Robust Speech Perception Hashing Authentication Algorithm Based on Spectral Subtraction and Multi-feature Tensor

Yi-Bo Huang[1], Qiu-Yu Zhang[2], Wen-Jin Hu[2]

(Corresponding author: Yi-Bo Huang)

College of Physics and Electronic Engineering, Northwest Normal University[1]
No.967, An-Ning East Road, Lanzhou 730070, China
(Email: huang_yibo@foxmail.com)
School of Computer and Communication, Lanzhou University of Technology[2]
No.287, Lan-Gong-Ping Road, Lanzhou 730050, China

(Received Apr. 20, 2017; Revised and Accepted Aug. 26, 2017)

Abstract

In order to make the speech perception hashing authentication algorithm has strong robustness and discrimination to content preserving operations and speech communication under the common background noise, a new robust speech perceptual hashing authentication algorithm based on spectral subtraction and multi-feature tensor was proposed. The proposed algorithm uses spectral subtraction method to denoise the speech which processed by applying pre-processing. Then, the algorithm acquires each speech component wavelet packet decomposition, MFCC and LPCC feature of each speech component are extracted to constitute the speech feature tensor. The feature tensor is decomposed tensor decomposition to reduce the complexity. Finally, speech authentication is done by generating the hashing values which use mid-value. Experimental results show that the proposed algorithm can denoise the speech effectively, and have good robustness and discrimination to content preserving operations, as well as able to resist the attack of the background noise, which is commonly heard during the communication.

Keywords: Background Noise; Multi-feature Tensor; Robust; Spectral Subtraction; Speech Perceptual Hashing

1 Introduction

Speech signal is easily to be disturbed in the transmission channel, in the speech instant messaging; the speech is usually affected by coding and decoding [17], 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 noise reduction technology and speech perceptual hashing authentication [2].

Therefore it is necessary to consider whether the speech feature can be extracted completely and accurately, and it is required that the calculation of the perceptual speech hashing robust should be the strongest, the coupling should be minimum, calculating should be easy. The extraction of the speech perception feature value is the key of speech perceptual authentication. In order to reduce the influence of noise on speech feature extraction, speech denoising technique is used in preprocessing. At present, the speech noise reduction methods mainly include: spectral subtraction, Wiener filtering method, Kaman filtering method, adaptive filtering method and so on. The current extraction of the speech perception feature is based on the human ear psycho acoustic model, the speech perceptual hashing feature value extraction and proceeding methods mainly include: the spectrum coefficient [11], linear predictive coding (LPC) [9], Mel-scale Frequency Cepstral Coefficients (MFCC) [4, 6] line spectrum frequency (LSF) [10], Energy to Entropy Ratio [19], frequency cepstral coefficients [12], Hilbert transform [21] and bark-bands energy [14]. The literature [4] proposed a Speech perception hashing algorithm that based on the Mel-scale Frequency Cepstral Coefficients and Nonnegative matrix factorization (NMF), the paper proposed a singular value decomposition then obtains the speech information, and then undergo the NMF. It reduces the mistakes and gives out a satisfactory outcome of the hashing function. The experiment shows that the Robustness is improved but the Distinction is poor, due to the principal component analysis method is used in the algorithm, the time complexity of the algorithm is large and it cannot meet the requirements of realtime speech authentication. Chen [3] proposed a speech perceptual hashing algorithm based on LPC combined with non-negative matrix factorization.

The algorithm has good ability of collision resistance, but it is not effective to distinguish the different speeches and content preserving operations. Zhang [22] proposed an efficient speech perception hashing algorithm based on a linear predictive residual coefficient of LP analysis combined with G.729 coding. The algorithm has good robustness, discrimination and high efficiency, but robustness is poor when the signal noise ratio is low. Li [8] 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 collision resistance is poor and performance at the low signal noise ratio is not good.

In order to solve the problem of robustness and discrimination in speech perception hashing authentication, we present a robust perceptual hashing based on spectral subtraction and multi-feature tensor 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. Firstly, preprocessing of the speech signal used spectral subtraction to denoise the speech signal noise. Secondly, introduces the method of MFCC coefficients and LPC cepstrum coefficients in the process of perception speech hashing, feature modeling based on multi-feature, Construction of feature tensor by multi-feature, finally, the authentication function is realized by using tensor decomposition and hashing structure.

The rest of this paper is organized as follows. Section 2 describes the basic theory of spectral subtraction for noise reduction and the basic algorithm of multi-feature. A detailed speech perceptual hashing authentication scheme is described in Section 3. Section 4 gives the experimental results as compared with other related method. Finally, we conclude our paper in Section 5.

2 Problem Statement and Preliminaries

2.1 Spectral Subtraction for Noise Reduction

The spectral subtraction speech enhancement is utilized broadly because it is simple and easy for the realtime processing [23]. The main idea of spectral subtraction is the independence of noise and speech signal, it will be Noisy speech power spectrum minus the noise power spectrum, and then get the pure speech spectrum.

$$y(t) = x(t) + n(t).$$

Let $x(t)$ be a speech signal, $n(t)$ is a noise signal, and $y(t)$ is a noisy speech signal.

$$Y(\omega) = S(\omega) + N(\omega). \tag{1}$$

Equation (1) is a frequency expression.

$$\begin{aligned} |Y(\omega)|^2 &= |S(\omega)|^2 + |N(\omega)|^2 + 2Re[S(\omega)N*(\omega)] \\ E(|Y(\omega)|^2) &= E(|S(\omega)|^2) + E(|N(\omega)|^2) \\ &\quad +2E(Re[S(\omega)N*(\omega)]). \end{aligned} \tag{2}$$

In Equation (2), $S(\omega)$ and $N(\omega)$ are completely independent. $N(\omega)$ submit to zero mean value normal distribution. Equation (3) can be written as:

$$|Y(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2 \tag{3}$$

$N(\omega)|^2$ can be estimated by Silent section. The estimated value of the original speech is defined as in Equation (4):

$$|S(\omega)| = [|Y(\omega)|^2 - |N(\omega)|^2]^{\frac{1}{2}} \tag{4}$$

2.2 LPCC Feature Coefficient Extraction

LPCC is a commonly used speech feature. This feature can be used to build the speech model, and the speech model is considered as the all pole model, which can be realized simply and easily in algorithm. But for the voiceless and nasal recognition effect is poor. We can directly derive the cepstrum from the linear prediction coefficient. The recurrence relation between LPC coefficient and LPCC coefficient is below:

$$\begin{aligned} c_0 &= a_1 \\ c_n &= a_n + \sum_{k=1} n - 1\frac{k}{n} c_k a_{n-k}, 1 \leq n \leq N \\ c_n &= \sum_{k=1} n - 1\frac{k}{n} c_k a_{n-k}, n > N. \end{aligned}$$

Here, c_0 is the DC component, a_n is the LPC coefficient, and c_n is the LPCC coefficient.

2.3 MFCC Feature Coefficient Extraction

When speech feature are extracted, the MFCC [13] is mostly used as the feature vector [5]. Mel scale describes the nonlinear feature of human ear's frequency perception. Its relation with the practical frequency of speech. The equation below:

$$Mel(f) = 2595 \log(1 + \frac{f}{700}), 0 \leq f \leq F_n.$$

In the equation above, f is the practical speech frequency; F_n is the Nyquist frequency of speech signal.

2.4 Wavelet Packet Transform

Discrete wavelet transform (DWT) has the ability to accurately characterize local details of speech signals [1, 20] Wavelet packet transform (WPT) as the further expansion of wavelet analysis theory. Wavelet packet decomposition can reflect the feature and nature of the signal. It is very suitable for the analysis and processing of

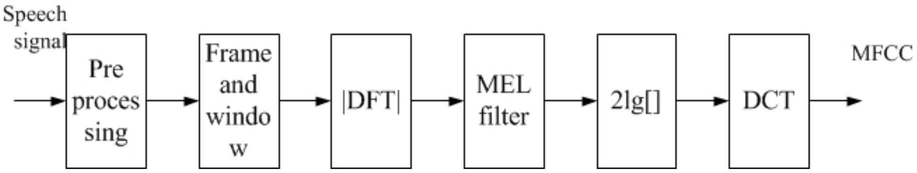


Figure 1: MFCC feature extraction process

speech signal types of non-stationary signal [15, 16, 18]. K level wavelet packet decomposition principle is shown in Figure 2. The subspace U_i^m is $U_m(t)$ and $U_{2m}(t)$'s closure spaces, speech signal through the recursive equation wavelet packet decomposition:

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

Here, $h(n)$ is a high pass filter group, and $g(n)$ is a low pass filter group, $g(n) = (-1)^n h(1-n)$, and that the two coefficients with orthogonal relation.

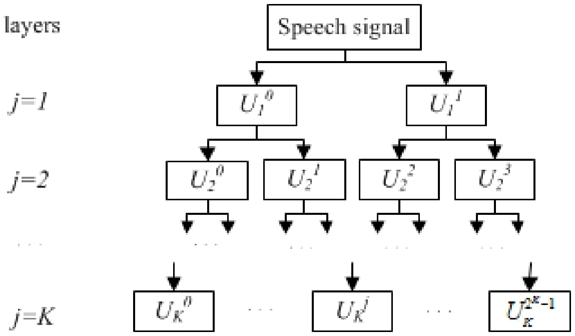


Figure 2: The decomposition graph of K-level wavelet packet

3 The Proposed Scheme

3.1 Establishment of Speech Tensor Model

Tensor can be considered as a product of vector space, and it is a higher order generalization of vector and matrix. The order of the tensor can be expressed as $X \in R^{N_1\times N_2\times L\times N_M}$. Tensor decomposition method is widely used in image processing, pattern recognition, data compression and so on [7]. It can better show the relationship among speech frame structure, decomposition scale and feature coefficient. Figure 3 shows the schematic diagram for the construction of speech tensor, which can directly describe the structure of speech tensor.

Describe the speech feature from three perspectives, which are respectively the speech frame, wavelet packet decomposition scale as well as MFCC and LPCC feature coefficients. The speech frame mainly describes the precedence relationship of speech and describes the relationship of speech feature from the time scale. The wavelet package decomposition scale conducts wavelet package decomposition for each frame of speech signal so as to different scales of approximate components and detailed components of each frame of speech signal. MFCC and LPCC feature coefficients conduct the feature extraction for the components decomposed by each wavelet package to obtain the component feature. Tensor construction can be carried out for a section of speech from the above three perspectives. The tensor constructed is the third order speech tensor of [speech frame × wavelet package decomposition scale × MFCC and LPCC feature coefficients].

Figure 4 shows the construction method for speech feature tensor adopted in this paper. The construction diagram consists of the speech signal preprocessing, speech feature extraction and speech feature tensor construction.

The Tucker decomposition model is the product of N-order tensor $X \in R_{I_1\times I_2\times L\times I_n}$ through Tucker decomposition to obtain the product of a lower dimensional core tensor G and N projection matrix $U^{(n)}$. Tucker decomposition model is below:

$$X \approx G\times_2 U^{(1)}\times_2 U^{(2)}\times L\times_N U^{(N)}.$$

$G \in R_{J_1\times J_2\times L\times J_N}$ is core tensor. The main information of the original tensor is retained in the core tensor. $U^{(n)} \in R_{I_n\times J_n}$ is projection matrix, $J_n \leq I_n$ and $U^{(n)}$ are orthogonal. Tucker decomposition can be used optimal decomposition to solve the optimization problem.

$$\begin{gathered} \min |X - G\times_1 U^{(1)}\times_2 U^{(2)}\times L\times_N U^{(N)}|^2, \\ G\geq 0, U^{(n)}\geq 0. \end{gathered}$$

If $J_n = rank_{(n)}X$, Tucker decomposition is meaningless. If $J_n < rank_{(n)}X$, Tucker decomposition is meaningful.

3.2 Quantization

Reconstruct the core tensor G to form the two-dimensional feature matrix. Calculate the sum of each column of matrixes:

$$R_h(j) = \sum_{i=1}^{r} H_{ij}^{(n)}, 1\leq j\leq k.$$

In the equation above, $H_{ij}^{(n)}$ signifies the feature coefficient in row j and line I, k is the number of rows of feature

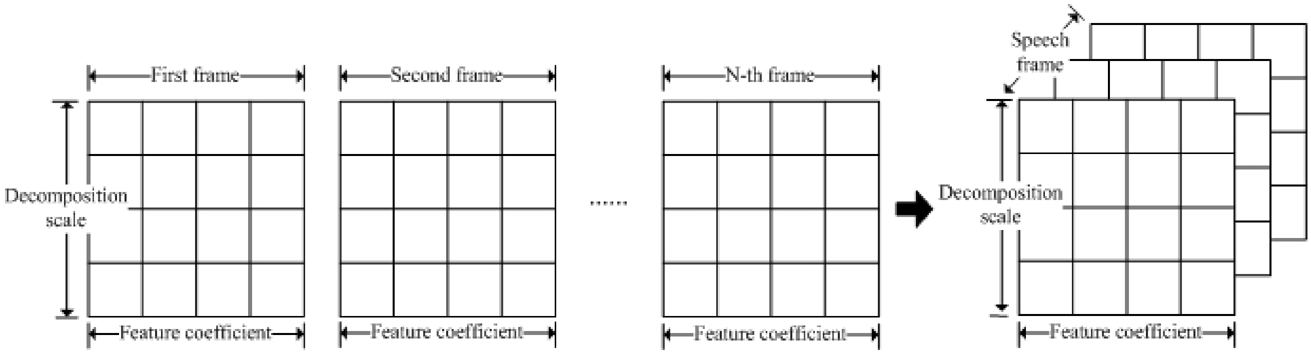


Figure 3: The structure graph of speech tensor

matrix. Quantize the coefficient formed and row matrix to form the hashing value $h(j)$ of speech segment;

$$h(j) = \left\{ \begin{array}{ll} 1 & R_h(j) > \hat{R_h}, 1 \leq j \leq k \\ 0 & \text{other} \end{array} \right\}$$

In the equation above, $\hat{R_h}$ is the mid-value.

3.3 Speech Perception Hashing Authentication Scheme

Figure 4 describes the construction of speech tensor. After the tensor decomposition, since the core tensor G is less than the original tensor X, the core tensor G can be considered as the compression form of original tensor X. In this algorithm, the core tensor G is used to describe the speech feature. The flow chart of speech perception hashing authentication based on Spectral Subtraction and Multi-feature tensor is shown in Figure 5.

The detailed steps of the algorithm are shown below:

Step 1: Preprocessing: conduct pre-emphasis on the speech in the speech library to be tested, enhance the useful frequency spectrum of high frequency, reduce the edge effect and eliminate noise.

Step 2: Spectral subtraction for noise reduction: the speech signal is processed by spectral subtraction, In the spectral subtraction experiment, the length of frame is 30ms, frame shift is 25ms, $NIS = 8$, $a = 3$, $b = 0.5$.

Step 3: Framing and windowing: in order to eliminate the inter frame loss during framing, conduct framing and add the Hamming window for speech $x(t)$; during framing, the frame length is L; when the frame moves at $L/2$, $s(n)$ can be obtained; later, add Hamming window for $s(n)$ to obtain $s_w(n)$, n is the frame number.

Step 4: Wavelet packet transform: carry out wavelet package decomposition for the speech frame. In this paper, the 3-order wavelet packet decomposition is carried out, 8 speech segments are obtained, and then calculate the MFCC coefficient and LPCC coefficient of each segment.

Step 5: Construction and decomposition of speech tensor: conduct the tensor construction of feature coefficient to obtain the speech feature tensor X, carry out Tucker decomposition for the feature tensor X to obtain the low-dimensional core tensor G and the project matrix $U^{(n)}$.

Step 6: Quantization: construct the core tensor G and thus obtain the sequence $R_h(j)$; quantize $R_h(j)$ to obtain the perception hashing sequence $h(j)$;

Step 7: Calculation and matching of perception hashing distance: suppose that there are two speech segments α and β, define the hashing mathematic distance is $D_h(:,:)$, which is shown below:

$$D_h(H_\alpha, H_\beta) = \sum |h_\alpha(j) - h_\beta(j)|, j = 1, 2, L, n.$$

Match according to the hypothesis testing of hashing mathematic distance $D_h(:,:)$ and hashing sequence $h(:)$ as follows:

K1: If the perception contents of the two speech segments α and β are the same:

$$D_h(H_\alpha, H_\beta) \leq \tau.$$

K2: If the perception contents of the two speech segments α and β are different:

$$D_h(H_\alpha, H_\beta) > \tau.$$

In the equations above, τ is matching threshold. The matching threshold can be used to determine whether the perception contents of speech signals are the same so as to realize the perception hashing authentication of speech signals.

4 Experimental Results and Analysis

The operating software environment is MATALB 2010b. The operating experimental hardware platform is Intel(R)

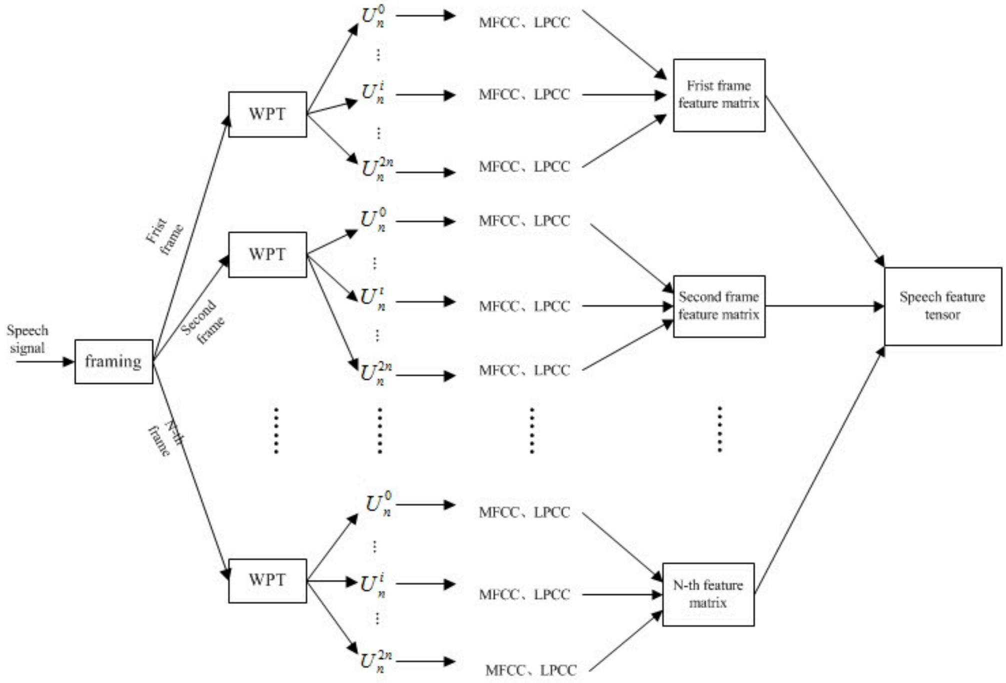


Figure 4: The structure graph of speech feature tensor

Core(TM) i5-4590 CPU 3.30GHz, with computer memory of 4GB. The speech data used in the experiment is the speech in the Texas instruments and Massachusetts institute of technology (TIMIT) speech library which is composed of different contents recorded English by men and women, and Noisex-92 noise library as noise library. The speech clip length is 4s. The speech library in this paper is a total of 1,280 speech clips. The content preserving operations are performed for the 600 speech clips, as shown in Table 1.

4.1 Discrimination Test and Analysis

Discrimination is mainly used to evaluate the reliability of the algorithm for distinguishing different speech contents read by different or same persons. Since the bit error rates (BER) of different speech segments are random variables, this experiment analyzes the discrimination of algorithm with the probability distribution curve. The BER of the perceptual hashing values of different speech contents basically obeys the normal distribution. By pairwise comparison of perceptual hash values for 600 speech clips, there are 179700 BER values are obtained. Compare every two of speech in the speech library and the diagram of BER normal distribution obtained is shown in Figure 6.

When the error rate is used as the distance measure, it should approximately abbey the normal distribution. It can be seen from Figure 6 that the probability curve of standard normal distribution overlaps the probability distribution of BER value of this algorithm, so the hashing distance obtained through this algorithm approximately obeys the normal distribution; namely, speech with different perceptions will generate different hashing values.

In the ideal condition, every speech segments with different contents will have its different perception hashing valve and every pair of hashing value matching should have a high error rate. Actually, there are always a few of BER data which are low and probably lower than the threshold value, then it will be wrongly judged as same content. According to Table 1, it can know that the false acceptation rate (FAR) increases with the enlargement of BER threshold value. Compared with the other two algorithms, the algorithm proposed in this paper has a strong collision resistance. When the threshold value $\tau = 0.25$, the collision probability is that 6 segments among 10^{10} speech segments may collide. When $\tau = 0.27$, the collision probability is that 1 segments among 10^8 speech segments may collide. When $\tau = 0.30$, the collision probability is that 6 segments among 10^7 speech segments may collide. It can be seen from Figure 6 that 2 segments among 10^5 speech segments will collide when the threshold value $\tau = 0.35$. As indicated in Table 1, compared with other two algorithms, the algorithm is very stable in the collision resistance. Therefore, this algorithm can correctly identify the authenticated speech segments.

The mean of matching between different speech is 0.4996 and the square is 0.0411, as shown in Figure 7, the μ and δ measured in the experiment are close to the theoretical results.

Table 1: Content preserving operation

Operating means	Operation method	Abbreviation
Volume Adjustment 1	Volume down 50%	V. ↓
Volume Adjustment 2	Volume up 50%	V. ↑
Resampling 1	Sampling frequency decreased to 8kHZ, and then increased to 16kHZ	R.8 → 16
Resampling 2	Sampling frequency decreased to 32kHZ, and then increased to 16kHZ	R.32 → 16
Narrowband noise 1	SNR=20db narrowband Gaussian noise, center frequency distribution in 0 4kHZ	G.N20
Narrowband noise 2	SNR=30db narrowband Gaussian noise, center frequency distribution in 0 4kHZ	G.N30
MP3 Compression 1	Re-encoded as MP3, and then decoding recovery, the rate is 32k	M.32
MP3 Compression 2	Re-encoded as MP3, and then decoding recovery, the rate is 48k	M.48
MP3 Compression 3	Re-encoded as MP3, and then decoding recovery, the rate is 128k	M.128
MP3 Compression 4	Re-encoded as MP3, and then decoding recovery, the rate is 192k	M.192

Table 2: The comparison results of FAR value

	[4]	Without applying spectral subtraction algorithm	The proposed algorithm
$tau = 0.20$	5.1875e-013	5.5635e-014	1.6299e-013
$tau = 0.25$	2.0985e-009	3.0135e-010	6.4908e-010
$tau = 0.27$	6.4075e-008	6.1722e-009	1.1930e-008
$tau = 0.30$	4.5267e-006	3.6602e-007	6.1098e-007
$tau = 0.32$	4.4124e-005	4.1395e-006	6.3326e-006
$tau = 0.35$	9.7400e-003	1.0136e-004	1.3820e-004

Table 3: The matching rate of speech authentication after the being kept operating content %

	[4]	Without applying spectral subtraction algorithm	The proposed algorithm
V. ↓	98.3	98.86	100
V. ↑	97.4	100	100
R.8 → 16	97.2	95.3	95.3
R.32 → 16	96.4	98.1	100
G.N20	78.1	86.7	91.6
G.N30	93.4	94.7	95.3
M.32	84.2	87.4	91.4
M.48	91.2	96.7	96.7
M.128	94.6	96.8	98.7
M.192	100	100	100

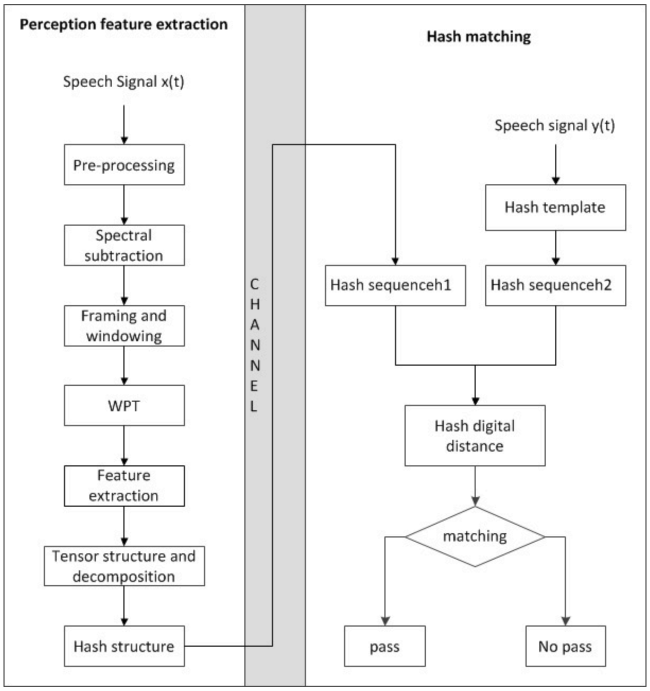


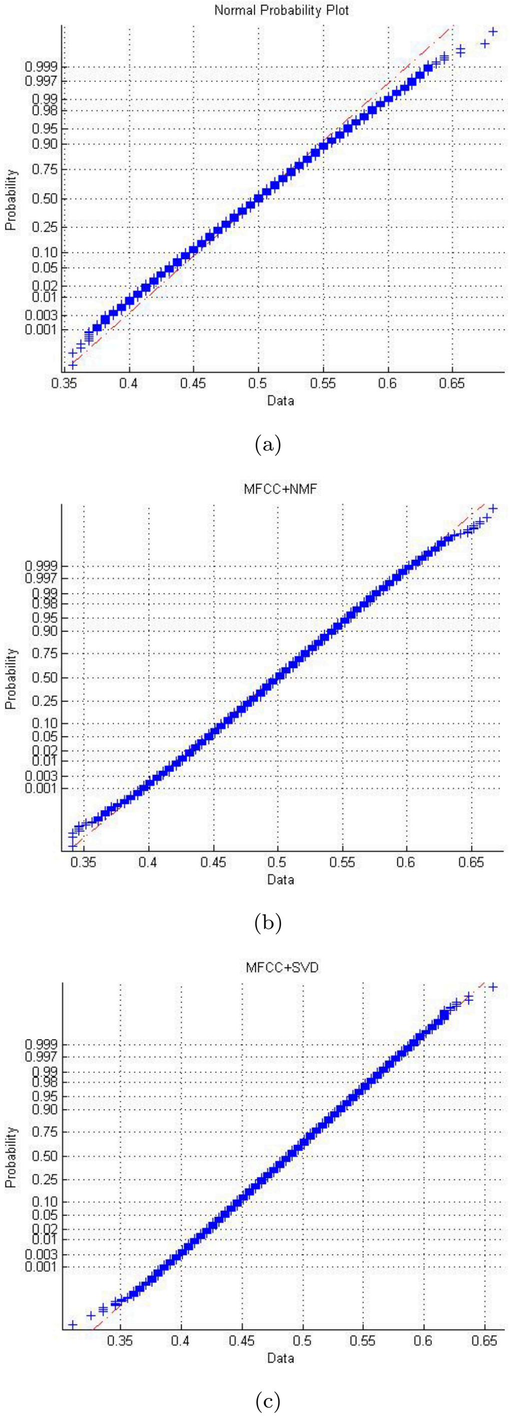
Figure 5: The flowchart of speech perception hashing authentication

4.2 Robustness Test and Analysis

The speech perception hashing robustness is mainly used to evaluate the reliability of the same speech after different preserving operations. The content preserving operations are performed for the 1280 speech clips, as shown in Table 1. The comparison results in various BER and the algorithm without applying spectral subtraction method are shown in Table 2.

As can be seen form Table 2, the proposed algorithm has good robustness for increasing and decreasing of the volume, filtering, resampling and re-encoding than that without applying spectral subtraction algorithm and [4], this is due to about content preserving operations have little effect on speech feature, so the algorithm has good robustness. However, the noise has great influence on the LPCC and MFCC coefficient, so the effect is not good on the speech added noise whether it is 20db or 30db. We can analyze the data from Table 3, 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, 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. According to the data in Table 2 and Table 3, as for the robustness of preserving content operation, the proposed algorithm is generally stronger than other two algorithms.

Based on the BER rate of content preserving operation, the false acceptance rate (FAR) and false reject rate (FRR) are obtained. Draw the FAR-FRR curve, which is shown in Figure 8.

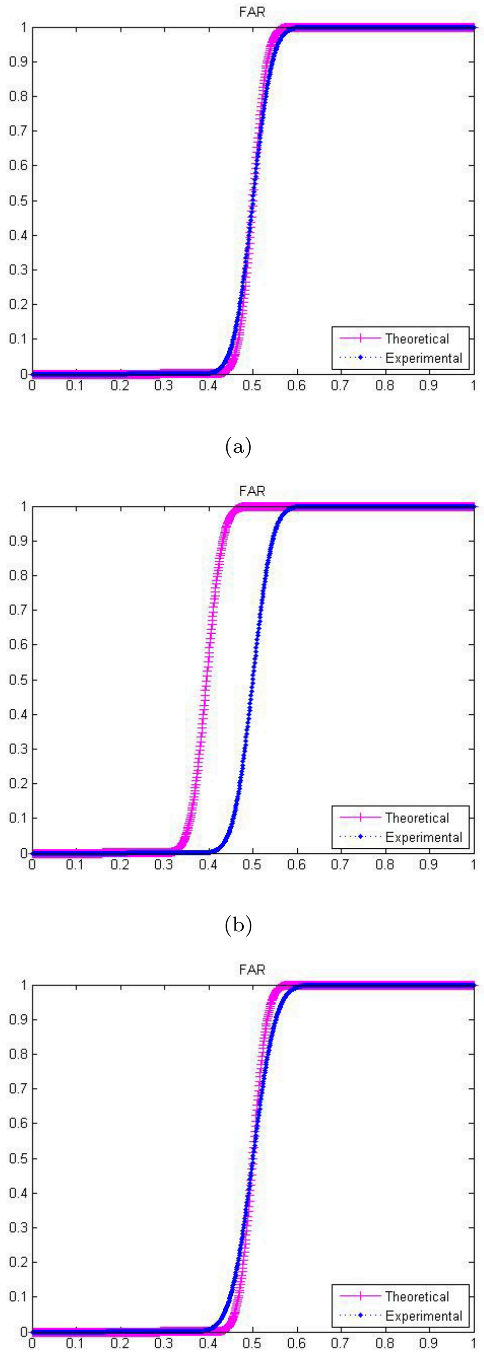


(a)

(b)

(c)

Figure 6: The BER normal distribution diagram. (a) The Proposed algorithm, (b) The algorithm of [4], (c) Without applying spectral subtraction algorithm.



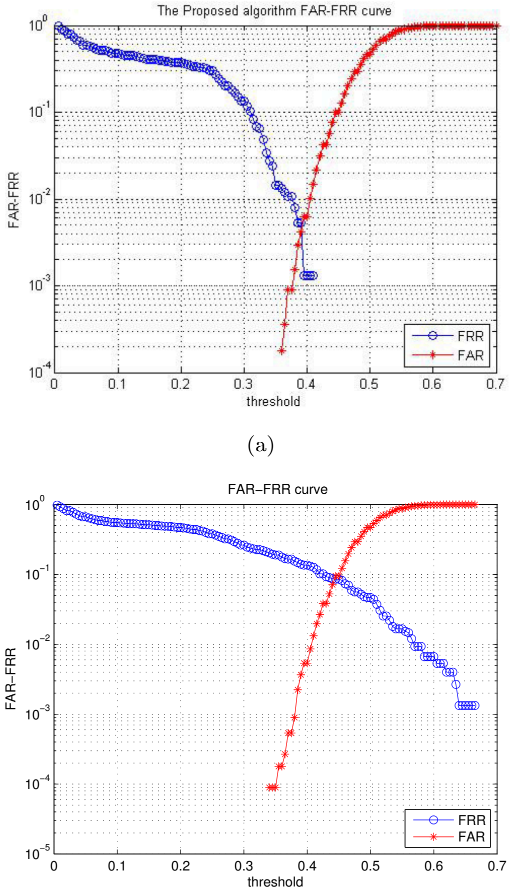
(a)

(b)

(c)

Figure 7: FAR curves of the algorithm. (a) The Proposed algorithm, (b) The algorithm of [4], (c) Without applying spectral subtraction algorithm.

This paper totally get 179,700 BER values by conducted pairwise comparison between perceptual hashing values form 600 different speech clips, and the false accept rate 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 [4] are shown in Figure 8. As indicated in the result in Figure 8, the FAR-FRR curve obtained by the proposed algorithm has a litter cross. The proposed algorithm's FRR curve does intersects with the FAR curve in 0.39, the FRR curve shows a significant convergence and a very wide judging domain; the judging threshold is between 0.35 and 0.40, showing a significant judging domain. Compared with the algorithms which without applying spectral subtraction algorithm, this algorithm has good robustness, and can correctly authenticate the same and different speech segments and, meanwhile, the algorithm authenticate the speech segments which go through the content holding operation and malicious attack. Therefore, compared with other two algorithms, this one has good discrimination and robustness.



(a)

(b)

Figure 8: The FAR-FRR curves of different perceptual hashing algorithm. (a) The Proposed algorithm, (b) Without applying spectral subtraction algorithm.

Table 4: BER mean and running time comparison

Operating	Mean	Variance	Max	Mean time	Mean	Variance	Max	Mean time
Algorithm	The proposed algorithm				Without applying spectral subtraction algorithm			
V. ↓	0.1019	0.3187	0.3187	72min40s	0.1253	0.3723	0.3723	65min42s
V. ↑	0.1121	0.3312	0.3312		0.0811	0.3163	0.3163	
R.8 → 16	0.1200	0.3750	0.3750		0.1012	0.5250	0.5250	
R.32 → 16	0.0981	0.3187	0.3187		0.0847	0.4125	0.4125	
G.N20	0.1980	0.3062	0.3363		0.2637	0.5563	0.5250	
G.N30	0.1371	0.3150	0.3212		0.1629	0.5062	0.5000	
M.32	0.1586	0.4700	0.4575		0.1750	0.4750	0.4750	
M.48	0.1301	0.3237	0.3238		0.0953	0.4188	0.4188	
M.128	0.1101	0.4313	0.4313		0.0829	0.3563	0.3563	
M.192	0.0846	0.3000	0.3000		0.1012	0.3500	0.3500	

Table 5: The algorithm efficiency (time/s)

	The proposed algorithm	Without applying spectral subtraction algorithm
Hashing structure	7.24	6.65
Hashing values	0.008	0.008
Total	7.248	6.658

4.3 Robustness for Common Background Noise

People usually talk in a noisy environment, so add Noisex-92 noise library which is common background noise, including white noise, pink noise, factory floor noise 1, factory floor noise 2, speech babble and Volvo noise. The signal-to-noise ratios of noises added are respectively 0db, 10db, 20db, 30db, 40db and 50db.

As shown in Figure 9, this algorithm has extremely strong robustness for Gaussian white noise attack and Babble noise attack; the robustness of this algorithm is obviously stronger than other two algorithm. Its robustness for other several noises is also in the middle level. As for the passing rate for pink noise attack, the passing rate of the three algorithms is all 100%.

As shown in Figure 9, the proposed algorithm has strong robustness for common background noise. In particular, its robustness for Gaussian white noise and Babble noise is significantly higher than the robustness of [4] and [3]. Its robustness performances for Volvo noise, Factory 1 noise, Factory 2 noise and Pink noise are in the middle level. The passing rate of authentication matching for various signal-to-noise ratios is very high. Compared with the NMF algorithm, the tensor decomposition algorithm has stronger robustness for pink noise attack and the feature value decomposed through tensor algorithm has higher stability. Therefore, the algorithm proposed in this paper has strong combination property of robustness for common noises, so it can meet the practical need of speech matching in our daily life.

4.4 Efficiency Analysis

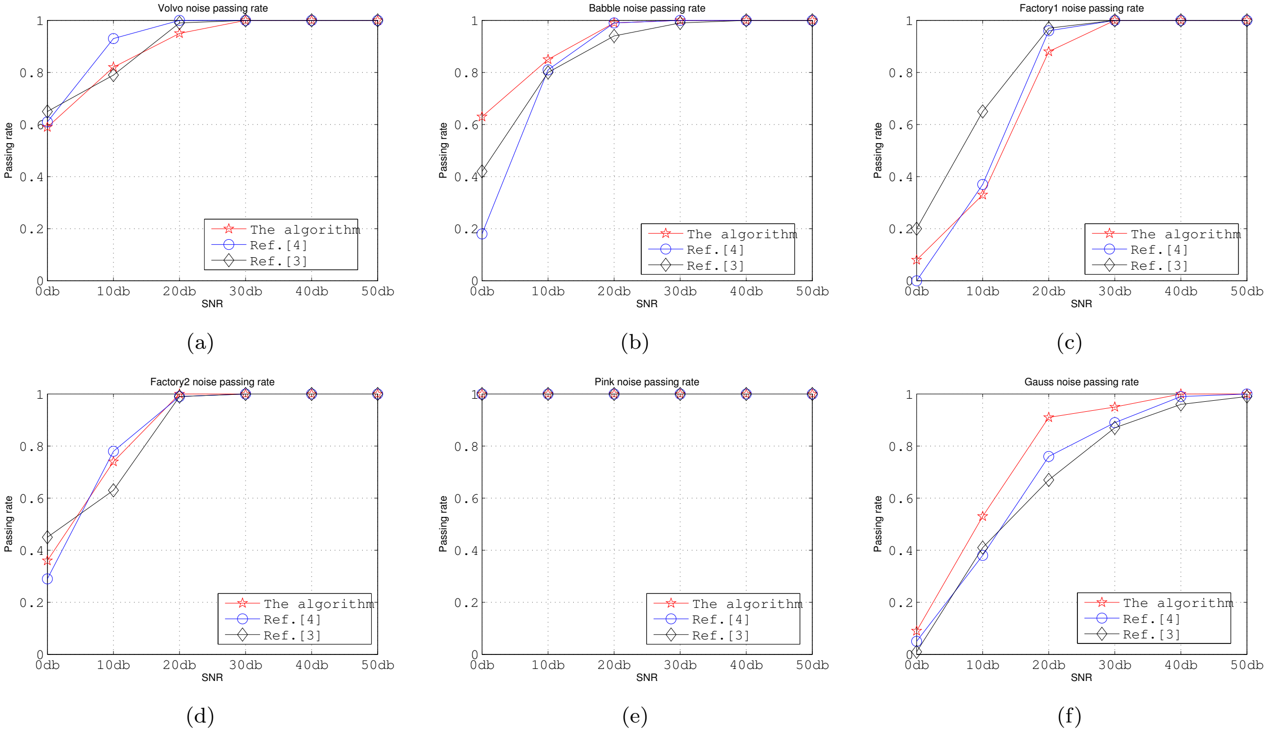
In order to measure the computational efficiency of the proposed algorithm, the researcher randomly extracted 200 segments of speech from the speech library to count the average operating time.

As shown in Table 5, compared with the without applying spectral subtraction algorithm, the proposed algorithm is required to conduct the spectral subtraction noise reduction, wavelet package decomposition prior to feature extraction and make tensor reconstruction for the feature extracted. Thus, the expenditure of operating time is long. On the premise of enhancing the robustness, compared with other algorithms, there is increase in the overall expenditure of operating time of this algorithm and its operating efficiency is largely affected, so this algorithm can only be applied to the occasions with low realtime requirement speech communication.

5 Conclusion

This paper proposed a robust perception hashing speech authentication algorithm based on Spectral subtraction and Multi-feature tensor. Through an experimental discussion and analysis, the proposed algorithm robust is better than the earlier methods as shown in the discussion, the conclusions below can be made.

As shown in the results of speech perception hashing discrimination and collision resistance experiments, the highest speech misidentification probability within the range of threshold. It means that the algorithm has good collision resistance performance and can meet the need of



(a)

(b)

(c)

(d)

(e)

(f)

Figure 9: The Speech authentication passing rate of being the common background noise attack. (a)Volvo noise, (b) Babble noise, (c) Factory1 noise, (d) Factory2 noise, (e) Pink noise, (f) Gaussian white noise.

practical application.

As shown in the robust experiment, compared with the other two algorithms, the robustness of the algorithm proposed is improved to a certain extent. After the content preserving operation is conducted, the algorithm can correctly match speech; there is an obvious judging domain in the FAR-FRR curve and the scope of the judging domain is 035. Compared with other algorithm, the FRR-FAR curve has small crossover. The proposed algorithm is especially in allusion to the common background noises during daily communications.

As shown in the experiment of speech against noise attack, this algorithm has strong robustness for common background noise, so it can meet the need of daily communications on varied dialogue backgrounds. Compared with other two kinds of algorithms, the algorithm for common background noise robustness is more stable. This algorithm can control the tensor size as required and model building is flexible. Besides, it can realize the speech content authentication and speaker authentication, the algorithm has a high practical value. Simulations show that the robustness 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. The main disadvantage of the proposed algorithm is that the efficiency is deduced. The next of the research objective is to improve the efficiency, decrease the impact of echo and Tamper detection and localization of speech.

Acknowledgment

This study was supported by the Natural Science Foundation of Gansu Province of China under grant 1606RJYA274, the National Nature Science Foundation of China under grant NSFC 61363078. The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] S. T. Ali, J. P. Antoine, J. P. Gazeau, *Coherent States, Wavelets, and Their Generalizations*, New York: Springer Publishing Company, 2013.

[2] 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 Application*, vol. 75, no. 4, pp. 2303–2325, 2016.

[3] N. Chen and W. G. Wang, “Robust speech hash function”, *ETRI Journal*, vol. 32, no. 2, pp.345–347, 2010.

[4] N. Chen, H. D. Xiao and W. G. Wan, “Audio hash function based on non-negative matrix factorization of Mel-frequency Cepstral coeffi-cients”, *IET Information Security*, vol. 5, no. 1, pp.7–8, 2013.

[5] M. A. Hossan, S. Memon, M. A Gregory, “A novel approach for MFCC feature extraction”, in *IEEE Xplore Conference: Signal Processing and Communication Systems*, pp.1–5, Gold Coast, Australia, 2010.

[6] 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*, vol. 43, no. 2, pp.124–128, 2015.

[7] J. Li, L. X. Jin and G. N. Li, "Hyper-spectral remote sensing image compression based on nonnegative tensor factorizations in discrete wavelet domain", *Journal of Electronics & Information Technology*, vol. 35, no. 2, pp. 489–493, 2013.

[8] J. F. Li, T. Wu and H. X. Wang, "Perceptual hashing based on NMF and MDCT coefficient of MFCC for speech authentication", *Journal of Beijing University of Posts and Telecommunications (in Chinese)*, vol. 38, no. 2, pp.89–93, 2015.

[9] P. Lotia and M. R. Khan, "Significance of complementary spectral feature for speaker recognition", *International Journal of Research in Computer and Communication Technology*, vol. 8, no. 2, pp.579–588, 2013.

[10] M. Nouri, N. Farhangian and Z. Zeinolabedini, "Conceptual authentication speech hashing base upon hypotrochoid graph", in *Proceedings of The Sixth International Symposium on Telecommunications*, pp. 1136–1141, Tehran, Iran, Nov. 2012.

[11] H. Ozer, B. Sankur and N. Memon, "Perceptual audio hashing functions", *EURASIP Journal on Applied Signal Processing*, vol. 12, pp. 1780–1793, 2005.

[12] V. Panagiotou and N. Mitianoudis, "PCA summarization for audio song identification using Gaussian mixture models", in *Proceedings of The 18th International Conference on Digital Signal Processing*, pp.1–6, Santorini, Greece, July 2013.

[13] J. W. Picone, "Signal modeling techniques in speech recognition", *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

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

[15] P. Sharma, K. Khan and K. Ahmad, "Image denoising using local contrast and adaptive mean in wavelet transform domain", *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 12, no. 6, pp.1450038.1–1450038.15, 2014.

[16] R. Sharma and V. P. Pyara, "A robust denoising algorithm for sounds of musical instruments using wavelet packet transform", *Circuits and Systems*, vol. 7, no. 4, pp. 459, 2013.

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

[18] Y. Yang and S. Nagarajaiah, "Blind identification of damage in time-varying systems using independent component analysis with wavelet transform", *Mechanical Systems and Signal Processing*, vol. 47, no. 1, pp. 3–20, 2014.

[19] Q. Y. Zhang, W. J. Hu, S. B. Qiao and Y. B. Huang, "Speech perceptual hashing authentica-tion algorithm based on spectral subtraction and energy to entropy ratio", *International Journal of Network Security*, vol. 19, no. 5, pp.752–760, 2017.

[20] Q. Y. Zhang, P. F. Xing, Y. B. Huang, R. H. Dong and Z. P. Yang, "An efficient speech perceptual hashing authentication algorithm based on wavelet packet decomposition", *Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, no. 2, pp. 311–322, 2015.

[21] Q. Y. Zhang, Z. P. Yang, Y. B. Huang, R. H. Dong and P. F. Xing, "Efficient robust speech authentication algorithm for perceptual hashing based on Hilbert-Huang transform", *Journal of Information and Computational Science*, vol. 11, no. 18, pp. 6537–6547, 2014.

[22] 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.

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

Biography

Yi-bo Huang received Ph.D candidate degree form 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 in-formation processing, information security, speech recognition.

Qiu-yu Zhang (Researcher/Ph.D supervisor), graduated from Gansu university of technology in 1986,and then worked at school of computer and communication in Lanzhou university of technology. He is vice dean of Gansu manufacturing information engineering research center, CCF senior member, a member of IEEE and ACM. His research interests include network and information security, information hiding and steganalysis, multimedia communication technology.