Robust Music Identification Using Long-Term Dynamic Modulation Spectrum

Kim, Hyoung-Gook*, Eom, Ki-Wan*
*Samsung Advanced Institute of Technology Computing Lab

(Received May 2 2006; Accepted May 22 2006)

Abstract

In this paper, we propose a robust music audio fingerprinting system for automatic music retrieval. The fingerprint feature is extracted from the long-term dynamic modulation spectrum (LDMS) estimation in the perceptual compressed domain. The major advantage of this feature is its significant robustness against severe background noise from the street and cars. Further the fast searching is performed by looking up hash table with 32-bit hash values. The hash value bits are quantized from the logarithmic scale modulation frequency coefficients. Experiments illustrate that the LDMS fingerprint has advantages of high scalability, robustness and small fingerprint size. Moreover, the performance is improved remarkably under the severe recording-noise conditions compared with other power spectrum-based robust fingerprints.

Keywords: *Audio fingerprinting system, Long-term dynamic modulation spectrum, Hash value*

I. Introduction

Music identification systems can be implemented by using audio fingerprint techniques. Robustness is one of the key issues for practical applications. A robust fingerprint should be able to represent an audio clip uniquely in order to avoid false positive matching, and it also should be robust against signal distortions, for example amplitude changing, filtering, acoustic transmission and channel distortions, severe background noise, perceptual audio coding, etc.

Several methods have been proposed to improve the robustness of audio fingerprint systems. These approaches can be summarized into 3 typical categories: robust features methods, statistical methods and confidence matching methods.

In the robust feature domain, mean energy, spectral flatness measure and spectral crest factor are proposed in [1, 3]. By adding the adaptive quantization, the system can tolerate the attacks of cropping and noise addition. Significant signal position feature is proposed in [7] and it uses the difference information of the peak positions of the spectrum centroid. Energy difference feature is used by [2] and the similar features in multiple sub-bands are developed into the system in [8].

In the statistical method domain, robust subspace spanning, which chooses the subspace that minimizes the effect of feature distortion, reports the hit rate of 83.4% on the cellular real world data [4]. Distortion discriminant analysis used in [5-6] reports extremely low false positive and false negative rates. [6] proposes a robust hash value conversion through statistics estimation, adaptive quantization and error correction decoding.

Confidence matching method is adopted in several systems via using high-confidence index bits, bit conversion, or soft hashing method in the stage of hashing indexing. The main objective is to improve the robustness of hashing [1, 8].

This paper will focus on the compressed domain robust feature issue. Modulation spectrum has been used in [9] for robust speech recognition and the mel-cepstrum modulation spectrum replaces the conventional delta and double delta dynamic features. It increases the speech recognition robustness significantly. Also in [10] a wavelet transformed modulation spectrum is tested in

Corresponding author: Kim, Hyoung-Gook
(hyounggook.kim@samsung.com)
SAIT, Mt. 14-1 Nongseo-dong, Gihevng-gu, Younggin-si
Gyunggi-do, 449-712

fingerprint applications. The substantial robustness improvement against compression and equalization distortions is reported.

In this paper the compressed domain modulation spectrum is proposed. It is designed for robust MP3 fingerprint. We also report the evaluation results compared with other robust audio fingerprint features against the real world environmental noise and channel distortions. The significantly high robustness is achieved.

II. System Overview

Figure 1 gives the block diagram of the system proposed in this paper. Two distinct phases, such as database generation stage and query stage, be distinguished.

During the database generation stage, the bit vector extraction module extracts the index bit vectors and the fingerprint bit vectors, both based on the modulation spectrum estimation. Each fingerprint bit vector is indexed by four index bit vectors. Then they are stored in the database by implementing a linear hash table data structure.

In the query stage, the query module also extracts the bit vectors of the query clip. Then the two-stage searching method firstly locates the positions of the fingerprint bit vectors indexed by the index bit vectors; secondly computes the distances between the indexed fingerprint bit vectors and those of the query clip. The music piece with the minimal distance is evaluated whether to be the retrieval result.

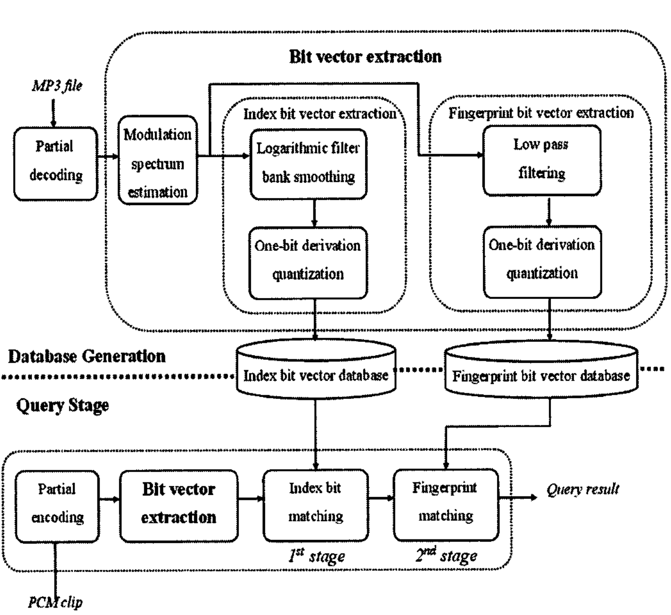

Bit vector extraction
MP3 file
Partial decoding
Modulation spectrum estimation
Index bit vector extraction
Logarithmic filter bank smoothing
One-bit derivation quantization
Fingerprint bit vector extraction
Low pass filtering
One-bit derivation quantization
Database Generation
Index bit vector database
Fingerprint bit vector database
Query Stage
Partial encoding
Bit vector extraction
Index bit matching
Fingerprint matching
Query result
1st stage
2nd stage
PCM clip

Figure 1. System block diagram.

III. Fingerprint Extraction

Unlike other music audio fingerprinting features based on the short-term spectrum estimation, which indicates the timbre properties of music, our feature is extracted from the long-term dynamic modulation spectrum (LDMS) estimation, which mainly characterizes the rhythmic property of music. The theoretical analysis proves that the modulation spectrum estimation is resistant to the linear and time-invariant convolutional distortion and the stationary additive noise. The fingerprint extraction process follows the steps below:

Step1. Partially decoding the MP3 frames to MDCT coefficients;

The MDCT coefficients are accessed directly from the compressed domain music data, via partial MP3 (MPEG 1 Layer 3) decoder. The detailed MP3 decoding description can be found in [11].

Step2. Performing a long-term Fourier transform on the MDCT sub-band signal along the time;

Modulation spectrum can be calculated from the MDCT coefficients (MS_{MDCT}) or the 32 sub-band values (MS_{SB}) by taking the DFT on them directly.

$$MS_{MDCT\,k,p}(q)=\sum_{n=0}^{Q-1}MDCT_k(p+n)e^{-j\frac{2\pi}{Q}nq} \tag{1}$$

In Equation (1) q is the modulation frequency index, p is the MS_{MDCT} frame index and k is the discrete frequency index of MDCT, from 0 to 575.

$$MS_{MDCT\,k,p}(q)=\sum_{n=0}^{Q-1}Sb_k(p+n)e^{-j\frac{2\pi}{Q}nq} \tag{2}$$

In Equation (2) q is the modulation frequency index, p is the MS_{SB} frame index and k denotes the 32 sub-band filters in the filter banks, from 0 to 31.

The MS_{MDCT} is invariant (except for scale) to the time-invariant filtering effect in nature; also it is very stable under the additive noise condition. Figure 2 shows the MS_{MDCT} snapshots from the noise corrupted same music clips. It can be seen that the MS_{MDCT} keeps almost the same even when the signal-to-noise ratio (SNR) is less than -1dB.

Step3. Getting the amplitude of the modulation spectrum;

Suppose that the results of long-term DFT at i-th frame are as

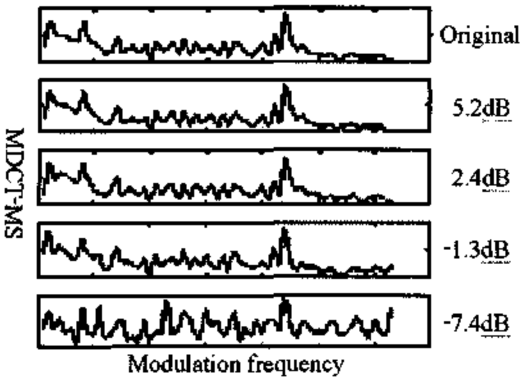


Figure 2. Invariance of MDCT MS to additive noise.

follows:

The real part is $\{MS_{MDCT,r}(i,0),\cdots\ !S_{MDCT,r}(i,N\text{-}1)\}$ and the image part is $\{MS_{MDCT,i}(i,0),\cdots\ !S_{MDCT,i}(i,N\text{-}1)\}$.

Then the amplitude of modulation spectrum is

$$AMS(i,n)=\sqrt{(MS_{MDCT,r}(i,n))^2+(MS_{MDCT,i}(i,n))^2}\quad n=0,\cdots,N-1 \tag{3}$$

where i is the index of frame, n is the index of modulation frequency, and N is the points of DFT.

Step4. Quantizing the smoothed amplitude by performing the one bit delta quantization;

The fingerprint bit vector *fbv* is transformed directly from *AMS* by taking the one-bit delta quantization along the modulation frequency dimension as Equation (3):

$$fbv(i)=\begin{cases}1, & AMS(i)>AMS(i+1)\\ 0, & otherwise\end{cases} \tag{4}$$

The index bit vector *ibv* is transformed to through the logarithmic filter bank smoothing process:

The logarithmic-scaled modulation frequency coefficients (*LMFC*) smoothed from *AMS* are transformed to the index bit vectors by taking the one bit delta quantization along the modulation frequency dimension as Equation:

$$ibv(i)=\begin{cases}1 & LMFC(i)>LMFC(i+1)\\ 0 & otherwise\end{cases} \tag{5}$$

Step5. Selecting M×N×T bits in N sub-bands (M bits per band per frame) of adjacent T frames to form the fingerprint block.

IV. Parameter Evaluation

The LDMS fingerprint is implemented by using several parameters: Q, *k*, *p* and *q*. The values influence the performance of a fingerprint system evaluated by scalability, robustness, speed and size. This section discusses the empirically optimal parameter selection with the emphasis on the robustness issue.

4.1. Data

1000 MP3 music songs are transformed from CD or downloaded from internet, including classical, rock, jazz, popular and folk genres. Environmental noise is recorded by a Samsung digital camera Digimax V4 on streets and in cars. Then around 2 hours of the wide band noise, when there is strong energy occupation in the up to 20kHz frequency span, are added to the decoded clean music data with different gain ratio, making noise corpus sets with different SNRs.

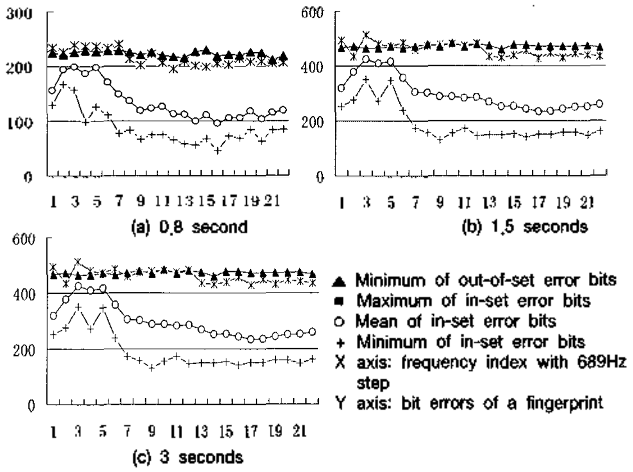
4.2. Parameter Design

4.2.1. Modulation Spectrum Length

Q is the DFT length taken on the MDCT coefficients. Longer Q results in higher modulation frequency resolution while on a longer time duration of music signals. The most appropriate duration is the trade-off between the hit accuracy and the computation resource requirement.

Figure 3 illustrates the error bits in different frequency range when the second DFT is taken over different Q length. The results of this experiment is tested on the 0~5dB set.

It is illustrated that 1 second signals have had good scalability, but under the strong noisy condition it is better to extract the fingerprint from 3 seconds signals because the in-set error bit



(a) 0.8 second

(b) 1.5 seconds

(c) 3 seconds

Figure 3. Error bits of different Q length.

curves and the out-of-set error bit curves can be separated marginally.

4.2.2. Frequency Range

k is the discrete frequency index, or the index of MDCT coefficients. Modulation spectrum can be taken only on several frequency lines with the advantages of

(1) Improving the robustness. Generally the energy density spectrum of the additive noise mainly resides in a limited low frequency range, as leaving the potential to select the more robust bands in which signal-to-noise ratio is approximately higher than in other bands.

(2) Fastening the searching speed potentially according to decreasing the fingerprint size

Figure 4 illustrates the different hit rate of LDMS fingerprint extracted in different frequency range.

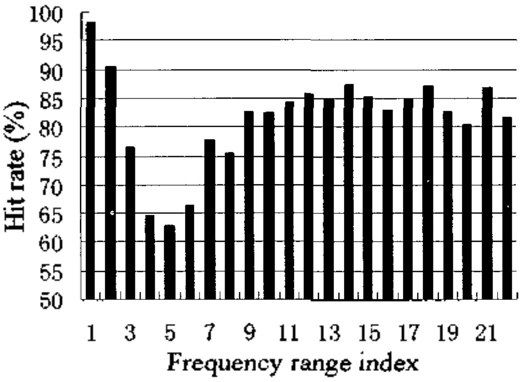


Figure 4. Hit rates of different frequency ranges (689Hz per each index step).

It is clear that the noise energy mainly resides in the frequency range from 1kHz to 4.5kHz. So the lowest bands below 1kHz or the higher bands above 4.5kHz are less prone to noise distortion. Although the 1kHz frequency range is incomplete for music recovery, it has included the first 5 octaves of the 7 octaves (the first 63 notes in the 88 piano notes), so the information should be sufficient for discrimination.

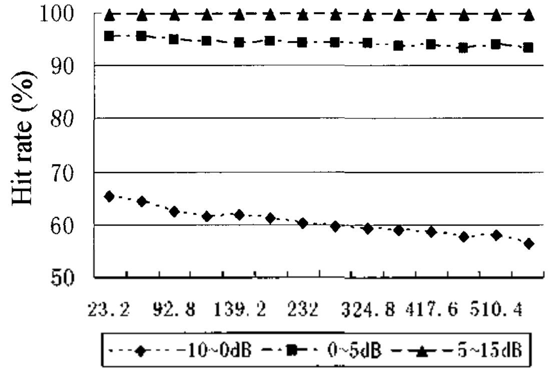


Figure 5. Hit rates of different frame shifts.

4.2.3. Modulation Spectrum Frame Shift

p is the LDMS frame index. The most appropriate value is the trade-off between the framing compensation and the fingerprint size and searching speed.

Figure 5 is the hit rate curves varied along with the different p values. It can be seen that the LDMS is a long-term feature, being stable during hundreds of milliseconds. So the longer frame shift can be used for modulation spectrum than other short-term features.

4.2.4. Modulation Frequency Range

q is the discrete modulation frequency index. Only the first half span, from 45Hz to 400Hz, of the valid modulation spectrum is used to make the fingerprint because the MDCT lines on which the modulation spectrum is estimated only occupy the frequency range below 1kHz.

V. Comparison Results

LDMS is compared with other robust features. Several features are considered here as follows:

(1) *SFM*, spectral flatness measure [3];

(2) *SCF*, spectral crest factor [3];

(3) *MMAX*, maximum of energy in multiple sub-bands;

(4) *MMG*, geometric average value of the energy in multiple sub-bands;

(5) *MNE*, arithmetic average value of the energy in multiple sub-bands;

(6) *DLE*, derivation of sub-band energy in time and frequency dimensions [8].

Among them, *SFM*, *SCF*, *MMAX*, *MMG* and *MNE* are estimated in 8 equal-spaced sub-bands, and then transformed to fingerprint bits via bit derivation; *DLE* is implemented in a fingerprint system as described in[8]. The fingerprint is then

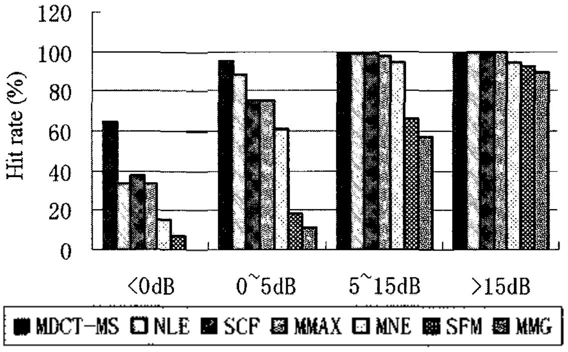


Figure 6. Hit rates of different robust features (False Positive Rate: 1.00%)

searched by minimum Hamming distance. Figure 6 shows the results.

It is clear that although all the features achieve high hit rate in case of above 15dB SNR, only the LDMS based on MS_{MDCT} fingerprint gains high hit rate under the severe noise conditions. Even when the SNR is under 0dB its hit rate is still above 60% while other fingerprints have totally failed. MS_{MDCT} hit rates are 100%, 99.77% and 95.41%, respectively, when the SNR is above 15dB, 5~15dB and 0~5dB, as proves its remarkable robustness as an audio fingerprint.

VI. Conclusions

This paper discusses the compressed domain modulation spectrum feature, and designs the fingerprint based on the robustness evaluation results. The LDMS fingerprint is compared with other robust fingerprints under the strong background noise conditions. The LDMS fingerprint illustrates the significant robustness performance.

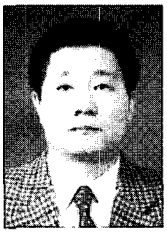
A simple bit derivation method is used here to quantize the LDMS feature into audio fingerprint bit stream. Other quantization methods deserve study to improve the robustness further. Our future work is to test the robustness of the proposed algorithm against various distortions, such as channel distortion. Furthermore experiments based on larger database will be tested in the further.

References

1. Rosa Lancini, Francesco Mapelli, and R. Pezzano, "Audio Content Identification By Using Perceptual Hashing", Proc. of the 2004 IEEE International Conf. on Multimedia and Expo, 7392, 2004.
2. F. Mapelli, and R. Lancini, "Audio Hashing Technique For Automatic Song Identification", Proc. of the International Conf. on Information Technology: Research and Education, 2003.
3. Jurgen Herre, Oliver Hellmuch, and Markus Cremer, "Scalable Robust Audio Fingerprinting Using MPEG-7 Content Description", IEEE Workshop on Multimedia Signal Processing, 2002.
4. Takayuki Kurozumi, Kunio Kashino, and Hiroshi Murase, "A Robust Audio Searching Method for Cellular-Phone-Based Music Information Retrieval", Proc. of the International Conf. on Pattern Recognition, 991~994, 2002.
5. Christopher JC Burges, J.C.Platt, and S.Jana, "Distortion Discriminant Analysis for Audio Fingerprinting", IEEE Trans. on Speech and Audio Processing, **11** (3), 165~174, 2003.
6. MK Mihcak, and R. Venkatesan, "A Perceptual Audio Hashing Algorithm: A Tool For Robust Audio Identification and Information Hiding", Proc. Of 4th International Information Hiding Workshop, 2001.
7. Andreas Ribbrock, and Frank Kurth, "A Full-Text Retrieval Approach to Content-Based Audio Identification", IEEE Workshop on Multimedia Signal Processing, 194~197, 2002.
8. Jaap Haitsma, and T. Kalker, "A Highly Robust Audio Fingerprinting System", Proc. of the International Conf. on Music Information Retrieval, 14~17, 2002.
9. Vivek Tyagi, Iain McCowan, Hemant Misra, and Herve Bourland, "Mel-Cepstrum Modulation Spectrum (MCMS) Features For Robust ASR", IEEE Workshop on Automatic Speech Recognition and Understanding, 2003.
10. S. Sukittanon, and L. Atlas, "Modulation Frequency Features For Audio Fingerprinting", Proc. of the International Conf. on Acoustics, Speech, and Signal Processing, 2002.
11. ISO/IEC 11172-3, "Coding of Moving Pictures And Associated Audio For Digital Storage Media, Part3: Audio". 1993.

[Profile]

•Kim, Hyoung-Gook



received the diploma degree in electrical engineering and the Ph. D. degree in computer science from the Technical University of Berlin, Berlin, Germany. From 1998 to 1999, he worked on speech recognition at Siemens AG. From 1999 to 2002, he was a Project Leader of the Speech Processing Laboratory at Cortologic AG. From 2002 to 2005, he served as Adjunct Professor of the Communication Systems Department, Technical University of Berlin. Since 2005 he joined Samsung Advanced Institute of Technology as a Project Leader. His research interests include music information retrieval, audiovisual content indexing and retrieval, speech enhancement, and robust speech recognition.

•Eom, Ki-Wan

received the B.S. degree in 1996 from the department of electronic engineering of Gwangju University, Gwangju, Korea, and the M.S. degree in 1998 and the Ph.D. degree in 2006 from the department of electronic engineering of Chonnam National University, Gwangju, Korea. In 2002, he joined Samsung Advanced Institute of Technology as a research staff member. His research interests are audio indexing, music processing, and speech synthesis.