Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7):379-386

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

MFCC-based perceptual hashing for compressed domain of speech content identification

Qiu-yu Zhang[1]*, Yang-wei Liu[1], Yan-jun Di[1], Qian-yunZhang[2] and Peng-fei Xing[1]

[1]School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China
[2]School of Communication & Information Engineering, Shanghai University, Shanghai, China

ABSTRACT

Current research on speech content identification aim primarily at raw wideband speech signals, which are generally transmitted in a compressed format. This makes it unable to meet the demand of speech content identification in compressed domain. This paper proposes a new speech perceptual hashing algorithm for speech content identification with compressed domain based on MFCC (Mel Frequency Cepstral Coefficient), to solve problems of real-time speech content identification and large quantity of voice message information over the mobile Internet. This algorithm extracts MFCC feature based on the raw wideband method. The process begins by extracting the MDCT coefficients, which are the intermediately decoded results of compressed speeches in MP3 format. These coefficients are translated to MFCC parameters and the binary hashing values are then generated from these parameters, combined with human auditory features. This algorithm uses highly compressed data to realize fast identification for speech content. Experimental results show that the proposed algorithm can realize tampering localization and increase 5% in efficiency when compared with raw wideband algorithms, with the precondition of robustness and discrimination.

Key words: Speech content identification; perceptual hashing; compressed domain; MFCC features; robustness

INTRODUCTION

With the development of information technology, the authenticity and integrity of voice products have been questioned when tools for digital media editing are processed [1]. Numbers of speech content identification algorithms in the compressed domain are much less than traditional ones based on non-compressed format. A perceptual hashing is an easily computable function that maps digital multimedia data into a compact digital digest. These functions are widely applied in information security, where they are used as new algorithms for identification, retrieval and identification over an opening and unreliable network [2, 3].

Since the parametric speech coding is completely different from the way of audio compression, the audio hashing algorithm is unsuitable for speech algorithm [4]. Current researches on speech perceptual hashing usually take the original speech data as input. This large computational complexity can’t meet the demand of real time application in speech communication terminals with limited resources [5]. In Ref. [6] the author proposed a method with MELP (mixed excitation linear prediction) coding, using some parts of speech bit streams to generate hashing values. With malicious content-tampering discriminating abilities and less computational complexity, this algorithm is suitable for real-time system of speech content identification. In Ref. [7] proposed a perceptual hash algorithm designed for AAC (Advanced Audio Coding) audio to keep MDCT-based algorithms highly robust to compressed audio and provided a solution for speech content identification in a compressed domain.

Major process of traditional algorithm for the compressed speech data is as follows [8]. The process begins by decoding the compressed speech into raw wideband data (PCM). Features from decoded frames are then extracted and further analysis to achieve content identification is finally made. Yet it continues to have a flaw of computational efficiency and complexity, therefore it can't afford real-time processing.

Digital audio in practical applications is usually encoded in compressed formats such as MP3, with the purpose of less data size, higher quality and easier transmission. Therefore research on compressed domain has positive significance [9]. For this reason we propose the following algorithm to extract the MFCC features using human auditory system and MPEG audio coding. The process begins by decoding the MP3 stream and translating MDCT coefficients from intermediate parameters. The MDCT coefficients of each frame are then translated to a 15-dimension MFCC coefficient vector after Mel filtering. Content integrity certification is finally verified by matching the extracted hashing values.

MDCT COEFFICIENTS IN COMPRESSED DOMAIN

As a major vector feature of speech in the frequency domain [10], MFCC is robust because of its full consideration to the human auditory. In this paper we translate MFCC feature in compressed domain and select it as characteristic parameters. Physiological studies have shown that the human ear is very sensitive to the frequency of audio, especially in the range of 200~5000 Hz [9]. A feature vector can be calculated by the original content of audio, but not MP3 compressed version of that content because of its process such as filtering and MDCT transformation.

We extract the features from MDCT coefficients which are intermediate parameters when decoding an MP3 file. As a way of time-frequency transformation, the MDCT-based method has been widely used in encoding audios, such as MP3, AAC, etc. [11]. In accordance to MPEG standards, audio stream is encoded frame-by-frame. A MP3 frame consists of 2 granules and each granule contains 576 samples per granules [10]. We can get MDCT coefficients either by decoding each frame, or by performing a modified Discrete Fourier transforms on the 32 sub-band PCM (Pulse Code Modulation) signals. Each of the sub-bands corresponds to 18 MDCT coefficients. It has been proved that MDCT coefficients can be acquired through linear superposition of original signal (with weighting windows) and the aliasing signal performed with SDFT (Shifted Discrete Fourier Transforms) [12]. Moreover, with the assumption that there is no time shifting and the frequency shifting is 50%, we can consider the original DFT as the nature of MDCT coefficients through linear transformation. Thus make it possible for us to extract perceptual features from MDCT coefficient, for the reason that it is an approximate version of frequency domain feature when we process audio stream using a sub-band filter [13].

CONTENT IDENTIFICATION FOR COMPRESSED DOMAIN SPEECH

A. Process of algorithm

The MFCC-based perceptual hashing for speech content identification is shown in Fig. 1.In this figure we get MDCT coefficients by processing the compressed speech with Huffman encoding.

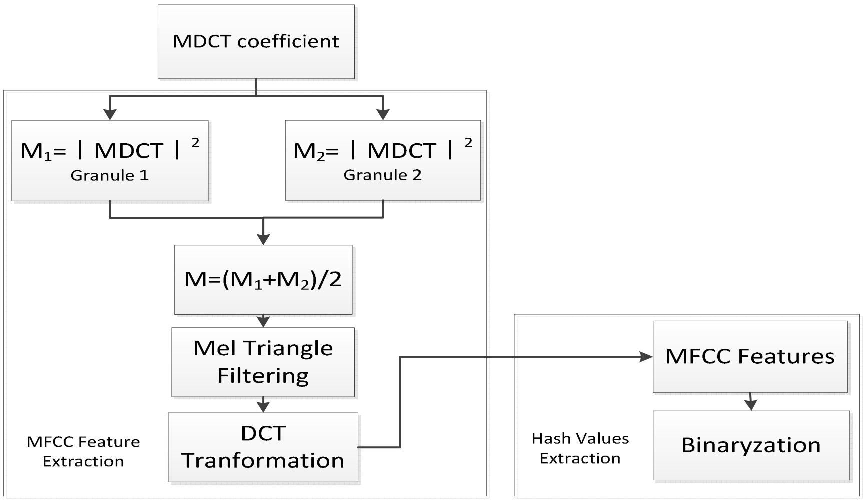

Fig. 1: Process of algorithm

B. MFCC feature extraction

MDCT coefficients can be regarded as an approximate version of linear spectrum of DFT [12]. Only considering the energy of these coefficients, we can extract features in the compressed domain according to MFCC algorithm in the non-compressed domain [14]. DFT coefficients with equal intervals are calculated after MP3 hybrid filtering. The difference is that none of these parameters divides the frequency spectrum into form of 2^n.

We propose this feature extraction method on the basic that MDCT coefficient contains enough information to describe frequency spectrum.Firstly, we redefine the MDCT coefficient of each frame to six critical bands, which are similar with the critical brand according to its bandwidth. Then an *n*-dimension MFCC feature parameters is calculated via triangle filtering the MDCT coefficients. In this paper, the redefine of critical band is not taken into account of when extracting features. A filter is performed before processing. Center frequency of Mel filtering and filter banks are determined based on the bits of calculated feature vectors. MDCT coefficient of each frame will finally be converted to a 15-dimension feature vector after Mel triangle filter banks followed by the cosine transformation.

It has lower resolution when compared to 576 parameters DFT in original audio without compression, but can afford identification for actual speech signals.

The details of the proposed method are expressed as follows.

Step1:*Intra-frame energy*. Because of the noise and estimation error of spectrum, the logarithmic energy of MDCT spectrum is calculated after a Mel filter to improve robustness. Considering time-varying in actual speech signals, the calculation process as follows is based on each frame.

The square of MDCT coefficients in each two granules of one frame is now calculated. The corresponding energy is denoted by $MDCT_1{}^2$ and $MDCT_2{}^2$ as shown in (1).

$$MDCT^2 = MDCT_1^2 + MDCT_2^2 \quad (1)$$

The mean value is calculated and the energy vector with 576 elements is given, which is accord with equal interval distribution in frequency domain.

Step2:*Mel triangle filtering*. Human perceptual auditory increases linearly with frequency in the range from 0Hz to 1000Hz, but they show a logarithmic relationship when frequency is above 1000Hz. We define 16 filters corresponding to the centre of Mel frequency to reduce computational complexity.

The upper - lower limit of filtering frequency (denoted by f_L and f_H) is mapped to the Mel frequency and the range is determined in (2).

$$\begin{cases} B(f_L) = 1125\ln(1 + f_L/700) \\ B(f_H) = \ln 1 + f_H/700) \end{cases} \quad (2)$$

In this formula we define a method to map the actual frequency to the Mel frequency, where B_H represents the upper bound of Mel domain and B_L represents the lower.

$$B_{Mel} = B_H - B_L \quad (3)$$

Using (3) we arrive at a Mel central frequency by dividing Mel bandwidth (B_{Mel}) into the number of filters equally and mapping central frequency of filters to corresponding frequency linear sequences.

$$FC(m) = \frac{N}{F_s} B^{-1}\left(B(f_L) + m \times \frac{B(f_H) - B(f_L)}{m+1}\right), 1 \le m \le 16 \quad (4)$$

In (4) we define $B^{-1}(b) = 700(e^{b/1125} - 1)$ as an inverse function of B and modified by inclusion of a factor of N/F_s in order to map center frequency to frequency linear sequence.

Here F_s is sampling rate and N =576 equal the number of MDCT coefficients in each granule.

The triangle filter is a function that calculates the component of frequency domain in range of Mel frequency and multiplies the MDCT energy amplitude by corresponding factors. Transfer function of Mel filter is shown in (5).

$$H_m(k)=\begin{cases}\dfrac{k-FC(m-1)}{FC(m)-FC(m-1)}, FC(m-1)\le k\le FC(m)\, k<FC(m-1)\, or\; k>FC(m)\\ \dfrac{FC(m+1)-k}{FC(m+1)-FC(m)}, FC(m)\le k\le FC(m+1)\end{cases} \tag{5}$$

The factors $1/(FC(m) - FC(m - 1))$ and $1/(FC(m + 1) - FC(m))$ can be seen as the filtering factors around center triangle filtering. These corresponding factors are different due to nonlinear bandwidth. Sequence number of filters is denoted by *m*.

See also the MDCT coefficient in Fig. 1, where *k* is corresponding to the coefficient ranging from 0 to 575.

Step 3:*Energy after filtering*. The triangle filter in the last step has a function of frequency division; therefore it can be used to process the energy coefficient in **Step1**. Given Noise Reduction, dynamic boundary of frequency spectrum and distribution of logarithmic energy spectrum, the output of the filter banks is calculated as (6).

$$X(m)=\ln(\sum_{k=0}^{575} MDCT^2\times H_m(k)),0\le m\le 15 \tag{6}$$

Where *m* and *k* represent sequence number of filters and $MDCT^2$ (possibly are 0 over high frequency).

Step 4:*Translation to cepstrum by DCT*. In order to assure the following decorrelation to different channels of MDCT spectrum, we perform a DCT transform on the output *X*(*m*) of filters.

$$Mel(n)=\sum_{m=0}^{15} X(m)\times\cos[\pi n(m+0.5)/16]),0\le n\le 15 \tag{7}$$

A 15-dimension MDCT vector is acquired using (7). However it makes distinguishing difference of these dimensions in content identification. In this paper we select the whole vector except for the first dimension considering its much less information.

C. Hashing values extraction

The 15-dimension coefficient vector is extracted from single frame of the speech signal as described in section B. Because of the real-time demand of speech signal and computation complexity of extracting hashing values frame-by-frame, the 10-dimension vector of every 10 frames is divided into a sub-band. Only binary sequences translated from eigenvector of these sun-bands are retained.

This method in (8) keeps robustness and unidirectivity as well as reduces the data quantity. Formula (8) formally defines the bits of the hashing string. The MDCT coefficient is denoted by $Mel_c(t, m)$, the *m*-th bit of the hash *H* in *t* sub-bands is denoted by *H*(*t*, *m*) and the threshold *T* is equal to zero.

$$H(t,m)=\begin{cases}1, Mel_c(t,m)\ge T\\ 0, Mel_c(t,m)<T\end{cases} \tag{8}$$

Finally we get a hashing block consisting of the m bit hashing string extracted from 10 subsequent frames (26*ms* per frame) with the above algorithm. The minimum precision of identification is 0.26*s* in speech content and tampering localization is achieved.

D. Hashing matching

Two derived threshold values denoted by τ_1 and τ_2 ($\tau_1<\tau_2$) will determine whether two 3 second speech clips are similar or tampered, by compared to bit error rates (BER) of hashing values which are extracted from the above clips. It will be declared either similar when BER is below a certain threshold τ_1, or tampered when BER is above τ_2. BER between τ_1 and τ_2 calls for a tampering localization detection.

RESULTS

In this paper, we present a full procedure of performance tests and their results. The database of speech clips in our experiment is shown in Table I.

Table I: Speech Clips

Sampling Rate	Bit Depth	Channel	Bit Rate
44100Hz	16 bits	mono	128kbps

The experiments environment platform is Windows7 operating system of Dell notebook, CPU is Inter Core i3-2450M, 2.4GHz and 2G memories, MATLAB R2010b.

A. Robustness analysing

All of the 1000 MP3 speech clips are processed as follows. Each of them can preserve the perceptual content of speech signals.

- Resample consisting of subsequent down and up sampling to 22.05 kHz and 44.10 kHz.
- Echo addition with attenuation of 60%.
- Increase the volume by 50%.
- Reduce the volume by 50%.
- Low-pass filtering using a fifth order Butterworth filter with cut-off frequencies of 2 kHz.

Thereafter the hash values are extracted from the speech clips which are processed with the first five content-preserving operations.

The BER between the hash values is then determined. The resulting bit error rates are shown in Fig. 2 (with same perceptual content).

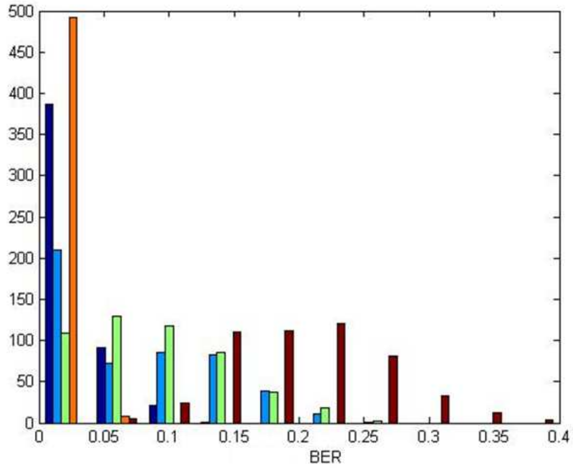

Fig.2: BER in 500 clips with same perceptual content

Here we arrive at a BER mostly below 0.3 from clips with same content, which ensure the robustness of the proposed method. Robustness is related to the extracted perceptual features as well as the threshold value.

Table II lists the ratio of clips declared equal using different threshold values. (These clips are subjected to different content-preserving operations).

Table II: Passing Rate

Threshold	*Volume down*	*Volume up*	*Echo*	*Resample*	*Low-pass Filter*
0.14	99.7%	77.6%	75.3%	100%	12.4%
0.18	100%	93.5%	93.4%	100%	30.2%
0.22	100%	98.7%	97.9%	100%	58.4%
0.27	100%	99.8%	100%	100%	85.4%
0.30	100%	100%	100%	100%	90.8%

Experimental results lead to the conclusion that we arrive at an extremely high identification precision. It also keeps high robustness to operations of resample and volume reducing with a threshold τ_2 of 0.3.

B. Discrimination analysing

In this paper we measure the discrimination ability for different speech contents with probability distribution because of the randomly variable BER.

Fig. 3 illustrates the comparison of the distribution of BERs and the normal distribution. It shows that BERs has a normal distribution approximately.

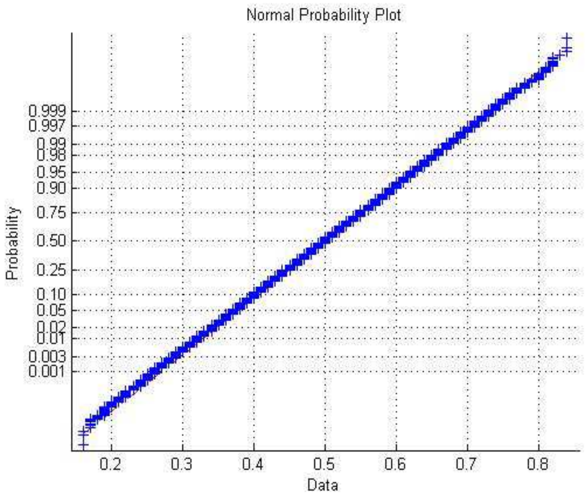

Fig. 3: Normal probabilityplot of BER among different speech content

The two contradictory parameters FRR and FAR can be used to measure the robustness and of ability discrimination respectively in proposed algorithm. In different applications it poses different emphases and FAR has slightly higher priority in our scheme to discriminate different and tampered clips.

Fig. 4 and Fig. 5 show the FRR-FAR curve of 500 speech clips which are randomly selected from speech database.

The cross point in Fig. 4 is cause by the weak robustness to low-pass filtering in the proposed method. The experimental results of other content-preserve processing are shown in Fig. 5.

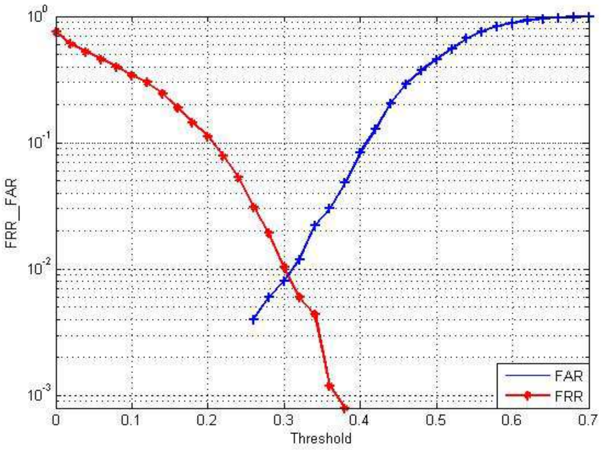

Fig. 4: FRR-FAR Curve with Low-pass filtering

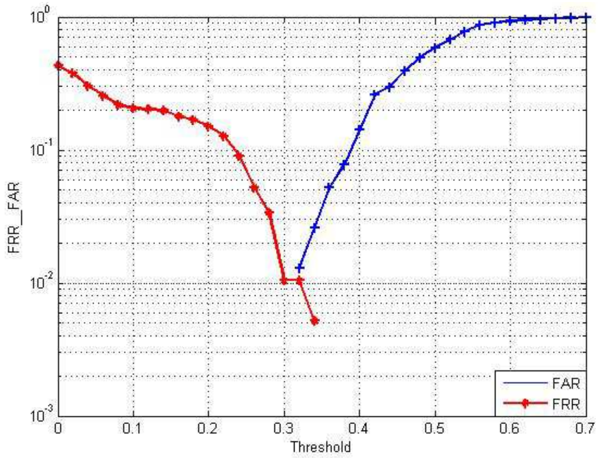

Fig. 5: FRR-FAR Curve without Low-pass filtering

It suggests that the proposed method is highly robust and able to discriminate between malicious content replacements and content-preserving operations, with the identification threshold value τ_2 of 0.30.

C. Performance analysing

The proposed method aims primarily at applications in communication terminals with limited resources. The efficiency of the algorithm can be measured with bit rate as (9).

$$\frac{44100}{576\times10}\times15\approx115bps \tag{9}$$

In this paper 15-dimension hash string is extracted from 10 frames (lasts 260*ms*) and leads to a low bit rate 115*bps*. This experiment process works on the platform of MATLAB 2010b, using 100 speech clips. Each clip is encoded at a 128k*bps* bit rate and lasts 4*s*. Experimental results show that the efficiency is increased by 5% compared with other raw wideband algorithm, which affords real-time applications.

D. Tamper location

A 7*s* clip randomly selected and cropped with two clips larger than 10 frames. Experimental results of malicious tampering are shown in Fig. 6.

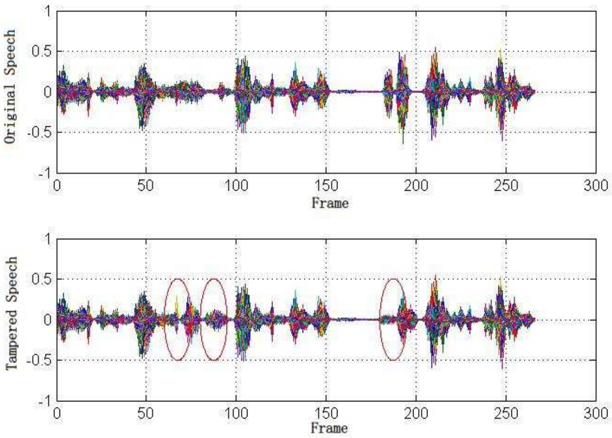

Fig.6: Tampering Detection and Location

Human speech rate is about 125 words of one minute and 480*ms* each word. The hash string in the proposed algorithm is extracted from 10 frames with time intervals of 260*ms*, which could be used to content tampering detecting and locating for one or more partial clips in speech signals.

CONCLUSION

In this paper we propose an identification algorithm for integrity identification of speech content in compressed-domain. This method is based on perceptual hashing algorithm and integrated with MFCC features, which are translated from intermediate parameters when decoding, named MDCT coefficient. Hash values are extracted from MFCC features based on raw wideband methods.

The experimental results show that the efficiency is increased by 5% compared with other raw wideband algorithms. The robustness and ability of discrimination is also maintained. As the precision of 260ms, the proposed method could be used in real-time identification as well as tampering detection and location. Based on the low cost of storage and computation we believe that this method has great value in certain applications.

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, No. 1310RJYA004). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

[1] Y. H. Jiao andX. M. Niu. *IEEE Signal Processing Letters*, **2009**, 16(9), 818-821.
[2] X. M. Niu and Y. H. Jiao. *Acta Electronica Sinica*, **2008**, 36(7), 1405-1411 (*in Chinese*).
[3] GUPTA S, CHO S, KUO C-C J. *IEEE Multimedia*, **2012**, 19(1), 50-59.
[4] Y. H. Jiao. Research on Perceptual Audio Hashing[Ph.D. dissertation]. Harbin:*Harbin institute of technology*,**2010**(*in Chinese*).
[5] J. Gu. Research on Key Technologies of Speech Perceptual Identification [Ph.D. dissertation]. Hefei: *University of Science and Technology of China*, **2009**(*in Chinese*).
[6] A. Shahbazi, A. H. Rezaie and R. Shahaazi. MELPe Coded Speech Hiding on Enhanced Full Rate Compressed Domain. In: *2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation*, Kota Kinabalu, Malaysia, **2010**, 267-270.
[7] Y. H. Jiao, M. Y. Li, B. Yang, *et al*. Compressed Domain Rubost Hashing for AAC Audio.In: *IEEE International Conference on Multimedia and Expo*, Hannover, **2008**, 1545-1548.
[8] Gary Grutzek, Julian Strobl, Bernhard Mainka, *et al*. Perceptual Hashing for the Identification of Telephone speech. In: *Proceedings of Speech Communication, 10. ITG Symposium*, Germany, **2012**, 1-4.
[9] Y. H. Jiao, Q. Li and X. M. Niu. Compressed Domain Perceptual Hashing for MELP Coded Speech. In: *Intelligent Information Hiding and Multimedia Signal Processing*, Harbin, China, **2008**, 410-413.
[10] L. Y. Chang and X. Q. Yu. *Journal of Computer Applications*, **2009**, 29(4), 1188-1192 (*in Chinese*).
[11] Y. J. Wang, L. Guo and C. P. Wang. *Journal of Chinese Computer Systems*, **2011**, 32(7), 1465-1469 (*in Chinese*).
[12] Y. Wang, Leonid P. Yaroslavsky, M. Vilermo. On the Relationship Between MDCT, SDFT and DFT. In: *Proceedings of the 5th International Conference on Signal Processing*, Beijing, China, **2000**, 44-47.
[13] Y. Liang, C. C. Bao. *Acta Electronica Sinica*, **2012**, 40(6), 2031-2038 (*in Chinese*).