A compressed-domain audio fingerprint algorithm for resisting linear speed change

Liming Wu[1], Wei Han[1, 2*], Songbin Zhou[2], Xin Luo[1], Yaohua Deng[1]

[1]*School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China*

[2]*Guangdong Institute of Automation, Guangzhou 510070, China*

Received 12 May, 2014, www.cmnt.lv

Abstract

Existing compressed-domain audio fingerprint algorithms have been able to be used to recognize audio information effectively according to hearing content, and are robust to common time-frequency domain distortion, including echo, noise, band-pass filtering, 32Kbps@MP3 and so on. However, they are poor in resisting linear speed change, which is a very common method for audio processing. In this paper, we propose a novel compressed-domain audio fingerprint algorithm. It is robust to large linear speed change via using auto-correlation function to reduce unaligned degree of MDCT spectrum sub-bands' energy. Besides, it is similar with existing compression-domain audio fingerprint algorithms on the other aspects.

Keywords: compressed-domain audio recognition, audio fingerprint, linear speed change, robustness

1 Introduction

As compressed format have become the main form in storage and transmission of audio files, extract fingerprint directly from compressed-domain audio for audio recognition appears more practical significance. Existing compressed-domain audio fingerprint algorithms [1-4] have owned good robustness to common time-frequency domain distortion, such as echo, noise, band-pass filtering, equalization, volume changing, 32Kbps@MP3, etc. However, they are insufficient for Linear Speed Change (LSC), which is an idiomatic mean of audio processing. Especially radio stations often play music with little of acceleration, the business reason is to shorten the time of playing music can bring more advertisements or other commercial purposes. On the other hand, most audiences maybe prefer the faster rhythm [5].

Nevertheless, it has not been seen in researching compressed-domain audio recognition after dealing with LSC. Only a handful of uncompressed-domain audio fingerprint algorithms [6-9] do it, and obtain good robustness. But this isn't conform to the status that compressed format has been mainstream.

In the remainder of this paper, Section 2 discusses the trouble caused by LSC, and introduces implementation process of the proposed compressed-domain audio algorithm. Section 3 expounds why auto-correlation function can be used to resist LSC. Section 4 utilizes experiments to test the performance of novel algorithm. Section 5 summarizes the full text.

2 Audio fingerprint scheme for resisting LSC

2.1 THE DIFFICULTY IN RECOGNITION DERIVED FROM LSC

Assume that original audio signal is $x(t)$, and its corresponding Fourier spectrum is $X(w)$. After dealing with LSC, the time-domain signal and frequency-domain signal will become $x'(t)$ and $X'(w)$ respectively.

$$x'(t) = x(t/\rho + t_0), \tag{1}$$

$$X'(w) = \int x'(t)e^{-jwt}dt = \int x(\rho/t + t_0)e^{-jwt}dt. \tag{2}$$

ρ is the scalability factor of LSC, and t_0 is the translational time.

After a series of calculation, the energy of $X'(w)$ can be expressed as follows:

$$|X'(w)| = \rho|X(\rho w)|. \tag{3}$$

To a large extent, MDCT spectrum is a linear approximation of Fourier spectrum, especially when only consider its energy [10]. Exactly, **MDCT** spectrum energy is selected to extract fingerprint in the proposed compressed-domain audio fingerprint algorithm. So hypothesize that δ indicates the linear relationship between MDCT spectrum and Fourier spectrum. Use $M(w)$ and $M'(w)$ individually denote the MDCT frequency-domain signal of original audio and distorted version leaded by *LSC*.

$$|M(w)| = \delta|X(w)|, \tag{4}$$

* ***Corresponding author*** e-mail: w.han@gia.ac.cn

$$|M'(w)| = \delta |X'(w)| = \delta\rho |X(\rho w)|. \quad (5)$$

Synthesize the above analysis, due to LSC will make the playing speed of audio faster or slower, which result in the change of MDCT spectrum energy. Worse, energy migration also happens as the variation of frequency of audio signal. However, most compressed-domain audio fingerprint algorithms extract audio fingerprint based on MDCT spectrum. This will induce that extracted audio fingerprint will have biggish difference before and after the distortion, which causes to reduce recognition rate.

Although all compressed-domain audio fingerprint algorithms have a large overlap degree between adjacent MDCT blocks [1-4], which will ensure that, even in the worst scenario, the unaligned extent of relevant blocks' border is very small. In other words, the sub-fingerprints of unknown audio clip waited to be identified are still very similar to the sub-fingerprints of the same clip in the database. Thus, algorithms can withstand a certain range of LSC. However, LSC will cause that MDCT coefficients are not aligned along time-axis and frequency-axis, and the unaligned degree would have a cumulative effect as times goes on. So that audio recognition will become more difficult with the increase of LSC degree.

2.2 THE PROPOSED COMPRESSED-DOMAIN AUDIO FINGERPRINT ALGORITHM

As shown in Figure 1 is the overview of the proposed compressed-domain audio fingerprint algorithm. In fact, for almost all compressed-domain audio fingerprint algorithms, most of their steps are similar, in addition to the fingerprint feature, which is derived from MDCT spectrum energy. In general, the eventual audio fingerprint is directly obtained by fingerprint feature with a chain of simple calculation. $F(n,m)$ denotes one bit fingerprint, n generally refers to a sub-band. And the fingerprint extracted from a sub-band is called a sub-fingerprint, whose length is m.

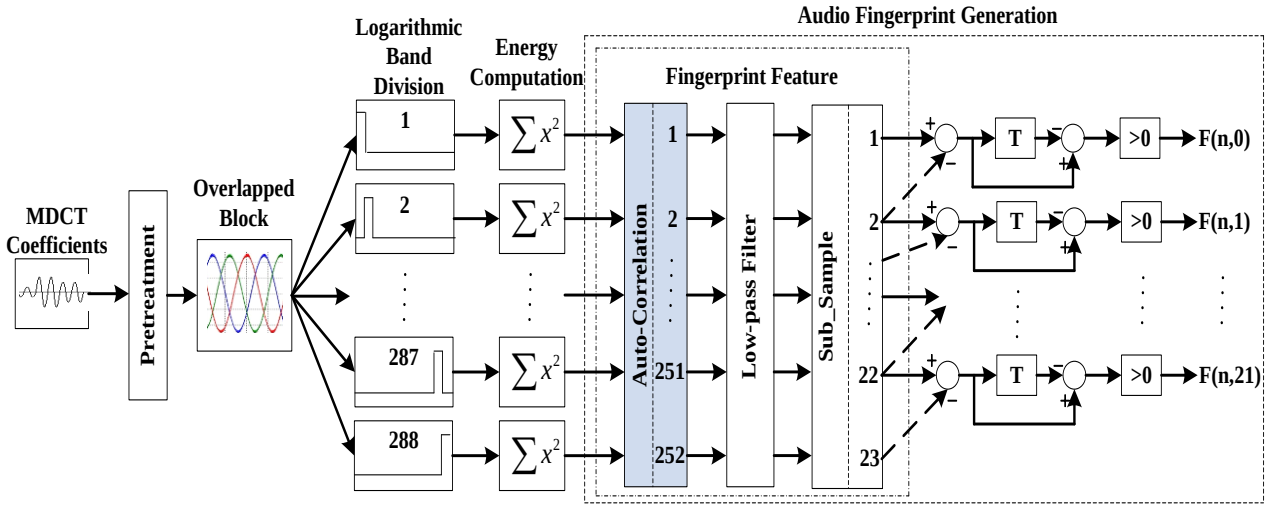

FIGURE 1 Overview of the proposed compressed-domain audio fingerprint algorithm

The proposed compressed-domain audio fingerprint algorithm will be started to introduce after MDCT coefficients originated from audio codec which can be accomplished according to reference [11]. At first, 12 frames MDCT coefficients acquired from the audio with MP3 format or *wave* format consist of one block with the length is about 0.31 seconds (when sampling rate is 44.1 KHz). The overlap degree is 95.83% between adjacent blocks, i.e., there is a section of hop distance because a MP3 frame contains two sections. Secondly, every section of MDCT spectrum (contains 576 coefficients) is divided into 288 sub-bands on the basis of logarithmic scale in the range of 300~2000Hz (the most relevant frequency range with the proposed fingerprint algorithm), and calculate the energy of each of them. Then sum these sub-bands' energy which locate in the same frequency range and pertain to an identical block. Assume that $SEN(i, j)$ represents the energy of the j-th sub-band which is belong to the i-th block, $S(m,n)$ indicates the energy of the n-th MDCT coefficient in the m-th section, $MDCT_p$ and $MDCT_q$ respectively represents the upper and lower bounds of MDCT coefficients which belong to the same sub-band. Therefore, the $SEN(i, j)$ can be calculated by Equation (6). Thus, in any block, an energy sequence composed of 288 sub-band's energy.

$$SEN(i, j) = \sum_{m=0.5(i+1)}^{12+0.5(i+1)} \sum_{n=MDCT_p}^{MDCT_q} |s(m,n)|^2 . \quad (6)$$

As has been analyzed in Section 2.1, LSC will lead to shift for the energy of MDCT sub-band. It is well known that auto-correlation function has the invariance to shift. Thus, it can be used to handle sub-band energy sequence of each block in order to resist LSC. And after this, the length of energy sequence reduced to 252. In order to improve the robustness, the results from auto-correction operation should be processed by a low-pass filter. Finally, the 252 energy values will be decreased to 23 energy data by down-sampling process. The purpose to shorten the length of energy sequence is to simplify the computational complexity at the time of fingerprint matching.

Suppose that *Ena*(*i*,*m*) (*i* is the number of MDCT block, *m*=0, 2,..., 21) represents the final 22 energy value. The fingerprint of an audio clip is actually a set of binary bit stream, and every bit of it defined by the Equation (7).

$$F(i,m)=\begin{cases}1 & if \quad Ena(i,m)-Ena(i,m+1)-Ena(i-1,m)+Ena(i-1,m+1)>0\\ 0 & if \quad Ena(i,m)-Ena(i,m+1)-Ena(i-1,m)+Ena(i-1,m+1)\le 0\end{cases}, \tag{7}$$

$$\rho_{ee}(x)=\sum_{j=1}^{M}e_1(k+j)e_2(x+j) \qquad 1\le x\le N-M\ . \tag{8}$$

As a result, a sub-fingerprint contains 22 bits can be extracted from a block. It does not own enough information to identify the corresponding complete audio clip. But fingerprint sequence, which is often referred to as a query fingerprint block, can do it. In the compressed-domain audio fingerprint algorithm of this paper, the length of audio fragment corresponds to a query fingerprint block is about 3s (114 MP3 frames). And a fingerprint block includes 204 sub-fingerprints, which has a total of 4488 bits. For example, if there is an audio clip with the length of 5 minutes, the length of unknown audio needed to index it in the fingerprint database only is 3s. Of course the unknown audio must be a part of original audio clip.

The Bit Error Rate (BER) is used to estimate the similarity between two audio clips. If the BER between query fingerprint block and one fingerprint block stored in the database beforehand is lower than the threshold *T*, it is considered to be a reliable match. A large number of experiments have proved that when the BER is less than *T*=0.35, it is seen that matching result is effective. Detailed index and match process can be implemented according to reference [12].

3 The principle of resisting LSC

Compared with the existing compressed-domain audio fingerprint algorithms, novel algorithm is added the auto-correlation processing. This is because auto-correlation function has invariance to shift, which can be proved as follows.

Suppose that *f*(*t*) represents audio signal, its auto-correlation coefficient $\rho_{ff}(x)$ is:

$$\rho_{ff}(x)=\int f(t)\,f(t+x)\,dt\ . \tag{9}$$

g(*t*)=*f*(*t*+*a*) generated by shift of *f*(*t*), its auto-correlation coefficient is $\rho_{gg}(x)$. $\rho_{gg}(x)=\rho_{ff}(x)$ can be inferred from the characteristic of auto-correlation function.

So, the auto-correlation coefficient of continuous function owns invariance to shift. But now is to deal with a discrete energy sequence. In order to approximate this characteristic of auto-correlation function, select a fixed sub-sequence e_1 from the complete energy sequence *e*(*n*) of each block (contains 288 energy values) and utilize it to do correlation calculation with any sub-sequence e_2 in the identical energy sequence. The auto-correlation coefficients $\rho_{gg}(x)$ of *e*(*n*) can be computed by Equation (8).

N stands for the length of entire energy sequence of each block, *M* represents the length of sub-sequence, and *k* denotes the starting position of a sub-sequence in the full sequence. In the proposed algorithms, *M*=36, *k*=54.

4 The effect of proposed algorithm

4.1 RESISTING LSC

Randomly choose 1000 audio clips (MP3 or *wav* format, stereo, 16 bit quantification, 44.1 KHz sampling rate) which belong to 10 different types of music, including DJ, electronic, classical, blues, jazz, folk, light music, hip-hop, country, rock and so on. The length of each clip is 20 seconds. Use various degree of LSC to distort each clip. Then severally extract fingerprint from the first 3 seconds in the distorted version and original version with existing algorithms and novel algorithm, and calculate each BER value between the fingerprints of an original version and its anamorphic version. Thus, 1000 BER data can be acquired for each treatment of LSC. The average value of these BER data embodies the ability of fingerprint algorithm to resist LSC. Figure 2 shows the relation between LSC and BER. It clearly shows that the proposed algorithm can resist LSC from –7% to +7%, which is obviously higher than the existing algorithms. In Figure 2 and Figure 3, the curve marked by "+" expresses the result of novel algorithm; the other curves demonstrate the results of exiting algorithms.

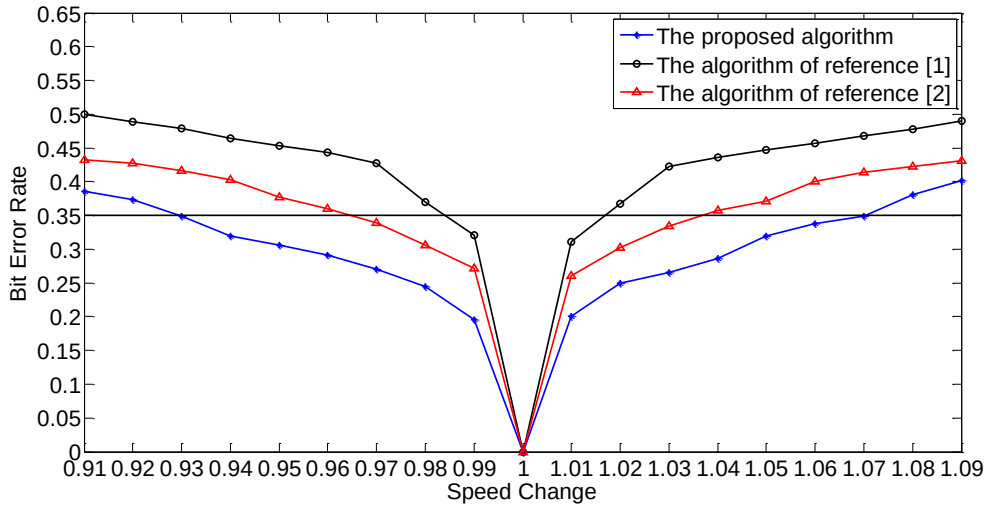

FIGURE 2 The robustness to LSC

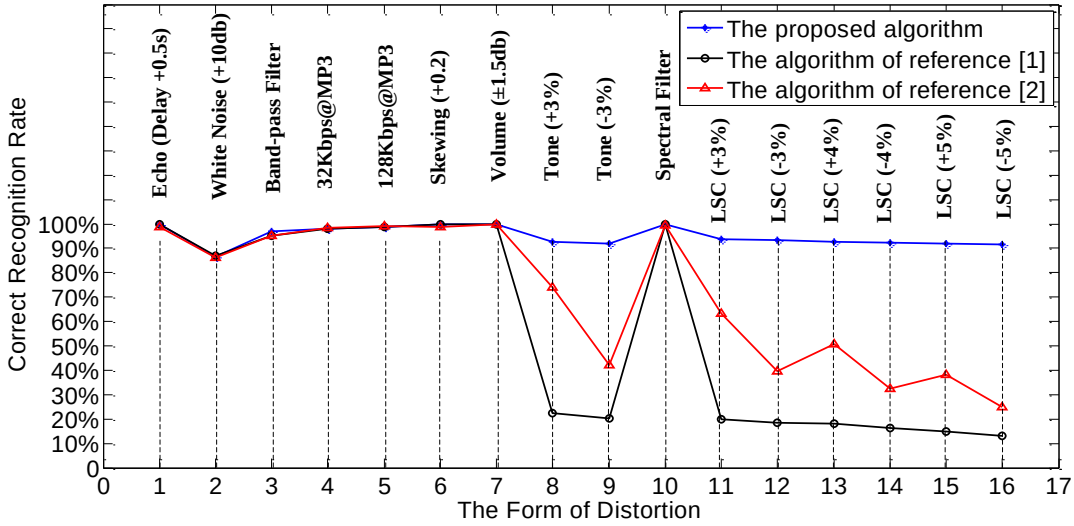

FIGURE 3 The comparison about recognition rate

4.2 THE ROBUSTNESS TO OTHER DISTORTION PROCESSING

Recognition rate, which reflects the reliability of algorithm, is the most important indicator to evaluate an audio fingerprint algorithm. Apply the following experiment to compare the recognition ability among novel algorithm, algorithm [1] and [2]. Experimental samples are the same as Section 4.1. Do 17 different distortion operations for each clip, including echo (delay 0.5s), noise (+10db), band-pass filter (300Hz~300Hz), 32Kbps@MP3, 128Kbps@MP3, offset (+0.2), volume (+1.5db), volume (–1.5 db), pitch (+3%), tone (-3%), spectrum filter (+1.0db), LSC (+3%), LSC (–3%), LSC (+4%), LSC (–4%), LSC (+5%), LSC (–5%). Randomly intercept a fragment with a length of 3s (or 114 MP3 frames) from each distorted audio clip. Thus there are 17000 unknown audio clips. Construct the fingerprint database according to the method of literature [12] for the fingerprints of original 1000 samples. And use the fingerprint of these 17000 unknown audio clips to search the homologous audio information in the database, the statistics of identification results are shown in Figure 3. The results clearly show that the novel algorithm is similar to or better than the existing algorithms for most of time-frequency-domain distortion. Besides LSC, novel algorithm also has better robustness to Tone change. This is due to Tone change is just a special case of LSC.

5 Conclusions

This paper proposes a novel compressed-domain audio fingerprint scheme. Experimental results show that the robustness of novel algorithm is similar to exiting algorithms for most audio distortion. But novel algorithm has better performance to deal with LSC, can resist it at the range from -7% to 7%. Generally, this is enough to handle the LSC in radio. Follow-up work of this paper will increase the resistance range to LSC, and improve the robustness for other common distortion.

Acknowledgements

This work was supported by Scientific Research Foundation of Guangdong Academy of Science for Young (Grant no. qnjj201306), High&New Technology Industrialization Project of Guangdong Province (Grant no. 2012B010100059), Science&Technology New Star of Zhujiang in Guangzhou City (Grant no. 2013J2200062).

References

[1] Wu L M, Han W, Deng U H 2013 *International Journal of Advancements in Computing Technology* **5**(9) 291-8
[2] Li W, Liu Y D, Xue X Y 2010 *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* Geneva Switzerland 627-34
[3] Chih-Chin Liu, Po-Feng Chang 2011 An efficient audio fingerprint design for MP3 music *The 9th International Conference on Advances in Mobile Computing & Multimedia* 190-3
[4] Sáenz-Lechón N, Osma-Ruiz V, Godino-Llorente J I, Blanco-Velasco M, Cruz-Roldán F, Arias-Londono J D 2008 *IEEE Transactions on Biomedical Engineering* **55**(12) 2831-5
[5] Cano P, Batlle E, Kalker T, Haitsma J 2005 A review of audio fingerprinting *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology* **41**(3) 271-84
[6] Liu J X, Zhang T X 2011 *International Conference on Computing, Information and Control* Wuhan China 360-68
[7] Zhu B, Li W, Wang Z, Xue X 2010 *18th ACM International Conference on Multimedia ACM Multimedia* Firenze, Italy 987-90
[8] Jin Soo Seo, Jaap Haitsma, Ton Kalker 2002 *Proceedints of the 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio* Leuven, Belgium 45-8
[9] Sun W, Lu Z M, Yu F X, Shen R J 2012 *International Journal of Digital Crime and Forensics* **4**(2) 49-69
[10] Wang Y, Yaroslavsky L, Vilermo M 2000 *Proceedings of the 5th International Conference on Signal Processing* Beijing China 44-47
[11] International Organization for Standardization 1993 *ISO/IEC 11172-3*
[12] Haitsma J, Kalker R 2002 *Proceedings of International Symposium on Music Information Retrieval* Paris, France 107-15

Authors

Liming Wu, born in January, 1962, Jieyang city, P.R. China

Current position, grades: professor of Guangdong University of Technology, Master Tutor.
University studies: graduated from the South China University of Technology in 2004 in China.
Scientific interests: speech processing, machine vision, optical, mechanical and electrical integration.
Publications: more than 80 papers.
Experience: teaching experience of 33 years, more than 20 scientific research projects.

Wei Han, born in July, 1987, Jingmen city, P.R. China

Current position, grades: assistant researcher of Guangdong Institute of Automation.
University studies: M.E. graduated from Guangdong University of Technology in 2013 in China.
Scientific interests: audio recognition, automatic control technology.
Publications: 8 papers.
Experience: 2 scientific research projects.

Songbin Zhou, born in July, 1978, Chaozhou city, P.R. China

Current position, grades: Doctor, Researcher of Guangdong Institute of Automation.
University studies: Dr.E. graduated from the South China University of Technology in 2008 in China.
Scientific interests: pattern recognition, intelligent sensing technology.
Publications: 15 papers.
Experience: 10 scientific research projects.

Xin Luo, born in April, 1989, Yichang city, P.R. China

University studies: Master's degree student of Guangdong University of Technology.
Scientific interests: Digital Image-processing, pattern recognition.
Publications: 3 papers.

Yaohua Deng, born in October, 1978, Huizhou city, P.R. China

Current position, grades: doctor, associate professor of Guangdong Institute of Automation.
University studies: Dr.E. graduated from in South China University of Technology in 2012 in China.
Scientific interest: intelligent measurement and control system, detection in flexible material processing, pattern recognition.
Publications: more than 20 papers.
Experience: teaching experience of 10 years, 6 scientific research projects.