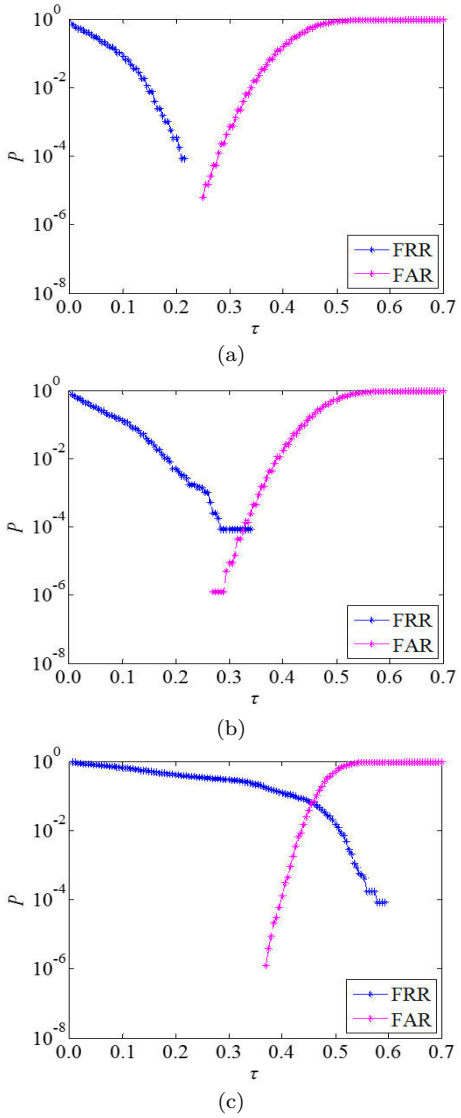

Figure 2: FAR-FRR curve of different algorithms. (a) The proposed algorithm, (b) The without applying spectral subtraction method, (c) The algorithm of Ref. [4].

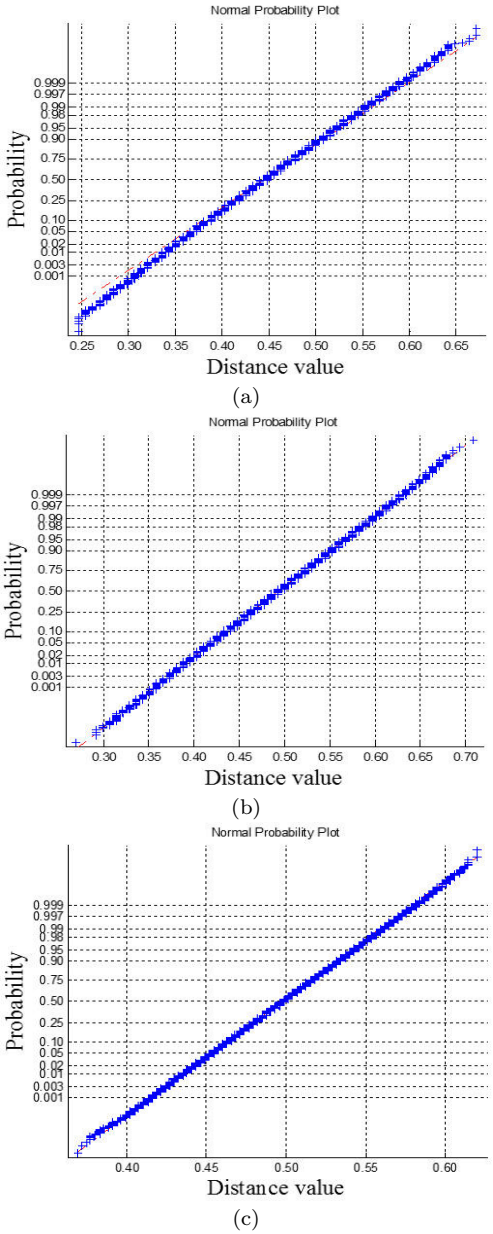

Figure 3: BER normal distribution diagram. (a) The proposed algorithm, (b) The without applying spectral subtraction method, (c) The algorithm of Ref. [4].

and efficiency can be solved very well.

4.2 Discrimination Test and Analysis

The BER of the perceptual hashing values of different speech contents basically obeys the normal distribution. By pairwise comparison of perceptual hash values for 1,280 speech clips, there are 816,003 BER values are obtained. The normal distribution of the BER values of the perceptual hashing sequences is shown in Figure 3.

According to the central limit theorem of De Moivre-Laplace, the hamming distance approximately obeys normal distribution. When adopting BER as the distance measure, the BERs approximately obey a normal distribution ($\mu = p, \sigma = \sqrt{p(1-p)/N}$), where N is the length of perceptual hashing sequence. The closer the BER distribution curve is to the normal distribution, the better the randomicity and collision resistance of the perceptual hashing sequence. In this paper, the length of perceptual hashing sequence is N=134. The theoretical normal distribution parameters mean and standard deviation μ=0.5, σ=0.0307 that are obtained according to the central limit theorem of De Moivre-Laplace. The experimental results demonstrate that the mean and standard deviation are μ_0=0.4452, σ_0=0.0463 in the proposed scheme. However, if without applying the spectral subtraction method in the proposed algorithm, the corresponding mean value is μ_1=0.4933, and the standard deviation is σ_1=0.0446. The FAR is calculated in order to verify the correctness of the

experiment. The expression is shown as follows:

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

where, τ is perceptual authentication threshold, μ represents the BER mean, σ is called BER variance, x is called false acceptance rate.

The comparison results of FAR value are shown in Table 4.

As shown in Table 4, the smaller the matching threshold τ is, the smaller the FAR value is. When the matching threshold τ=0.23, there are approximately 1.67 speech clips misjudged in 1×10^6 speech clips, it demonstrates that the algorithm could meet the requirement of perceptual hashing authentication. By comparison with the algorithm of without applying spectral subtraction methods it can be obtained that the FAR is far lower than the algorithm that applying spectral subtraction method. It is because that when applying spectral subtraction some speech clips are regarded as noise therefore the distinction is decreased. So it is necessary to improve the spectral entropy method to reduce the FAR. When the algorithm can distinguish between the different speeches and the content preserving operations completely, the τ=0.2 and FAR is 1.111×10^{-5} in Ref. [10], the τ=0.3 and the FAR is 9.731×10^{-5} in Ref. [11], so the FAR of the proposed algorithm is lower than the Ref. [10, 11].

4.3 Tamper Detection and Localization

The speech instant messaging of mobile terminals are vulnerable to malicious tampering and attack of criminals. In order to achieve safe and reliable speech content authentication, the speech perception hash algorithm needs to possess the function of tamper detection and location ability for preventing illegal malicious attack and tampering. Generally, illegal malicious operation will cut or tamper part of the speech, errors under the content preserving operations of the speech are often distributed uniformly. However, errors caused by illegal malicious operation usually cause a greater impact in part of the area. So we can determine whether it has been tampered by comparing the hash value. Since the algorithm adopted in this paper is the binary perceptual hash value. So we can judge if there exist tampering by comparing perceptual hash value.

Calculated according to the standard speed 220 words per minute, if there are two or greater speech frames perceptual hashing values are different, we can affirm that it is tampered. This is because that the generally speaking speed is much faster than the standard. And it is also judged as tampering part in the case of the previous and latter frame is different, and the middle frame is same. Because when computing the perceptual hashing values, the previous frame hash value is affected by the latter frame hash value. In order to verify the sensitivity of the algorithm to malicious attacks or tamper, in the experiment, we select a clip of 4 s speech randomly; different speech from the same speaker is used to replace 10% speech clips. Figure 4 is the schematic diagram of perceptual hashing value of tamper localization, where the red elliptic curves contain regions that are tampered. It can be known that the algorithm has a certain ability of tamper detection, and has a good accuracy of tamper detection and localization.

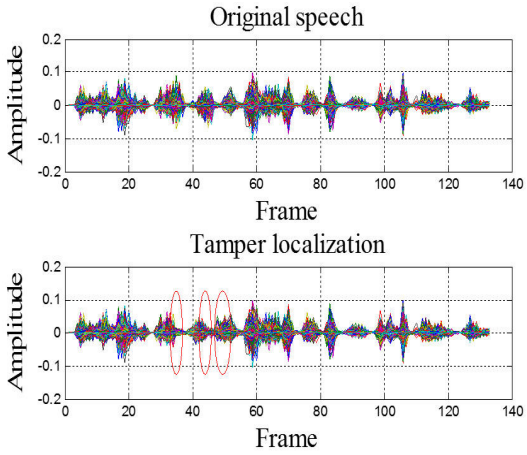

Figure 4: Tamper localization schematic diagrams

4.4 Efficiency Analysis

In order to assess the computational complexity and efficiency of the proposed algorithm, the average run-time is required when performing 1,280 speech clips which are selected randomly from the speech library. The comparison results of the proposed algorithm with the algorithm in Ref. [3, 4, 10] are show in Table 5. In Table 5, the file lengths are 4 s.

Table 5: Comparison of operating efficiency of the algorithm

Algorithms	**Basic frequency** (GHz)	**Average running time** (s)
proposed	2.30	0.0503
Without applying spectral subtraction	2.30	0.0299
Ref. [3]	3.30	0.9008
Ref. [4]	2.27	0.1603
Ref. [10]	2.50	0.1304

As shown in Table 5, the proposed algorithm efficiency is three times more faster than the Ref. [4], two times more faster than the Ref. [10], and nearly 18 times faster than the Ref. [3]. The proposed algorithm has high efficiency and low complexity, and the size of perceptual hashing sequence is 134, which is almost 1/15 of ($N = 64 \times 8 \times 4$) the algorithm in Ref. [8]. And the size of perceptual hashing sequence in the algorithm of Ref. [4, 10] is 360, which

Table 4: The comparison results of FAR value

τ	Proposed	Without applying spectral subtraction	Ref. [10]	Ref. [11]
0.10	4.4688×10^{-14}	5.8061×10^{-19}	2.939×10^{-12}	2.976×10^{-15}
0.15	9.0999×10^{-11}	6.9481×10^{-15}	1.144×10^{-8}	2.631×10^{-12}
0.20	5.9217×10^{-8}	2.4126×10^{-11}	1.111×10^{-5}	9.687×10^{-9}
0.23	1.6763×10^{-6}	1.7784×10^{-9}	-	-
0.25	1.2435×10^{-5}	2.4465×10^{-8}	2.715×10^{-4}	1.484×10^{-6}
0.30	8.5614×10^{-4}	7.3185×10^{-6}	1.682×10^{-3}	9.731×10^{-5}

shown that the summary of the proposed algorithm is powerful. Therefore, the proposed algorithm can meet the requirements of real-time and low complexity of speech communication, which can be applied to the mobile devices with limited bandwidth speech communication terminal and lower hardware configuration in mobile computing environment.

5 Conclusions

An efficient speech perceptual hashing authentication algorithm is proposed based on the spectral subtraction and energy to entropy ratio. The algorithm uses the spectral subtraction method to denoise the speech signal, and then the energy to entropy value that obtained by the method of energy to entropy rate as the perceptual feature which is used to construct the hash sequence and the speech is authenticated. Finally the robustness, discrimination and efficiency of the applied spectral subtraction method and without applying spectral subtraction method are analyzed. Simulations show that the robustness (especially noise) of the proposed algorithm is superior to that without applying spectral subtraction method, but the efficiency is reduced by nearly 1 times and the FAR is increased. In the different speech content preserving operations, the proposed algorithm can effectively resist on the conventional operations, such as resampling, echo, filtering, etc. Especially the effect is good at the volume adjustment and resampling. The proposed algorithm can fully distinguish the different speeches and content preserving operations, at the same time, the false accept rate is low, the efficiency is high, the summary of the proposed algorithm is powerful, and it has a good accuracy of tamper detection and localization.

The main disadvantage of the proposed algorithm is that the efficiency is reduced and the FAR is increased after applying the spectral subtraction method. The next of the research objective is to improve the spectral subtraction in order to decrease the impact of Gaussian noise and reduce the FAR of the algorithm, as well as achieve the approximate recovery and encryption of the speech tampering.

Acknowledgments

This study is supported by the National Natural Science Foundation of China under grant NSFC 61363078, the Natural Science Foundation of Gansu Province of China (No. 1606RJYA274). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] J. Chen, S. Xiang, H. Huang, and W. Liu, "Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet," *Multimedia Tools and Applications*, vol. 75, no. 4, pp. 2303–2325, 2016.

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

[3] N. Chen, H. D. Xiao, J. Zhu, J. J. Lin, Y. Wang, and W. H. Yuan, "Robust audio hashing scheme based on cochleagram and cross recurrence analysis," *Electronics Letters*, vol. 49, no. 1, pp. 7–8, 2013.

[4] N. Chen and W. G. Wan, "Robust speech hash function," *ETRI journal*, vol. 32, no. 2, pp. 345–347, 2010.

[5] J. Deng, W. Wan, R. Swaminathan, X. Yu, and X. Pan, "An audio fingerprinting system based on spectral energy structure," in *Proceedings of the IET International Conference on Smart and Sustainable City (ICSSC'11)*, pp. 1–4, Shanghai, China, July 2014.

[6] 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.

[7] Y. B. Huang, Q. Y. Zhang, Z. T. Yuan, and Z. P. Yang, "The hash algorithm of speech perception based on the integration of adaptive MFCC and LPCC," *Journal of Huazhong University of Science and Technology (Natural Science Edition) (in Chinese)*, vol. 43, no. 2, pp. 124–128, 2015.

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

[9] Y. H. Jiao, Q. Li, and X. M. Niu, "Compressed domain perceptual hashing for MELP coded speech," in *Proceedings of the IEEE International Conference on*

Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP'08), pp. 410–413, Haerbin, China, Aug. 2008.

[10] J. F. Li, H. X. Wang, and Y. Jing, "Audio Perceptual Hashing Based on NMF and MDCT Coefficients," *Chinese Journal of Electronics*, vol. 24, no. 3, pp. 579–588, 2015.

[11] J. F. Li, T. Wu, and H. X. Wang, "Perceptual Hashing Based on Correlation Coefficient of MFCC for Speech Authentication," *Journal of Beijing University of Posts and Telecommunications (in Chinese)*, vol. 38, no. 2, pp. 89–93, 2015.

[12] P. Lotia and D. M. R. Khan, "Significance of Complementary Spectral Features for Speaker Recognition," *International Journal of Research in Computer and Communication Technology*, vol. 2, no. 8, pp. 579–588, 2013.

[13] X. Lu, S. Matsuda, M. Unoki, and S. Nakamura, "Temporal modulation normalization for robust speech feature extraction and recognition," *Multimedia Tools and Applications*, vol. 52, no. 1, pp. 187–199, 2009.

[14] M. Nouri, N. Farhangian, Z. Zeinolabedini, and M. Safarinia, "Conceptual authentication speech hashing base upon hypotrochoid graph," in *Proceedings of the 6th IEEE International Conference on Symposium Telecommunications (IST'12)*, pp. 1136–1141, Glance, Iran, Nov. 2012.

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

[16] V. Panagiotou and N. Mitianoudis, "PCA summarization for audio song identification using Gaussian Mixture models," in *Proceedings of the 18th IEEE International Conference on Digital Signal Processing (DSP'13)*, pp. 1–6, Santorini, Greece, July 2013.

[17] M. Ramona and G. Peeters, "Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection," in *Proceedings of the 2011 IEEE Int. Conference on Acoustics Speech and Signal Processing (ICASSP'11)*, pp. 477–480, Prague, Czech, May 2011.

[18] S. J. Xiang and J. W. Huang, "Audio watermarking to D/A and A/D conversions," *International Journal of Network Secruity*, vol. 3, no. 3, pp. 230–238, 2006.

[19] B. Q. Xu, Q. Xiao, Z. X. Qian, and C. Qin, "Unequal protection mechanism for digital speech transimission based on turbo codes," *International Journal of Network Security*, vol. 17, no. 1, pp. 85–93, 2015.

[20] Q. Y. Zhang, Z. P. Yang, Y. B. Huang, S. Yu, and Z. W. Ren, "Robust speech perceptual hashing algorithm based on linear predication residual of G.729 speech codec," *International Journal of Innovative Computing, Information and Control*, vol. 11, no. 6, pp. 2159–2175, 2015.

[21] Y. Zhang and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Communication*, vol. 55, no. 4, pp. 509–522, 2013.

[22] 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.

Biography

Qiu-yu Zhang (Researcher/PhD 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, multimedia communication technology.

Wen-jin Hu graduated from Shenyang Ligong University, Liaoning, China, in 2010. He received M.Sc. degrees in Communication and information system from Lanzhou University of Technology, Lanzhou, China, in 2014. His research interests include audio signal processing and application, multimedia authentication techniques.

Si-bin Qiao 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.

Yi-bo Huang received Ph. D. candidate degree from Lanzhou University of Technology 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.