AUTOMATIC MUSIC SUMMARIZATION IN COMPRESSED DOMAIN

*Xi Shao[#] *, Changsheng Xu[#], Ye Wang* , Mohan S Kankanhalli**

[#]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{shaoxi, xucs}@i2r.a-star.edu.sg
*School of Computing, National University of Singapore
{wangye, mohan}@comp.nus.edu.sg

ABSTRACT

A novel compressed domain automatic music summarization approach is presented in this paper. The proposed method works directly in the compressed domain. Only the encoded subband samples are extracted and processed for characterizing music content and discovering the music structure. The experimental results and the evaluation by a subjective study have shown that the summarization based on MPEG-1 Layer 3 (MP3) music is comparable to the summarization based on uncompressed PCM music samples.

1. INTRODUCTION

With the rapid increase of the size of digital multimedia data collections and greatly increased availability of multimedia content for the general user, how to create a concise and informative extraction that best summarizes an original digital content is a challenge in music content analysis and is extremely important in large-scale information organization and processing. Nowadays, many music companies are putting music products on websites and customers can purchase them on line. But from the customer point of view, they would prefer to listen to the highlights of the music before they make a decision on whether to purchase or not. Although summaries are available on some websites, they are generated manually, which needs expensive manpower and is time-consuming. Therefore, it is crucial to come up with an automatic summarization approach for music.

The first music summarization system [1] was developed on the MIDI format. But MIDI is a synthesizer and structured format and is different from sampled audio format such as wav which is highly unstructured. Therefore, MIDI summarization method cannot be applied to real music summarization. Several automatic music summarization methods [2, 3, 4, 5] have been proposed in recent years, but all these methods summarized music in uncompressed domain. Due to the huge size of music data and the limited bandwidth, audio/music analysis in compressed domain is in great demand. Research in this field is still in its infancy and there are many open questions to solve.

There are a number of approaches proposed for compressed domain audio processing. Most of the wok focuses on compressed domain audio segmentation [6, 7]. Compared to compressed domain audio processing, compressed domain music processing is much more difficult, because a musical song consists of many types of sounds and instrument effects. Wang and Vilermo [8] have used the window type information encoded in MPEG-1 Layer 3 side information header to detect beats. The short windows are used for short but intensive sounds to avoid pre-echo. They found that the window-switching pattern of pop-music beats for their specific encoder at bit-rates of 64-96 kbps gives (long, long-to-short, short, short, short-to-long, long) window sequences in 99% of the beats.

However, there is no music summarization method available for compressed domain. Due to the large amount of compressed domain music (e.g. MP3) available nowadays, automatic music summarization in compressed domain is in high demand.

In this paper, we proposed a novel approach to summarize the music directly in compressed domain. Firstly, we extract features directly from compressed domain. Then we use clustering method to capture the music structure. Finally, the music summary is generated based on clustering results and music domain knowledge.

2. FEATURE EXTRACTION

In our experiment, MPEG-1, Layer 3, 44.1 KHz sampling rate, mono channel music samples are used. The reason we conduct summarization using MP3 samples is twofold. Firstly, considering the large amount of MP3 music available nowadays, research in this field is more relevant in practical terms. Secondly, the summarization method conducted on MP3 can be easily used with other layers as well as any filterbank-based perceptual coder.

Feature extraction is very important to music summarization. Considering the features [2, 5] that have been used to characterize music content for

summarization in uncompressed domain, we have developed similar features in the compressed domain to simulate those features commonly used in the uncompressed domain. The analysis is performed on one MP3 granule (512 samples, about 13ms).

2.1. Amplitude Envelope

The amplitude Envelope describes the energy change of the signal in the time domain and is generally equivalent to the so called ADSR (attack, decay, sustain and release) of a musical sound.

Here, we use the root mean squared (RMS) values [6] on a granule resolution to approximate the amplitude Envelope.

$$RMS(t) = \sqrt{\frac{1}{I}\sum_{i=0}^{I-1} S_i^2(t)} \qquad (1)$$

where I is the number of MDCT coefficients (for MP3, I=576), and $S_i(t)$ is the MDCT coefficient at time index t for MDCT coefficient index i. Figure 1 illustrates that the RMS value carries approximately the same information with RMS Energy [5] obtained from PCM samples. The horizontal axis represents the granule index. For comparison purposes, the window size for corresponding PCM samples is 576.

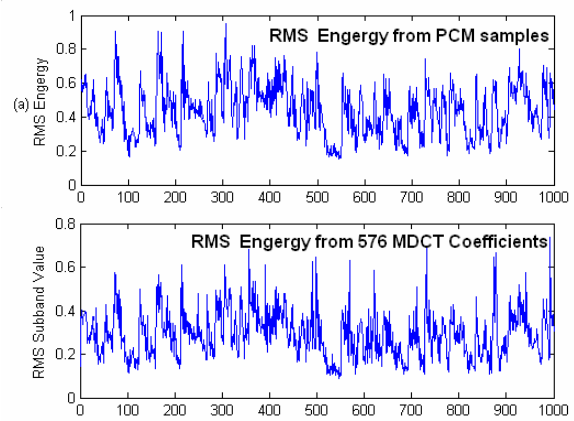

Figure 1 Amplitude Envelope for
(a) PCM samples and (b) MP3 Samples

2.2. Spectral Centroid

The spectral centroid is the balancing point of the MDCT spectral coefficient energy distribution. It is thus calculated as the first moment of the MDCT coefficient energy distribution [10]:

$$C(t) = \frac{\sum_{i=0}^{I-1}(i+1)S_i(t)}{\sum_{i=0}^{I-1} S_i(t)} \qquad (2)$$

where I is the number of MDCT coefficients, and $S_i(t)$ is the MDCT coefficient at time index t for MDCT coefficient index i.

The spectral centroid determines the frequency area around which most of signal energy is concentrated and is thus closely related to the time-domain zero crossing rates (ZCR) that is constantly used to characterize music content for music summarization. Figure 2 shows similar behavior for ZCR from PCM samples and Spectral Centroid from MP3 samples.

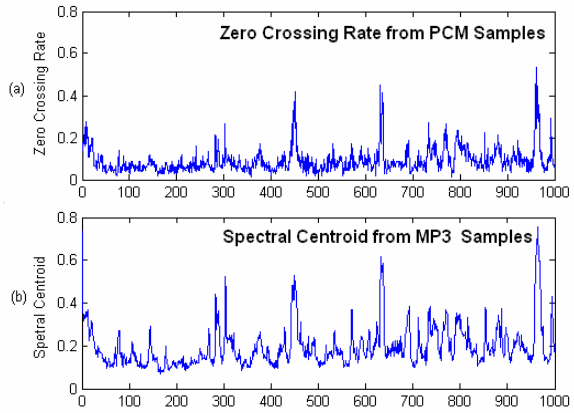

Figure 2 (a) Zero Crossing Rates VS (b) Spectral Centroid

2.3. Cepstral Features

The mel-frequency cepstrum has proven to be highly effective in recognizing structure and modeling the subjective pitch and frequency content of audio signals [2]. For PCM samples, the mel-cepstral features can be illustrated by the Mel-Frequency Cepstral Coefficients (MFCCs), which are computed from the FFT power coefficient. The power coefficients are filtered by triangular bandpass filter banks. These filter banks have a constant mel-frequency interval. But for MP3, the 32 subbands are subdivided into linearly spaced spectral coefficients. Therefore, in our experiment, we group the 576 MDCT coefficients of each granule into 6 newly defined subbands to approximate critical band widths (as Table 1 shows). Although we cannot achieve the same high frequency resolution as that computed using the PCM samples, they are good enough for our summarization purpose.

Table 1 Subbands defined to approximate critical band

Subband	Frequency interval (HZ)	Index of MDCT coefficient
1	0-459	0-11
2	460-918	12-23
3	919-1337	24-35
4	1338-3404	36-89
5	3405-7462	90-195
6	7463-22050	196-575

These MDCT coefficients are filtered by a triangular bandpass filterbank which is 50% overlapped and covers the frequency range of 0-22.05 KHz, and this results in 11 filterbanks covering the 6 new defined subbands.

Denoting the output of the k-th filterbank by O_k (k=1, 2… K, K=11), the MFCCs are calculated as:

$$c_n = \sqrt{\frac{2}{K}} \sum_{k=1}^{K} (\log O_k) \cos[n(k-0.5)\pi / K] \qquad n = 1,2,...L \quad (3)$$

where L is the length of the cepstrum. In our experiment, we set L=10.

The actual features used for the summarization are the mean and variance of amplitude Envelope and spectral centroid in a larger window (30 granules about 400ms for 44.1k Hz sampling rate). While for cepstral feature, we use the mean value of MDCT coefficients over 30 granules to calculate the MFCCs. The mean value of MDCT coefficients over a window is defined as:

$$\overline{S_i(t)} = \frac{1}{M} \sum_{m=0}^{M-1} \left| S_i(t+m) \right| \qquad i = 0,1,...575 \qquad (4)$$

where M is the window size, and here we set M=30.

3. AUTOMATIC SUMMARIZATION

Based on calculated features of each frame, we use clustering method to group the music frames to get the structure of the music content. The feature vector for each frame window can be denoted as:

$$V_i = (MAE_i, VAE_i, MSC_i, VSC_i, MFCC_i) \quad i = 1,2,...N$$

where MAE_i and VAE_i denote the mean and variance of amplitude Envelope, MSC_i and VSC_i denote the mean and variance of spectral centroid, and $MFCC_i$ denotes the 10 MFCCs coefficients. N is the total number of frame window.

The clustering method is similar to that used in [5] for music (PCM) summarization in uncompressed domain, but there is a difference in the framing scheme between PCM and MP3.

The time resolution for PCM and MP3 is different. For PCM samples, we can arbitrarily adjust the window size; but for MP3 samples, the resolution is a granule, which means we can only increase or decrease the window size by at least with one granule (corresponding to 576 PCM samples). To conceal the side effect, we segment PCM samples into fixed-length and overlapping windows to generate summary. But for MP3 samples, we group 15 MP3 frames (30 granules) as a bigger window frame which is non-overlapping.

The details of the clustering algorithm can be found in [5]. After clustering, the structure of the music content can be obtained. Each cluster contains frames with similar features. Summary can be generated in terms of this structure and music domain knowledge. According to music theory, the most distinctive or representative musical themes occurs repetitively in an entire music work. The scheme of summary generation can be found in [5].

4. EXPERIMENTS AND EVALUATION

The aim of music summarization is to extract the most common and salient themes of a given music. For this purpose, ideally, the summary lasting for long time should contain the summary lasting for short time. Therefore, we perform content study on original music and the summaries with different length.

Table 2 shows the content of our testing MP3 music "Top of the world" (by Carpenter). Music summaries are extracted using different length. The result is shown in Figure 3. The vertical axis represents the summary length, and the horizontal axis represents the frame number. The color bar in the figure corresponds to the frames extracted from the original MP3 music. The results show that the music summaries are located at the beginning of the first verse portion and the later part of two chorus portions. This excerpt was selected because most salient themes of music occurred commonly in the memorable verse theme and the later part of the chorus. From Figure 3, it can be seen that all the short summaries are included in the long summaries. This illustrates the fact that our proposed music summarization method can capture the main themes of the music work.

Table 2 The content of "Top of the world"

Section	Range(Frame Number)	Content
1	0-20	Introduction
2	21-176	Verse by the female singer
3	177-227	Chorus by two singers
4	228-248	bridge
5	249-450	Verse by the female singer
6	451-504	Chorus by two singers
7	505-513	Epilogue

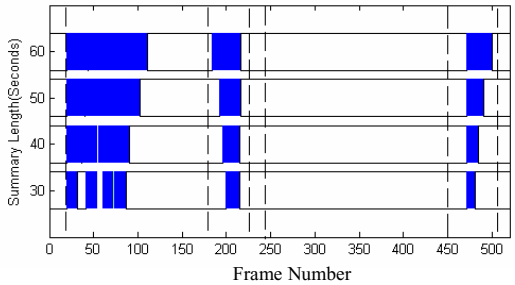

Figure 3 Experiment results on "Top of the world"

Since there is no ground truth available to evaluate the quality of a music summary, we employed a subjective study [11] to evaluate the performance of our music

summarization approach. There are various attributes that are considered in an ideal summary and the summary must be evaluated based on these attributes.

a. *Clarity*: This pertains to the clearness and comprehensibility of the music summary. A good music summary should capture the gist of the original music.
b. *Conciseness*: This pertains to the terseness of the music summary.
c. *Coherence*: This pertains to the consistency and natural drift of the segments in the music summary.

Four genres of music were used in the test. They are pop, classical, rock and jazz. Each genre contains five music samples. The aim of providing different music of different genres is to determine the effectiveness of the proposed method in creating summary of different genres. The length of music testing samples is from 2m52s to 3m33s. The length of the summary for each sample is 30s.
20 subjects with audio experience are invited. Before the tests, the subjects could listen to each testing sample as many times as needed till he/she grasped the theme of the sample. Then the subjects listened to summaries generated from test samples and rated the summaries in three categories (Clarity, and Conciseness and Coherence) on a scale of 1-5, corresponding to worst and best respectively. The average grade of summaries in each genre from all subjects is the final grade of this genre. In order to make comparison, we also asked subjects to rate summaries generated using our previously proposed method [5] based on corresponding PCM samples.

Table 3 Results of User Evaluation

Genre	Clarity		Conciseness		Coherence	
	I	II	I	II	I	II
Pop	4.6	4.3	4.0	4.2	4.5	4.6
Classic	4.0	4.1	3.6	3.8	3.8	3.5
Rock	4.6	4.2	4.5	4.3	4.3	4.1
Jazz	3.7	3.6	3.9	3.6	3.8	3.6

I: Music summarized on PCM samples
II: Music summarized on MP3 samples

From the evaluation results in Table 3, it can be seen that the summarization conducted on MP3 samples is comparable with the summarization conducted on PCM samples for all genres of music testing samples. We believe it is due to the features we selected in compressed domain carrying approximately same information as features in uncompressed domain. However, we still notice that the summarization based on MP3 is slightly inferior to summarization based on PCM samples in terms of evaluation results.

5. CONCLUSIONS AND FUTURE WORK

We have presented an automatic compressed domain music summarization method. The experimental result and the evaluation by a subjective study have shown that our proposed summarization approach is comparable to the summarization methods in uncompressed domain.
In the future, we need to explore more features such as pitch and timbre [10] in the compressed domain. We also need to improve and optimize the clustering and summary generation method. In addition, we will explore more aspects of music theory and apply it to music summarization.

6. REFERENCES

[1] R. Kraft, Q. Lu and S. Teng, Method and apparatus for music summarization and creation of audio summaries, US Patent 6,225,546, 2001.

[2] B. Logan and S. Chu, Music summarization using key phrases, In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Orlando, USA, 2000.

[3] M. Cooper and J. Foote, Automatic music summarization via similarity analysis, In *Proceedings of International Conference on Music Information Retrieval*, Paris, France, 2002.

[4] G. Peeters, A. Burthe and X. Rodet, Toward automatic music audio summary generation from signal analysis, In *Proceedings of International Conference on Music Information Retrieval*, Paris, France, 2002.

[5] Xu, C., Zhu, Y., and Tian, Q., Automatic music summarization based on temporal, spectral and cepstral features, In *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 117-120, Lausanne, Switzerland, 2002.

[6] G.Tzanetakis, P.Cook, Sound analysis using MPEG compressed audio, In *Proceedings of IEEE International Conf on Acoustics, Speech, and Signal Processing*, pp.761-764, Vol.2, Istanbul, Turkey, 2000.

[7] N.V.Patel and I.K.Sethi, Audio characterization for video indexing, In *Proceedings of SPIE Storage and Retrieval for Still Image and Video databases IV*, Vol.2670, pp.373-384, San Jose, CA, 1996.

[8] Ye Wang, Miikka Vilermo, A compressed domain beat detector using MP3 audio bit streams, In *Proceeding of ACM Multimedia*, pp194-202, Ottwa, Ontario, Canada, 2001.

[9] David. Pan, A Tutorial on MPEG/Audio Compression, *IEEE Multimedia*, Vol.2, No.2, 1995, pp.60-74.

[10] Silvia Pfeiffer, Thomas Vincent, Survey of compressed domain audio features and their expressiveness, In *Proceedings of SPIE-IS&T Electronic Imaging,* Vol.5021, pp133-147, Santa Clara, CA, 2003.

[11] John .P.Chin , Virginia A. Diehl and Kent L.Norman, Development of an instrument measuring user satisfaction of the human-computer interface, In *Proceedings of SIGCHI'88,* pp.213-218 , New York, 1988.