Spectrogram-based Efficient Perceptual Hashing Scheme for Speech Identification

Qiu-Yu Zhang, Tao Zhang, Si-Bin Qiao, and Dong-Fang Wu
(*Corresponding author: Qiu-Yu Zhang*)

School of Computer and Communication, Lanzhou University of Technology
No.287, Lan-Gong-Ping Road, Lanzhou 730050, China
(Email: zhangqylz@163.com)
(*Received Sept. 19, 2017; revised and accepted Mar. 27, 2018*)

Abstract

In order to meet the requirements of discrimination, robustness and high-efficiency identification of existing speech identification algorithms in mobile speech communication, an efficient perceptual hashing scheme based on spectrogram for speech identification was proposed in this paper. Firstly, a fraction of spectrogram is cut, which represents low frequency information of input speech signal and is less susceptible to common content-preserving manipulations such as MP3 compression, noise addition and volume adjustment etc. Secondly, the local binary pattern (LBP) algorithm is applied to produce a LBP feature image. Finally, the perceptual hashing sequence is obtained by employing an image perceptual hashing algorithm to the LBP feature image. Experimental results show that the proposed approach has a good discrimination, robustness and identification efficiency. It can satisfy the real-time identifying requirements in mobile speech communication.

Keywords: Entropy Rate; LBP; Perceptual Hashing; Spectrogram; Speech Identification

1 Introduction

With the development of multimedia technology and network communication technology, the transmission and storage of speech information become more and more convenient. However, some speech information contains much private information, such as court testimony and military order. Therefore, validating their authenticity becomes a critical issue to multimedia identification techniques [5].

Traditional cryptography hashing algorithms are very sensitive to the changes of speech content because of introducing some distortions while processing speech signal, such as resample and compression, which have an adverse effect on speech content identification. Speech perceptual hashing identification technologies can protect speech information by verifying its authenticity, which can guarantee speech information services more safe and reliable [2, 4]. So, this technology is recently receiving big attention in the area of research.

A lot of spectrogram-based audio fingerprinting algorithms have been proposed in recent year. Rafii *et al.* [12] proposed an audio fingerprinting system to handle different kinds of live version audio, and the fingerprinting is extracted from a binary image, which is obtained from a log-frequency spectrogram by using an adaptive threshold method. Though the system shows a good identification precision to different genres of live music, its robustness is not illustrated in detail. To satisfy robustness of audio fingerprinting system, Zhang *et al.* [15] proposed a feature extracting method based on spectrogram through utilizing scale invariant feature transform (SIFT) local descriptor and the locality sensitive hashing (LSH). Due to the stability of SIFT, the proposed algorithm achieves a high discrimination and robustness, but its time complexity is high. Being similar to [15], SIFT is employed to extract 128 features of spectrogram in [16]. Experimental results show that the system has a good identification rates when the audio lengths are stretched from 65% to 150%. However, due to use Euclidean distance to match the features of 128-dimension descriptors, it is still time-consuming.

Besides there are a few audio fingerprinting algorithms based on the feature of spectrogram, some audio perceptual hashing algorithms with respect to other features have been proposed. Chen *et al.* [3] introduced an audio perceptual hashing algorithm based on Zernike moment. Experiment results show that the algorithm achieves a good discrimination and perceptual robustness. However, generating hashing process in his paper takes too much time. Huang *et al.* [6] proposed a speech perceptual hashing algorithm based on linear prediction analysis, which has a high running efficiency but not a good robustness. Li *et al.* [9] introduced a hashing generating approach by utilizing the correlation of Mel-frequency cepstrum coefficients (MFCC). Instead of traditional hamming distance, it takes advantage of similarity metric function to implement hashing matching and shows a good robustness to

re-sampling and MP3 compression. However, the algorithm is computationally expensive. In [8], the combination of modified discrete coefficients transform (MDCT) and non-negative matrix factorization (NMF) is considered as a hashing yielding scheme. It exhibits a good robustness but a poor discrimination. To develop an efficient speech identification system, Zhang *et al.* [14] proposed a speech perceptual hashing algorithm in terms of discrete wavelet packet decomposition (WPD). Although the system can discriminate different speech files, it lacks robustness in the noisy environment. By exploiting linear prediction coefficients (LPC) of speech signal to obtain local features, Chen *et al.* [1] proposed a robust hash function, which gains a better discrimination but has poor effect to resist some speech content-preserving distortions, such as filtering and noise addition.

Aiming at the problems mentioned above, obtaining a compromise between discrimination and robustness of algorithm, and meeting the requirement to enhance identification efficiency, we proposed an efficient speech perceptual hashing identification algorithm based on spectrogram. Firstly, a spectrogram produced by a 4 s original speech is obtained like Figure 1(a). Figure 1(b) is a spectrogram of the original speech signal contaminated by 30 dB white Gaussian noise. Through comparing Figure 1(a) with Figure 1(b), it is obvious that noise has little influence on low frequency portion (namely the bottom half of the spectrogram), then, which is cut to get Figure 1(c). In addition, Figure 1(c) is converted to acquire a feature image (Figure 1(d)) by using LBP algorithm. As can be seen from Figure 1(d), most of textural features are extracted. Lastly, an image perceptual hashing algorithm proposed in [7] is applied to get hashing sequences of LBP feature image, and which are matched to finish speech identification. The experimental results demonstrate that the proposed algorithm can satisfy the real-time need of mobile speech communication.

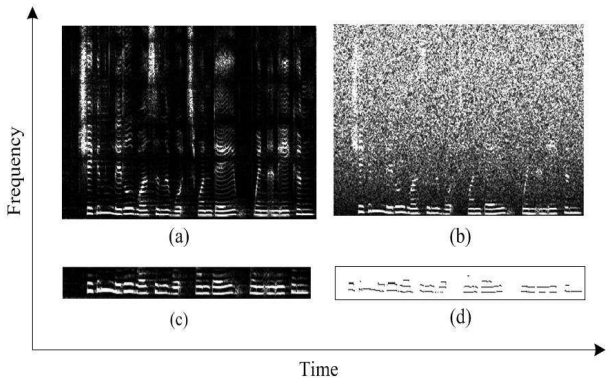

Figure 1: Spectrogram analysis: (a) Spectrogram of an original 4 s speech clip; (b) Spectrogram of adding 30 dB noise to original speech; (c) An image of being cut out from (a); (d) LBP feature image of (c)

The remaining part of this paper is organized as follows. Section 2 does several preliminaries, which are mainly introducing three theories, including LBP descriptor, 2D-DCT and SVD, which will be exploited in this paper. The detailed proposed algorithm is described in Section 3. Subsequently, Section 4 gives the experimental results and performance analysis as compared with other related methods. Finally, we conclude our paper in Section 5.

2 Problem Statement and Preliminaries

2.1 LBP Descriptor

LBP [13] is an effective image texture description method. Owing to having some characteristics, such as simple calculation, rotation and gray invariance, LBP is applicable to real-time system. The method is described briefly as follows: a central pixel point i_c is defined in a local 3×3 neighborhood of a monochrome texture image. Next, i_c is in comparison with one of the joint 8 pixel values i_n (i=1, 2, ..., 8), if i_c is less than in, b_n is set to 1, otherwise to 0 (see Equation(1)). Equation (2) is utilized to get final LBP code values.

$$b_n = \begin{cases} 1 & i_n - i_c > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$LBP(x_c, y_c) = \sum_{n=0}^{7} 2^n b_n. \quad (2)$$

In this paper, LBP descriptor is used to extract texture information of spectrogram.

2.2 Two-Dimensional Discrete Cosine Transform (2D-DCT)

Discrete Cosine Transform (DCT) [11] is used to approximate to an image via different amplitude and frequency. Two-dimensional discrete cosine transform (2D-DCT) can be obtained by computing twice one-dimensional DCT in the two directions of row and column. 2D-DCT is frequently applied in image processing since it is characterized by lossless compression and energy concentration. Generally, after 2D-DCT, the main energy of an image is concentrated in the part of low frequency, which is located in top left corner of a 2D-DCT coefficient matrix and representing the stable features of an image. The definition of 2D-DCT is as follows:

$$D(u,v) = \frac{2}{\sqrt{MN}} c(u)c(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cos\frac{(2x+1)u\pi}{2M} \times \cos\frac{(2y+1)v\pi}{2N} \quad (3)$$

The 2D-DCT inverse transform is given by:

$$f(x,y) = \frac{2}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} c(u)c(v)D(u,v) \cos\frac{(2x+1)u\pi}{2M} \times \cos\frac{(2y+1)v\pi}{2N} \quad (4)$$

where $D(u,v)$ is a 2D-DCT transform coefficient matrix, the size of input digital image $f(x,y)$ is $M \times N$, $0 \le x \le M-1$, $0 \le y \le N-1$, $0 \le u \le M-1$, $0 \le v \le N-1$, $c(u)$ and $c(v)$ are transform parameters and their values are defined as follows:

$$c(u), c(v) = \begin{cases} 1/\sqrt{2} & u, v = 0 \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

In this paper, 2D-DCT is utilized to extract the low frequency energy of LBP feature image.

2.3 Singular Value Decomposition

Singular value decomposition (SVD) [10] is a method of algebraic feature extraction and numerical analysis. Thanks to the stability of singular value, it is widely applied in the field of image compression and digital watermark. Supposing A is a $M \times N$ gray image. The SVD transform of A is given by:

$$A = USV^T = \sum_{i=1}^{r} \lambda_i u_i v_i^T \quad (6)$$

where U and V are $M \times M$ and $N \times N$ orthogonal matrices respectively, S is a $M \times N$ matrix. λ is singular value of A and satisfies Equation (7). r is the number of non-zero singular value. $u_i$and v_i are left singular vector and right singular vector corresponded by λ_i.

$$\lambda_1 \ge \lambda_2 \ge \cdots \lambda_r = \cdots = \lambda_M \quad (7)$$

In this paper, SVD is performed to the 2D-DCT low frequency coefficient matrix for obtaining the left singular vector and right singular vector corresponded by the biggest singular value.

3 The Proposed Scheme

The two principal components of speech identification system are hashing generation and hashing matching, and the procedures of proposed perceptual hashing algorithm are shown in Figure 2. The whole steps of the algorithm are depicted as follows: an input speech signal yields a hashing sequence, which is matched with other hashing sequences that are stored in a reference hashing database. And matching results are analyzed to identify input speech content. The specific hashing generation process can be seen from Figure 3. Firstly, one-dimensional speech signal is converted to a two-dimensional spectrogram, due to the high frequency part of speech signal is vulnerable to some distortions, so an image block representing low frequency portion of input speech signal is cut out. Next, by using LBP method, the image block is extracted texture features to gain a LBP feature image. Subsequently DCT coefficient matrix is obtained by utilizing 2D-DCT to the LBP feature image. It is divided into many smaller matrices with the same size, and some of them representing low frequency part of LBP feature image are recombined into a new matrix. Finally, after doing SVD to it, a hash sequence is derived.

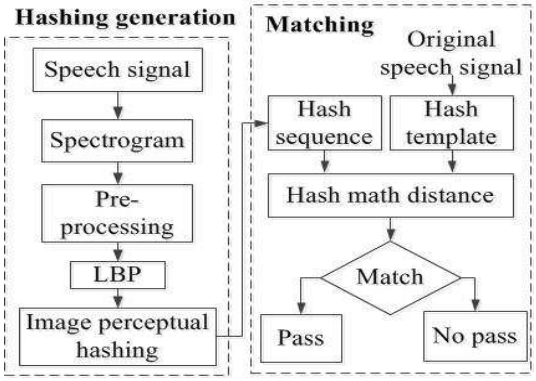

Figure 2: Block diagram of speech perceptual hashing identification algorithm

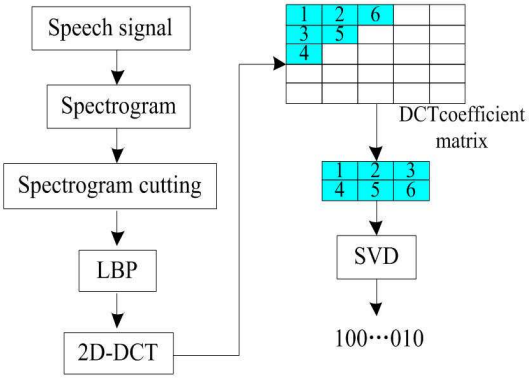

Figure 3: Block diagram of hashing generation of input speech signal

3.1 The Process of Hashing Generation

Assuming the original speech signal is s, the steps of hashing generation are depicted as follows:

Step 1: Short Time Fourier Transform (STFT) is used to obtained the spectrogram of s, which in matrix form is $S_i = \{S_i(k) \mid i = 1, 2, \ldots, M, k = 1, 2, \ldots, L\}$, where M and L represent the number of rows and columns of the matrix, respectively.

Step 2: Cutting the spectrogram which represents the low frequency part of s, expressed as $Cs_i = S_i$, where, $i = 1, 2, \ldots, N$, and N is the number of rows after S_i is cut.

Step 3: The LBP is performed on Cs to obtain a $M_1 \times N_1$ LBP feature image $L(m, n)$. where M_1 and N_1 are the number of rows and columns of $L(m, n)$ and $m = 1, 2, \ldots, M_1, n = 1, 2, \ldots, N_1$.

Step 4: The 2D-DCT is performed on $L(m, n)$ to obtain a 2D-DCT coefficient matrix $D(m, n)$. In order to extract its top-left-corner low frequency energy conveniently, it is divided into 25 same-sized $p \times q$ matrix blocks $\Psi_{i,j}(p, q)$, and some of them representing low frequency energy of $L(m, n)$ are recombined to derive a new matrix C, which is shown as follows:

$$C = \begin{bmatrix} \Psi_{1,1} & \Psi_{1,2} & \Psi_{2,1} \\ \Psi_{3,1} & \Psi_{2,2} & \Psi_{1,3} \end{bmatrix} \quad (8)$$

where, i and j are block position indices of 2D-DCT block matrix, and each block has p rows and q columns. In this paper, $i = 1, 2, \ldots, 5, j = 1, 2, \ldots, 5, p = 1, 2, \ldots, 6, q = 1, 2, \ldots, 74$.

Step 5: The SVD is performed on C to get a left singular value vector u_1 and a right singular value vector v_1 corresponded by the biggest singular value. Next they are transposed respectively and combine into a feature vector F, which is shown as follows:

$$F(k) = [u_1^T \ v_1^T] \quad 1 \le k \le l \quad (9)$$

where T and k are referred to matrix transpose and feature index respectively, total number of feature is l.

Step 6: F is quantified to obtain a perceptual hashing sequence by using Equation (10).

$$h(k) = \begin{cases} 1 & F(k) \ge 0 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where k is hashing index, l is the length of hashing codes.

3.2 Hashing Match

After yielding hashing sequences, the normalized hamming distance, shown in Equation (11), is utilized to match them. The bit error rate (BER) is the ratio between the length of mismatches and the total length of hashing vector, and it is equal to normalized hamming distance in numerical value.

$$BER(h_1, h_2) = D(h_1, h_2) = \frac{1}{l} \sum_{k=1}^{l} \mid h_1(k) - h_2(k) \mid \quad (11)$$

where h_1 and h_2 are two hashing sequences randomly extracted from two speech clips s_1 and s_2, k is hashing index, l is the length of hashing sequence.

In order to estimate the performance of whole perceptual hashing system, a statistical hypothesis testing method is defined as follows:

Given two randomly selecting speech clips s_1 and s_2

H_0**:** if s_1 and s_2 are same two perceptual contents,

$$BER \le \tau$$

H_1**:** if s_1 and s_2 are two different perceptual contents,

$$BER > \tau$$

where τ is perceptual threshold. By setting a reasonable τ and computing BER of two clips s_1 and s_2, if BER$\le \tau$, the two clips can be treated as same two perceptual contents, identification is passed, otherwise not passed.

4 Experimental Results and Analysis

In this section, the performance of the proposed algorithm will be evaluated. The experimental speech data comes from the Texas Instruments and Massachusetts Institute of Technology (TIMIT) speech database and the Text to Speech (TTS) speech database. There are different 1280 speech clips in experimental database recorded by 640 men and 640 women. The format of each speech clip is wav with the length 4 s, which is of the form of 16 bits PCM, mono and sampled at 16 kHz. Experimental hardware environment is Intel(R) Core(TM) i5-3230, 4-core processor, 8 G and 2.6 GHz, software environment is the MATLAB 2013a under Win7 operating system. Next, the proposed method compares with three algorithms in [8, 14, 1], for convenience, which are abbreviated as MDCT-NMF [8], WPD-QT [14] and LPC-NMF [1] respectively. Some parameters involved in this experiment are shown in Table 1.

Table 1: The parameters used in this experiment

Hashing algorithm	Parameters
Proposed	M=257, L=372, N=32, M_1=255, N_1=30, l=234
MDCT-NMF	M=360, L=177, N=100, r=1
WPD-QT	M=64000, N=256, n=16
LPC-NMF	M=360, N=12, r=1

As shown in Table 1, in MDCT-NMF algorithm, speech signal is divided into M frames with L samples, N is the number of lower MDCT coefficients of each frame, and r is the dimension-reduction number of NMF. In WPD-QT algorithm, the length of speech signal is M, wavelet packet coefficients matrix is split into N identical $n \times n$ square blocks. In LPC-NMF algorithm, framing number is M, and N is the order of LPC, and r is the dimension-reduction number of NMF.

4.1 Discrimination Analysis

In this phase, 1, 280 different speech clips are used to calculate BERs in pairs, therefore a total of 818,560 BERs can be obtained and follow the distribution shown in Figure 4. Supposing the generation of binary sequence is random (independent and identical distributed), consequently, these BERs follow binomial distribution (l, μ), where l represents the length of hashing sequence and μ is the probability of that a 0 or 1 is extracted. According to central limit theorem, if l is large, the BERs obey the normal distribution with a mean μ of 0.5 and the standard deviation $\sigma = \sqrt{\mu(1-\mu)/l} = \sqrt{1/4l}$. In our experiment, l=234. After substituting l into above equation, it is found that theoretical standard deviation σ value is 0.0327 (experimental mean and standard deviation value are 0.4904 and 0.0332, respectively), which shows that experimental results are pretty close to theoretical values.

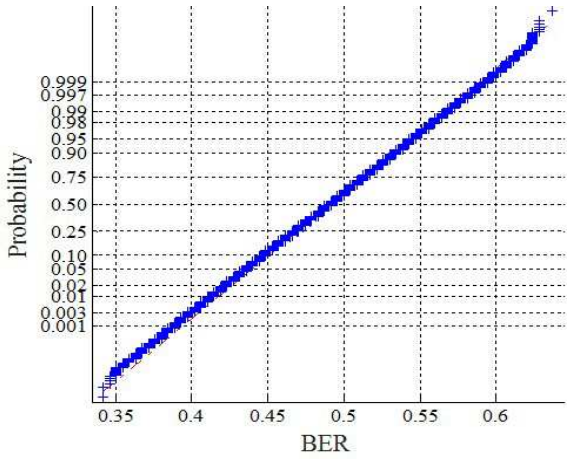

Figure 4: BER normal distribution diagram

The false accept rate (FAR) is commonly used to evaluate discrimination of a speech perceptual hashing system. It refers to the probability that the BER of two different perceptual contents is less than τ, and it is given by Equation (12).

$$FAR(\tau) = \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)}{2\sigma^2}} dx \qquad (12)$$

where τ is BER threshold, μ and σ are mean and standard deviation of BERs.

Through comparing proposed algorithm with three other algorithms, the Table 2 shows the FARs under different thresholds, and the conclusion may be drawn: when $\tau < 0.3$, with respect to the discrimination, the proposed method is better than MDCT-NMF and LPC-NMF and close to WPD-QT. This is primarily due to the following reasons, MDCT-NMF method is sensitive to the change of frame size, when sampling rate and frame number are set to 16 kHz and 360 in our experiment respectively, its FARs decreased dramatically; In WPD-QT method, the wavelet packet transform reflects frequency variation of speech signal well, so it has smaller FARs under different thresholds; In LPC-NMF method, there is a certain amount of error when LPC is used to describe vocal tract character. Therefore, the method shows a lower discrimination to different speech signal; for proposed method, different speech signals have obviously distinct spectrograms, so it gains a better discrimination.

Entropy rate (ER) is a comprehensive evaluation criterion on discrimination of perceptual hashing algorithm. It principally overcomes the disadvantages where the discrimination of algorithm is susceptible to hashing size. It ranges from 0 to 1, and the larger its value indicates the higher capacity of discrimination. It can be calculated from following Equation (13) and Equation (14).

$$ER = -[q \log_2 q + (1-q)\log_2 (1-q)] \qquad (13)$$

$$q = \frac{1}{2}\left(\sqrt{\frac{|\sigma^2 - \sigma_1^2|}{\sigma^2 + \sigma_1^2}} + 1\right) \qquad (14)$$

where σ and σ_1 are theoretical and experimental standard deviation of BERs respectively, q is experimental mean value.

As can be observed in Table 3, the entropy rate of proposed method is larger than MDCT-NMF and LPC-NMF except WPD-QT. Thus, by above analyses, proposed algorithm displays a better discrimination.

4.2 Robustness Analysis

Unlike the discrimination analysis, the robustness analysis phase requires to compare BERs yielded from original speech clips with their content-preserving manipulating speech. There are 10 types of content preserving operations shown in Table 4. In order to vividly demonstrate the influence of different content preserving distortions to the 4 s original speech spectrogram (it is shown on Figure 1(a)), six spectrograms are manifested in Figure 5. Compared with Figure 1(a), it can be clearly seen from Figure 5(b), Figure 5(c), Figure 5(d) and Figure 5(e) that the upper part of the spectrogram that reflects high frequency information of the speech signal is subjected to MP3 compression, filtering, noise and echo addition while their lower portions are seldom swayed by these distortions. Moreover, in Figure 5(a), because increasing volume causes the energy of speech fingerprint to rise, some new textures appear, which tends to be useless and interferes with extracting useful textual features. As also can be observed, there is no significant difference between Figure 5(f) and Figure 1(a), this indicates that resampling operation has less impact on Figure 1(a).

The mean and maximum of BERs of the proposed algorithm and three other methods in different content-keeping manipulations are presented in Table 5. In MDCT-NMF approach, when the sampling rate (its value in the original paper is 44.1 kHz while it is 16 kHz in our experiment) is reduced, the length of frame is decreasing on the condition of having same frame number, which results in containing less information in each frame and

Table 2: FAR under different threshold

τ	MDCT-NMF	WPD-QT	LPC-NMF	Proposed
0.10	2.94×10^{-21}	5.47×10^{-31}	1.48×10^{-22}	3.17×10^{-32}
0.15	1.14×10^{-16}	4.60×10^{-24}	1.05×10^{-17}	5.74×10^{-25}
0.20	1.11×10^{-12}	4.60×10^{-18}	1.73×10^{-13}	1.09×10^{-18}
0.25	2.75×10^{-09}	5.50×10^{-13}	6.75×10^{-10}	2.23×10^{-13}
0.30	1.68×10^{-06}	7.97×10^{-09}	6.25×10^{-07}	4.88×10^{-09}

Table 3: The comparison of entropy rate

Algorithm	MDCT-NMF	WPD-QT	LPC-NMF	Proposed
ER	0.5449	0.9510	0.6730	0.8308

Table 4: Content preserving operations

Type	Parameters	Abbreviation
Volume adjustment 1	-50%	V1
Volume adjustment 2	+50%	V2
Resampling 1	16-8-16 (kHz)	R1
Resampling 2	16-32-16 (kHz)	R1
Echo addition	100 ms, 0.5	E
Narrowband noise	AWGN,40 dB	NN
Low-pass filter 1	Butterworth filter, 3.4(kHz)	LP1
Low-pass filter 2	FIR filter, 3.4(kHz)	LP2
MP3 compression 1	32 kbps	M1
MP3 compression 2	128 kbps	M1

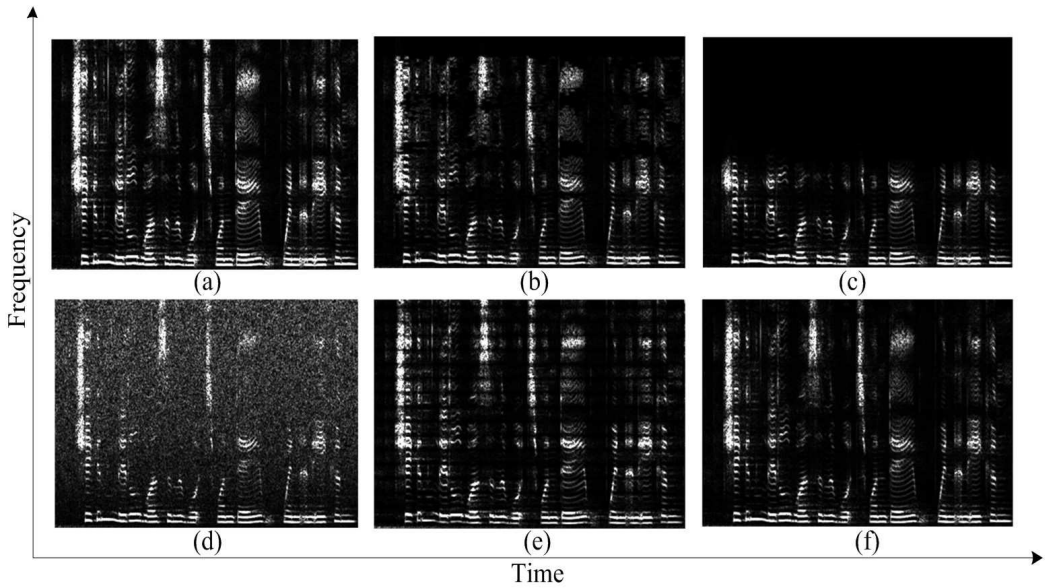

Figure 5: Spectrogram of (a) the 50%-volume-adding version of the 4 s original speech clip; (b) the 32 kbps-MP3-compressing version of the 4 s original speech clip; (c) the Butterworth-filtering version of the 4 s original speech clip; (d) 40 dB-noise-adding version of 4 s original speech clip; (e) echo-adding version of 4 s original speech clip; (f) resampling (16-8-16 kHz) version of the 4 s original speech clip

its degrading robustness. Although the wavelet packet transform can offer a more precise decomposition to signal frequency, the decomposing coefficients of signal high frequency are susceptible to noise. So the WPD-QT method is poor on noise resistance. Since the linear prediction analysis applies several past speech sampling values to approximate to current ones, therefore the change of amplitude has bad influence on LPC-NMF method. Because noise makes the spectrogram blurry, echo makes its textures overlapping. Therefore, the influence of noise and echo to proposed method is obvious. By the analysis above and combining the data in Table 5, the following information is obtained: in terms of robustness, the proposed algorithm outperforms LPC-NMF except the operation of volume addition, and it is also better than MDCT-NMF and WPD-QT on resisting the distortions caused by noise addition and filtering, but it is a little weaker than MDCT-NMF on resisting echo distortion.

In contrast to discrimination analysis, the false reject rate (FRR) is employed to estimate the robustness of a perceptual hashing system. It is the probability that the BERs of same two perceptual contents are more than τ. And its formula can be got from Equation (15).

$$FRR(\tau) = 1 - \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)}{2\sigma^2}} dx \qquad (15)$$

where τ is BER threshold, μ and σ are mean and standard deviation of BERs.

In order to describe the discrimination and robustness of the proposed algorithm more adequately, two kinds of BERs are utilized for probability analysis and drawing FAR-FRR curve in a coordinate system, one from the discrimination analysis in Section 4.1 and another from the robustness analysis in Section 4.2. Then, the robustness and discrimination of system can be evaluated by observing whether FAR curve and FRR curve cross, because if they have an intersection, the system cannot judge whether two speech clips are same perceptual contents in intersection area. Figure 6 shows the comparison of FAR-FRR curve between proposed approach and three other methods.

As can be seen in Figure 6(d), there is no one intersection on FAR-FRR curve of the proposed algorithm. Therefore, assuming the matching threshold τ is set to 0.3, when BER<0.3, the identification system can judge that two speech contents are perceptually same, otherwise different, which indicates the proposed algorithm has a better overall discrimination and robustness. For the discrimination analysis in Section 4.1, MDCT-NMF algorithm has a poor discrimination, and on the robustness analysis in Section 4.2, WPD-QT algorithm has a poor robustness against noise. Furthermore, there are bad effects on resisting the operations of noise adding and filtering in LPC-NMF algorithm. All these drawbacks of the three methods result in appearing an intersection on FAR-FRR curve in Figure 6(a), Figure 6(b) and Figure 6(c).

Through the analysis above, it is proved that the proposed algorithm shows satisfactory results on overall robustness and discrimination.

4.3 Efficiency Analysis

For illustrating the complexity and running efficiency of proposed algorithm, the running time is used to evaluate them with 100 speech clips selected randomly from the original speech database, which is took in the process of hashing generation and matching.

As can be seen from Table 6, compared with three other algorithms, the proposed algorithm takes a less time to generate and match hashing sequences. And its running efficiency is 4 times than LPC-NMF, 13 times than WPD-QT and 48 times than MDCT-NMF, which indicates that the proposed method obtains higher running efficiency. Furthermore, the hashing size in proposed algorithm is 234, 360 in LPC-NMF and MDCT-NMF, 250 in WPD-QT, which shows that the proposed algorithm has a stronger compaction.

From the analyses given above, the proposed algorithm has the advantages of high speed and few data, therefore it can meet the efficiency requirement of real-time speech communication.

5 Conclusions

In this paper, we proposed an efficient perceptual hashing scheme based on spectrogram for speech identification. By leveraging computer-vision method, the proposed algorithm adopts LBP to make texture information of a sub-spectrogram block more salient. As well as an image perceptual hashing method is utilized to generate hashing sequences from the sub-spectrogram block, which represents low frequency information of speech signal and is not sensitive to common content keeping distortions. Experimental results show that the proposed scheme achieves better discrimination to different speech clips and good robustness against some routine speech operations, such as noise addition, MP3 compression and filtering. Furthermore, the proposed scheme shows a high running efficiency and stronger compaction. This enables the proposed approach to be used in mobile speech real-time environment well.

There exist some issues to be handled in the proposed algorithm. For example, the robustness resisting echo and volume adjustment need to be further improved. And the performance analysis of speech fragment tampering attack has not yet been taken into account.

In future work, we will focus on extracting less and more salient features to overcome various degradations in spectrogram and research the tampering attack performance of proposed algorithm.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61363078), the Natu-

Table 5: Robustness test

Type	MDCT-NMF		WPD-QT		LPC-NMF		Proposed	
	Mean	Max	Mean	Max	Mean	Max	Mean	Max
V1	0.0040	0.0722	0.0008	0.0022	0.0016	0.0330	0.0179	0.0940
V2	0.0256	0.1056	0.0082	0.0100	0.0415	0.0917	0.0479	0.1624
R1	0.0012	0.0194	0.0036	0.0352	0.0260	0.1250	0.0094	0.0470
R2	0.0098	0.0750	0.0489	0.2266	0.1219	0.4028	0.0263	0.0149
E	0.0923	0.1827	0.1066	0.2305	0.2015	0.3000	0.1260	0.2051
NN	0.1357	0.2086	0.1452	0.5273	0.3464	0.5250	0.0918	0.2094
LP1	0.1422	0.2500	0.0864	0.2617	0.4098	0.5389	0.0784	0.1667
LP2	0.1615	0.2583	0.0924	0.2695	0.4303	0.5500	0.0813	0.1851
M1	0.0218	0.0722	-	-	0.1147	0.2920	0.1097	0.1838
M2	0.0035	0.0389	-	-	0.0727	0.2810	0.0248	0.0855

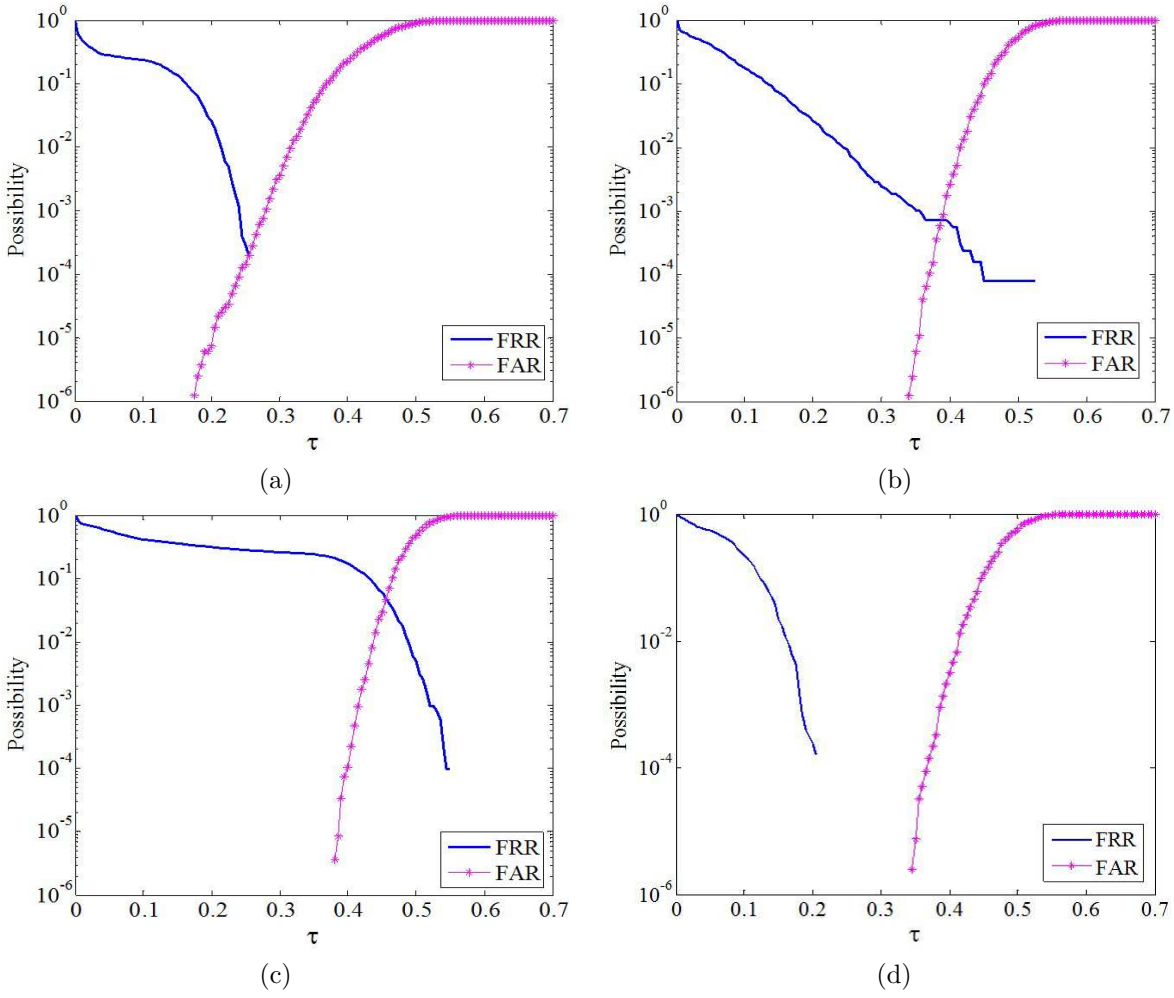

Figure 6: BER normal distribution diagram. (a) FAR-FRR curve of MDCT-NMF algorithm; (b) FAR-FRR curve of WPD-QT algorithm; (c) FAR-FRR curve of LPC-NMF algorithm; (d) FAR-FRR curve of proposed algorithm

ral Science Foundation of Gansu Province of China (No.1310RJYA004). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

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

[2] N. Chen, H. D. Xiao, J. Zhu, J. J. Lin, Y. Wang, and

Table 6: Running time

Algorithm	MDCT-NMF	WPD-QT	LPC-NMF	Proposed
File length(s)	4	4	4	4
Dominant frequency (GHz))	2.5	2.5	3.3	2.6
Total (s)	130.4	36.81	12.47	2.67

W. H. Yuan, "Robust audio hashing scheme based on cochleagram and cross recurrence analysis," *Electron Letters*, vol. 49, no. 1, pp. 7–8, 2013.

[3] N. Chen and H. D. Xiao, "Perceptual audio hashing algorithm based on zernike moment and maximum-likelihood watermark detection," *Digital Signal Processing*, vol. 23, no. 4, pp. 1216–1227, 2013.

[4] W. R. Ghanem, M. Shokir, and M. Dessoky, "Defense Against Selfish PUEA in Cognitive Radio Networks Based on Hash Message Authentication Code," *International Journal of Electronics and Information Engineering*, vol. 4, no. 1, pp. 12–21, 2016.

[5] Y. B. Huang, Q. Y. Zhang, and W. J. Hu, "Robust speech perception hashing authentication algorithm based on spectral subtraction and multi-feature tensor," *International Journal of Network Security*, vol. 20, no. 2, pp. 206–216, 2018.

[6] Y. B. Huang, Q. Y. Zhang, and Z. T. Yuan, "Perceptual speech hashing identification algorithm based on linear prediction analysis," *Telkomnika Indonesian Journal of Electrical Engineering*, vol. 12, no. 4, pp. 3214–3223, 2014.

[7] S. S. Kozat, R. Venkatesan, and M. K. Mihcak, "Robust perceptual image hashing via matrix invariants," in *International Conference on Image Processing (ICIP '04)*, pp. 3443–3446, Oct. 2004.

[8] J. F. Li, H. X. Wang, and J. Yi, "Audio perceptual hashing based on NMF and MDCT coefficients," *Chinese Journal of Electronics*, vol. 24, no. 3, pp. 579–588, 2015.

[9] 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*, vol. 38, no. 2, pp. 89–93, 2015.

[10] N. M. Makbol, B. E. Khoo, and T. H. Rassem, "Block-based discrete wavelet transform-singular value decomposition image watermarking scheme using human visual system characteristics," *IET Image Process*, vol. 10, no. 1, pp. 34–52, 2016.

[11] S. S. Nassar, N. M. Ayad, H. M. Hamdy, H. M. Kelash, H. S. El-sayed, M. A. M. El-Bendary, F. E. A. El-Samie, and O. S. Faragallah, "Efficient audio integrity verification algorithm using discrete cosine transform," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 1–8, 2016.

[12] Z. Rafii, B. Coover, and J. Y. Han, "Robust audio hashing scheme based on cochleagram and cross recurrence analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 644–648, May 2014.

[13] C. S. Rao and S. B. G. T. Babu, *Image Authentication Using Local Binary Pattern on the Low Frequency Components*, India, 2016.

[14] 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 Information Hiding Multimedia Signal Process*, vol. 6, no. 2, pp. 311–322, 2015.

[15] X. Zhang, B. L. Zhu, L. W. Li, W. Li, X. Q. Li, W. Wang, P. Z. Lu, and W. Q. Zhang, "Sift-based local spectrogram image descriptor: A novel feature for robust music identification," *EURASIP Journal on Audio Speech & Music Processing*, vol. 2015, no. 6, pp. 1–15, 2015.

[16] B. L. Zhu, W. Li, Z. R. Wang, and X. Y. Xue, "A novel audio fingerprinting method robust to time scale modification and pitch shifting," in *In Proceedings of the 18th ACM international conference on Multimedia*, pp. 987–990, Oct. 2010.

Biography

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

Zhang Tao. 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.

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

Wu Dong-fang. In 2015, Wu Dongfang obtained his bachelor of engineering degree from Northwest University

for Nationalities. Currently, he is studying for his master's degree at Lanzhou University of Technology. His research focuses on the industrial control network security.