
# PERCEPTUAL HASHING ALGORITHM FOR SPEECH CONTENT IDENTIFICATION BASED ON SPECTRUM ENTROPY IN COMPRESSED DOMAIN

Zhang Qiu-yu, Liu Yang-wei, Huang Yi-bo, Xing Peng-fei and Yang Zhong-ping

School of Computer and Communication

Lanzhou University of Technology, 730050

Lanzhou, Gansu

Email: zhangqylz@163.com; cc1000cc@126.com

***Submitted: Nov. 5, 2013*** ***Accepted: Feb. 8,2014*** ***Published: Mar. 10, 2014***

***Abstract- This paper proposes a new perceptual hashing algorithm for speech content identification with compressed domain based on MDCT (Modified Discrete Cosine Transform) Spectrum Entropy. It aims primarily to solve problems of large computational complexity and poor real-time performance that appear when applying traditional identification methods to the compressed speeches. The process begins by extracting the MDCT coefficients, which are the intermediately decoded results of***

***compressed speeches in MP3 format. In order to reduce the computational complexity, these coefficients are divided into sub-bands and the energy of MDCT spectrum is then calculated. Sub-bands of MDCT spectrum energy are then mapped to a similar mass function in information entropy theory. The function will be used as a perceptual feature and set to extract binary hash values. Experimental results show that the proposed algorithm keeps greater robustness to content-preserving operations while also maintaining efficiency. As a result of the partial decoding process, the real-time performance can meet the requirements of applications in real-time communication terminals.***

**Index terms*:* Perceptual speech hashing algorithm; Spectrum entropy; Modified discrete cosine transform; Compressed domain**

## I. INTRODUCTION

Human usually acquire information from the outside world by means of languages, images and words. Language signals contain the largest amount of information that can spread quickly and globally. Audio also plays an important role in area of human computer interaction systems. The information gathered from audio signals can be more trustable, helpful, and is some cases unique providers of information [1]. With the development of computers and network communication technology, the way of speech signal transmission and storage has also altered. The authenticity and integrity of speech signals have been questioned when tools for digital media editing are processed over an open and unreliable network. In the world of multi-media information security, these signals have become crucial to related studies. Traditional identification algorithms for speeches such as Signatures and Digital Watermark always focus on the integrity of digital structure rather than content of speech signals [2]. When studying issues of content identification and recognition in communication terminal applications like smart phones, prior training of a speech library is required [3]. Traditional identification algorithms are not suitable for real-time content identification in communication terminal applications.

A perceptual hash is a function that maps digital multimedia data based on human auditory models, into a compact digital digest. It was a method proposed by Ton Kalker in 2001 [4]. In recent years [5, 6] it has become central in research. With robustness, content-preserving operations and malicious content tampering discriminating abilities, this method can afford integrity identification of uncompressed music and speech content. The efficiency of the algorithm meets the demand of real-time applications in audio authentication, retrieval and recognition over an opening network because of the progress in related research.

The current perceptual hash algorithms are generally designed to accommodate uncompressed raw bandwidth audio placed in PCM format. Frequency domain characters are calculated by time-frequency transformation and used to extract perceptual features, including sub-band energy, MFCC (Mel Frequency Cepstral Coefficient), LPC (linear predictive coding), etc. [7][8][9]. When apply traditional algorithms to a compressed domain, transmission and storage of speech signals are unable to meet the demands of speech content identification if they exist in compressed formats, such as MP3. Standards often cannot be met because the complete decoding process calls for large computational complexity and differences between the hash values extracted from compressed and uncompressed speech signals [10], affect the accuracy of identification.

In order to solve these issues, Li Ming-yu from the Harbin Institute of Technology proposed a method to achieve identification of audio content in a compressed domain [11]. MDCT coefficients are extracted from MP3 files as intermediately decoding results in this method. The summation of the coefficients is calculated by a sub-band division and used to quantify binary hash values. Jiao Yu-hua developed a method of security and performance evaluating [12]. Perceptual encoding keeps these two MDCT-based methods highly robust to compressed audio and provides a solution for speech content identification in a compressed domain. Yet it continues to lack a large amount of identification data and high bit rate.

In summary, this paper proposes a MDCT-based algorithm for speech content identification using spectrum energy and Entropy in a compressed domain. This algorithm focuses on speech

signals compressed in MP3 format and their demand for efficiency and robustness. Experimental results show that the proposed algorithm remains highly robust in content-preserving operations and in turn, reduces the authentication data, while improving calculation efficiency.

## II. BACKGROUND

### *A. MP3 DECODING PROCESS*

Because of its high-fidelity audio quality and little amount of data, perceptual coding is widely used to compress wideband audio. As a function of time-frequency transformation and core step, MDCT (Modified discrete cosine transform) has been used in audio encoding processes such as MPEGI Layer-3 (MP3) and MPEG Advanced Audio Coding (AAC) [13]. The encoding process of a MP3 format is shown in Figure 1.

Audio data is encoded frame-by-frame, in accordance to MPEG standards. One MP3 frame consists of 2 granules, where each granule contains 576 samples per channel. Input of original sampling signals is firstly mapped to 32 sub-bands through the poly-phase filter banks. Each sub-band has the same bandwidth and corresponds to 18 coefficients [14]. A window function with two different lengths is used during transform, depends on whether echo suppression is needed. Both of the window functions export 576 MDCT coefficients that are listed in order of low to high frequency belts. MDCT coefficients can be acquired through frame decoding or by performing a modified Discrete Fourier transforms on 32 sub-band PCM (Pulse Code Modulation) signals, each at 18 MDCT. 576 MDCT coefficients are similarly extracted from the decoding process as inverse-quantization results are extracted in order.

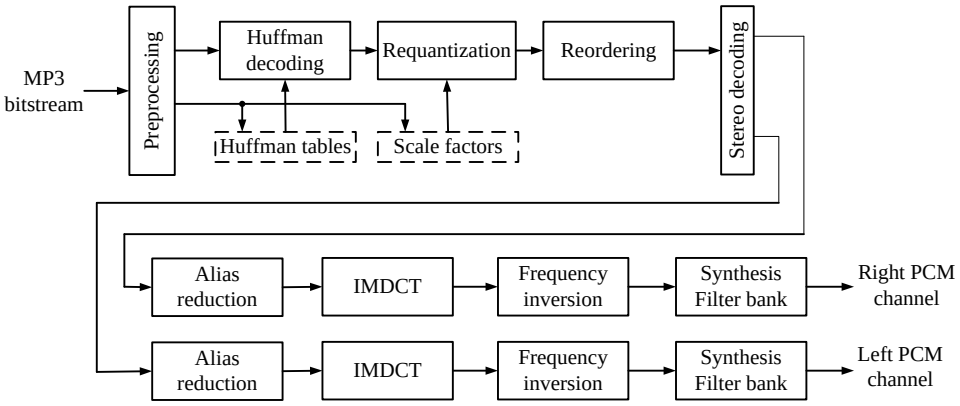


Figure 1. MP3 encoding process

Audio data is encoded frame-by-frame, in accordance to MPEG standards. One MP3 frame consists of 2 granules, where each granule contains 576 samples per channel. Input of original sampling signals is firstly mapped to 32 sub-bands through the poly-phase filter banks. Each sub-band has the same bandwidth and corresponds to 18 coefficients. A window function with two different lengths is used during transform, depends on whether echo suppression is needed. Both of the window functions export 576 MDCT coefficients that are listed in order of low to high frequency belts. MDCT coefficients can be acquired through frame decoding or by performing a modified Discrete Fourier transforms on 32 sub-band PCM (Pulse Code Modulation) signals, each at 18 MDCT coefficients. 576 MDCT coefficients are similarly extracted from the decoding process as inverse-quantization results are extracted in order.

### *B. SPECTRUM ENERGY*

The human cochlea can be equally compared to filter banks that simulate recognizable frequency to 26 critical bands ranging from 20 to 20000Hz. The human ear is sensitive to the energy of audio signals rather than audio phases; therefore sub-band energy of speech signals can be used as perceptual features to extract hash values. According to a Parseval theorem, periodic signals can be equivalent to superposition of each harmonic. As shown in (1), the power of the original signal is equal to the sum of square of Fourier coefficients:

$$E = \sum_{\pi=-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| X(e^{j\omega}) \right|^2 d\omega \quad (1)$$

It has been proven that MDCT coefficients can be obtained through linear superposition of the original signal (with weighting windows) and the Aliasing signal established, via SDFT (Shifted Discrete Fourier Transforms) [15]. This is shown in (2), where *u* and *v* represent shifts in time and the frequency domain.

$$SDFT_{u,v} = \sum_{k=0}^{2N-1} a_k e^{[i2\pi(k+u)(r+v)/2N]} \quad (2)$$

Moreover, the original DFT with half frequency shifting is considered the nature of MDCT coefficients through linear transformation. Additionally, it is assumed that time shifting does not occur. For this reason, it is possible to extract perceptual features from MDCT coefficients in an approximate version of frequency of the domain feature, when audio data using sub-band filtering [16] is processed.

### *C. SPECTRUM ENTROPY*

Entropy is closely related to audio content and thus has the ability to accurately represent features of audio signals. It is the expected information content in a sequence and the average of all the information contents weighted by their probabilities to occur. The information theory defines entropy as shown in (3).

$$H_i = \sum_{k-1}^{n} p_k \log p_k \quad i = 1,2,\dots,n \quad (3)$$

The entropy of a signal is also a measure of how unpredictable it is, the entropy should be minimum when the signal is a value since the signal is most predictable and the corresponding Probability Density Function is a unitary impulse and its entropy is zero. On the opposite case, if the signal has a uniform distribution then its entropy is maximum due to the fact that the sample values are most unpredictable and its entropy would be $log(n)$ . For a frame of 2.9721sec at a sampling rate of 44100 samples per second and a sample size of 16bits, each possible value would have to appear exactly twice.

In speech signals, the human voice has a clearer spectrum structure because of formant frequency and smaller entropy. The spectrum of environmental noise tends to have higher entropy [15]. While the spectrum entropy curves remain unchanged, entropy decreases as noise increases. Liu

Ya-duo [16] proved that the curve remains similar, even when distortion appears as time-frequency domain signal processing performs original compressed MP3 data flow. This characteristic of robustness makes it possible for MDCT spectrum entropy to be used as constant features to extract hash values.

## III. PROPOSED HASHING ALGORITHM

### *A. PROCESS OF ALGORITHM*

The proposed perceptual hash algorithm for speech content identification is based on MDCT coefficients in a compressed domain, as shown in Figure 2. MDCT coefficients are extracted using Libmad (MPEG Audio Decoder) [19], between reordering and alias reduction during the decoding process.

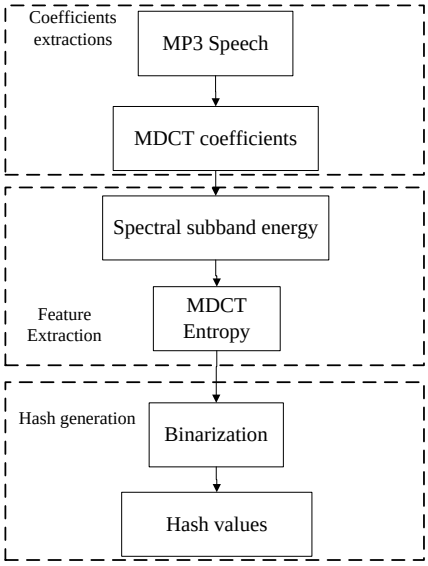


Figure 2. Process of algorithm

### *B. PERCEPTION FEATURES*

Although speech signal is a non-stationary stochastic process strictly after certain pretreatment (a general method for audio signal processing), it can be considered stationary. In wideband perceptual hash algorithms, pretreatments of window functioning, framing and aliasing are used

to makes the speech signal stationary in every 10-30*ms* clips. Given the fixed frame structure of MP3 files, we divide N granules into a sub-band with 50% overlapping as shown in Figure 3. This sub-band division functions similarly to pretreatments.

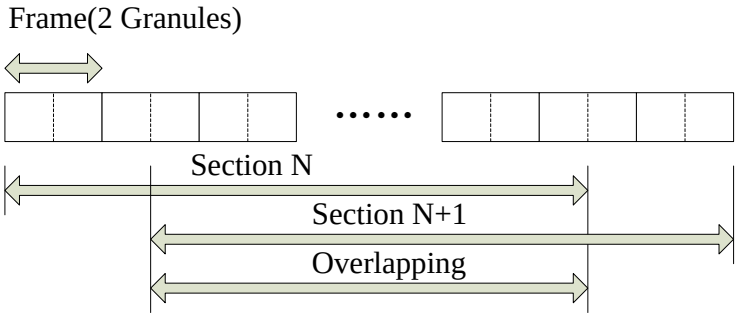


Figure 3. Sub-band division

With greater sub-band bandwidths, the algorithm will have higher robustness and lower accuracy. According to Kalker's theory [4], a 256-bit binary hash string can represent a complete, 3 seconds, audio clip. If we denote the number of granules of MDCT coefficients by *L* and the number of granules in each sub-band by *N*, *N* will be calculated as *N*=int (*L*/123.5). N is set for easy calculation.

MDCT coefficients in each sub-band function as do the Fourier coefficients, after pretreatments in the original power spectrum. If we denote energy of the *j*-th granule in *i*-th sub-band by $SBE_{ij}$, the sub-band energy is calculated as (4), where *G* (*i*, *n*) represents the *n*-th coefficient in *i*-th granule and *N* represents number of divided sub-bands.

$$SBE_{ij} = \sum_{m=\frac{N(i-1)}{2}+1}^{\frac{N(i+1)}{2}} \sum_{n=1}^{32} |G(m,n)|^2 \tag{4}$$

Here we divide MDCT spectrum energy of each granule by the energy of corresponding sub-band and denote the result by $p_{ij}$.

$$p_{ij} = \frac{SBE_{ij}}{\sum_{j=1}^{\frac{N(i+1)}{2}} SBE_{ij}} \quad i = 1,2,3\ldots,256 \tag{5}$$

In (5) the summation of $p_{ij}$ amounts to 1. This is similar to the mass function shown in (3). Entropy of MDCT spectrum of *i*-th sub-band is formally defined in

$$H(i) = -\sum_{j=1}^{N} p_{ij} \log_2 p_{ij} \quad i = 1,2,3,\ldots,256 \tag{6}$$

### *C. HASH GENERATION*

Perceptual features are translated by the methods of spectrum energy and entropy in the last section and used to generate an array of 256 elements. This allows for the characteristic of uni-directivity and compressibility. In order reduce the computational complexity and achieve higher robustness, this paper takes a method of comparing neighboring sub-bands entropy in order to achieve quantification of binary hash value string, as shown in (7).

$$Hash(i) = \begin{cases} 0, H(i) < H(i+1) \\ 1, H(i) \geq H(i+1) \end{cases} \tag{7}$$

### *D. HASH MATCHING*

Threshold values denoted by $\tau$ will determine whether a pair of 3 seconds speech signals are similar or tampered with, in comparison to bit error rates (BER) of hashing values extracted from the signals. Bit error rate, also namely the normalized hamming distance, is the ratio of error bits with total bits. It will be declared either similar when BER is below a certain threshold $\tau$, or tampered when BER is above $\tau$.

$$BER = \frac{\sum_{i=1}^{N} (hash_{new} \oplus hash_{origin})}{N} \tag{8}$$

## IV. EXPERIMENRAL RESULTS AND DISCUSSION

### *A. EXPERIMENTAL ENVIRONMENT*

In this paper, we present a full procedure of performance tests and their results. The database of 400 speech clips in our experiment, including clips with different content of Chinese and English and same content read by different people, is shown in Table 1. Each clip is compressed into MP3 format and lasts 3 seconds.

Table 1: Speech Clips

| Sampling Rate | Bit Depth | Channel | Bit Rate |
|---|---|---|---|
| 44100Hz | 16 bits | mono | 64kbps |

*B. DISCRIMINATION ANALYSIS*

In this paper, due to the random variable BER, the ability to discriminate different speech content is measured by the probability distribution. Hash values are extracted from 400 speech clips within the database and then compared in pairs. The 79800 BER results are shown in Figure 4, where the comparison of the distribution of BERs and the normal distribution is illustrated.

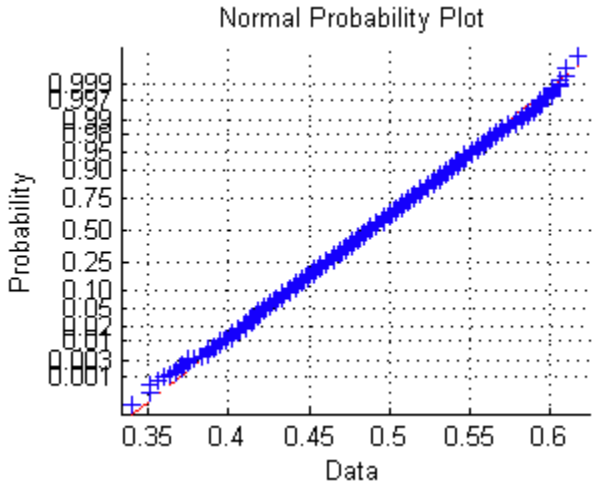


Figure 4. BER probability

It shows BER between speeches and with different content has a generally normal distribution. The probability distribution parameters are a mean value of $\mu$= 0.4869 and standard deviation of $\sigma$ = 0.0351, both calculated upon the Matlab platform. The false acceptance rate (FAR) of proposed algorithm is calculated as (9).

$$FAR(\tau) = \int_{-\infty}^{\tau} f(\alpha|\mu,\sigma)d\alpha = \int_{-\infty}^{\tau} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\alpha-\mu)^2}{2\sigma^2}} d\alpha \quad (9)$$

Table 2: False accept rate comparison

| | **This paper** | **Ref.[11]** |
|---|---|---|
| FAR | 1.6686e-18 | 1.0389e-15 |

In an ideal situation and where speech content is inconsistent, any pair of the hash values will have a high error rate. However, in a realistic situation, this will not often occur. There will always exist a small amount of data that is declared similar and consequently cause the BER to stand relatively lower than threshold. FAR increases in accordance to a higher threshold $\tau$. We arrive at a low FAR=1.6686e-18 at a threshold $\tau$=0.18, revealing that less than 2 clips were falsely claimed as similar to entire $10^{18}$ clips. Experimental results prove that the proposed algorithm meets the accuracy demands of speech identification in practical applications.

*C. PERCEPTUAL ROBUSTNESS ANALYSIS*

All of the 100 MP3 speech clips are subjected to the following procedures:

- Increase the volume by 50%.
- Reduce the volume by 50%.
- Resample consisting of subsequent down and up sampling to 22.05 kHz and 44.10 kHz.
- Echo addition with attenuation of 60%, time delay of 300*ms* and initial strength of 20% and 10%
- Noise addition with center frequency of 0~4 KHz
- Low-pass filtering, using a fifth order Butterworth filter with cut-off frequency of 2 kHz.

Each of the operations can preserve the perceptual content of speech signals except for the last signal.

Hash values are extracted from speech clips processed with the first five content-preserving operations and the BER between the hash values are determined. The values are extracted from clips with the same perceptual content. The resulting bit error rates are shown in Figure 5 (with same perceptual content).

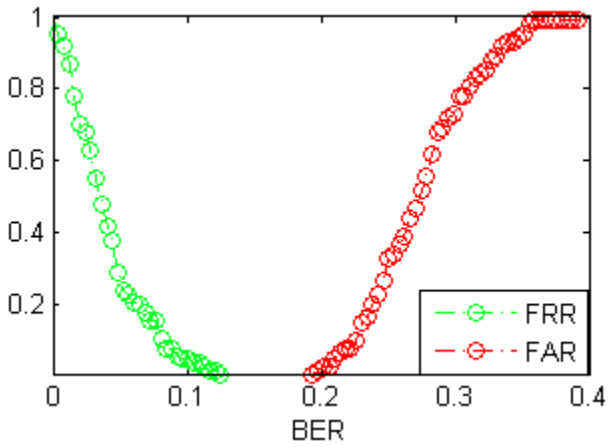


Figure 5. FRR-FAR curves

Afterwards, FAR curve of speech clips are subjected to low-pass filtering that is drawn from within the same FRR curve, coordinate system. The interval of discrimination between 0.14 and 0.19 makes it possible for the proposed algorithm to certify clips performed by content-preserving operations. It also allows for the discrimination of clips that have been subjected to malicious content tampering. When compared to the curves in Ref. [12] as shown in Figure 6, the interval decreases and the threshold $\tau$ will be set with a smaller value. However the final result of speech identification is not affected.

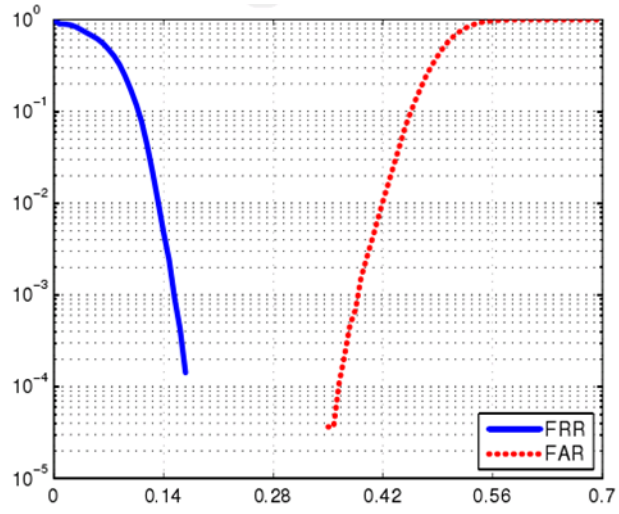


Figure 6. FRR-FAR curves in Ref. [12]

Table 3 lists the average bit error rate of each operation. When compared to algorithms in Ref. [11, 12, 20], the BER of content-preserving operations decreases significantly while it increases in content tampering operation of low-pass filtering. Experimental results show that the proposed algorithm keeps greater robustness to content-preserving operations while also maintaining discrimination abilities to content tampering operations.

Table 3: Average BER

| **Operation** | **BER** | **Ref.[11][12]** | **Ref.[20]** |
|---|---|---|---|
| Volume down | 0.0096 | 0.0451 | 0.0721 |
| Volume up | 0.0179 | N/A | N/A |
| Echo addition | 0.1872 | N/A | 0.148 |
| Resampling | 0.0068 | 0.0527 | 0.0041 |
| Noise addition | 0.0415 | N/A | 0.010 |
| Low-pass filtering | 0.2746 | 0.1943 | 0.0578 |

*D. EFFICIENCY ANALYSIS*

This algorithm is proposed with the purpose of dealing with speech communication terminals that have limited resources. Therefore the analysis of certification efficiency is necessary. An open source decoder named Libmad is used in to operate on 64kbps speech clips and is prepared as partial decoding. MDCT coefficients decoded from 100 seconds clips are translated to hash values in simulation platform of Matlab. The time consumption of each algorithm stage equates to 4% of the complete decoding process. This is shown in Figure 6. The proposed algorithm is much more efficient than traditional algorithms that have been fully decoded the compressed audio into PCM stream.

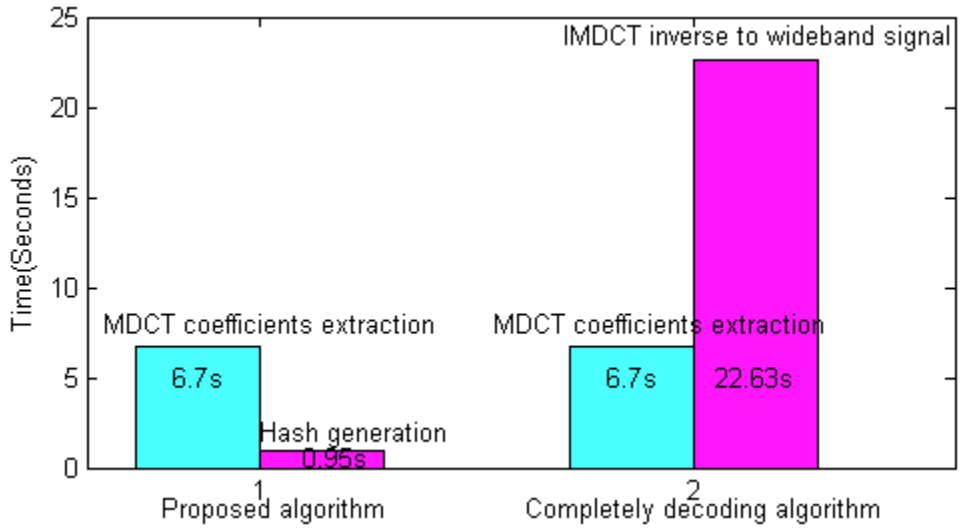


Figure 6. Time consuming column

Identification time for each pair of 3 seconds speech clips is stable at 0.01s. Stability is only reached with the preparation of the MDCT coefficient extraction. The proposed algorithm maintains good real-time performance after several experiments have been conducted.

In this paper 256-bit hash string is extracted from a 3 seconds speech clip and leads to a low bit rate of 245bps, with a sampling rate of 44.1kbps and calculated by (10).

$$\frac{44100}{576\times80}\times 256 \approx 245\text{bps} \tag{10}$$

Table 4 shows that the efficiency is increased compared with other algorithms in Ref. [11, 12].

Table 4: Bit rate of Algorithm

| | **This paper** | **Ref.[11]** | **Ref.[12]** |
|---|---|---|---|
| Bit Rate(bps) | 245 | 383 | 383 |

Figure 7 shows a GUI program in Matlab platform. It has simple identify function that determines whether a speech clip is tampered or content-preserved.

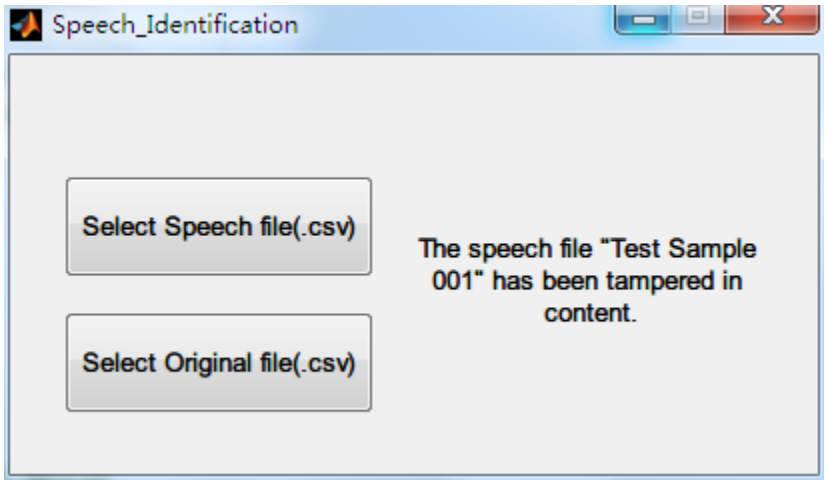


Figure 7. GUI Program of Identification

## V. CONCLUSIONS

In order to develop perceptual hash algorithms for speech content identification in compressed domain, this paper proposes an algorithm based on the entropy of spectrum energy. By partially decoding speech signals and operating the MDCT spectrum with sub-band division and overlapping, the MDCT coefficient was first translated in order to achieve a similar function as with the Fourier transform. Then the sub-band energy is calculated and mapped to an approximate mass function p. Hash values are finally extracted from the entropy of the function p. The proposed perceptual hash algorithm aims to mainly encompass the demands of robustness, discrimination and real-time performance. Experimental results show that the algorithm is highly robust when under echo and noise attack. As the bit rate decreases in comparison to Ref. [11, 12], real-time performance of the proposed algorithm improves. There is potential for further research, where the security of compressed speech can be explored in depth.

## ACKNOWLEDGMENTS

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

## REFERENCES

[1] F. Karray, M. Alemzadeh, J. A. Saleh and M. N. Arab, "Human-Computer Interaction: Overview on State of the Art", *International Journal on Smart Sensing and Intelligent Systems*, Vol. 1, No. 1, pp. 137-159, 2008.

[2] X. Niu and Y. Jiao, "An overview of perceptual hashing", *Acta Electronica Sinica*, Vol. 36, No. 7, pp. 1405-1411, 2008.

[3] N. K. Verma, R. K. Sevakula, J. K. Gupta, S. Singh, S. Dixit and A. Salour, "Smartphone Application for Fault Recognition", *International Journal on Smart Sensing and Intelligent Systems*, Vol. 6, No. 4, pp. 1763-1782, 2013.

[4] J. Haitsma, T.Kalker and J. Oostveen, "Robust Audio Hashing for Content Identification", *International Workshop on Content-Based Multimedia Indexing*, Vol. 4, pp. 117-124, 2001.

[5] G. Grutzek, J. Strobl, B. Mainka, F. Kurth, C. Porschmann and H. Knospe, "Perceptual hashing for the identification of telephone speech", *Speech Communication; 10.ITG Symposium, Proceedings of.VDE*, Germany, 2012, pp.1-4.

[6] Y. Jiao, L. Ji and X. Niu, "Robust speech hashing for content authentication", *IEEE Signal Processing Letters, Signal Processing Letters*, IEEE, Vol. 16, No. 9, pp. 818-821, 2009.

[7] J. Gu, L. Guo, H. Liang and L. Cheng, "Effective robust speech authentication algorithm based on perceptual characteristics", *Journal of Chinese Computer Systems*, Vol. 7, pp. 1461-1466, 2010.

[8] L. Ghouti and A. Bouridane, "A robust perceptual audio hashing using balanced multiwavelets", in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP'06)*, France, 2006, pp. 209-212.

[9] R. Lancini, F. Mapelli and R. Pezzano, "Audio content identification by using perceptual hashing", in *Proceedings of IEEE International Conference on Multimedia and Expo(ICME'04)*, Taipei, 2004, pp. 739-742.

[10] P. J. O. Doets, M. M. Gisbert and R. L. Lagendijk, "On the comparison of audio fingerprints for extracting quality parameters of compressed audio", *Electronic Imaging 2006, International Society for Optics and Photonics*, 2006, pp. 60720L-60720L-12.

[11] M. Li, "MDCT-based compressed domain perceptual audio hashing", Harbin, *Harbin institute of technology*, 2008.

[12] Y. Jiao, "Research on perceptual audio hashing", Harbin, *Harbin institute of technology*, 2009.

[13] P. Noll, "MPEG digital audio coding", *IEEE Signal Processing Magazine*, Vol. 14, No. 5, pp. 59-81, 1997.

[14] L. Chang, X. Yu, W. Wan, C. Li and X. Xu, "Research and realization of speech segmentation in MP3 compressed domain", *Jounal of Computer Applications*, Vol. 29, No. 4, pp. 1188-1192, 2009.

[15] Y. Wang, L. Yaroslavsky and M. Vilermo, "On the relationship between MDCT, SDFT and DFT", in *Proceedings of the 5th International Conference on Signal Processing*, Beijing, 2000, pp. 44-47.

[16] Y. Liang, C. Bao, B. Xia, Y. He, X. Zhou and N.Li, "Compressed domain speech enhancement based on Gaussian mixture model", *Acta Electronica Sinica*, Vol. 40, No. 10, pp. 2031-2038, 2012.

[17] H. Misra, S. Ikbal, H. Bourlard and H. Hermansky, "Spectral entropy based feature for robust ASR", in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'04)*, Canada, 2004, pp. I-193-6.

[18] Y. Liu, W. Li, X. Li, Z. Wang and R. Feng, "A robust compressed-domain music fingerprinting technique based on MDCT spectral entropy", *Acta Electronica Sinica*, Vol. 38, No. 5, pp. 1172-117, 2010.

[19] Underbit Technologies, Inc, "MAD: MPEG Audio Decoder, http://www.underbit.com/ products/mad", 2013.

[20] J. Haitsma and T. Kalker, "A Highly Robust Audio Fingerprinting System", in *Proceedings of International Symposium on Music Information Retrieval (ISMIR '02)*, Paris, 2002, pp. 107 – 115.