A Music Summarization Scheme using Tempo Tracking and Two Stage Clustering

Sangho Kim, Sungtak Kim, Suk-bong Kwon, and Hoirin Kim
School of Engineering
Information and Communications University
Daejon, the Republic of Korea
{ksh, stkim, sbkwon, hrkim}@icu.ac.kr

***Abstract*—In this paper, we present effective methods for music summarization which automatically extract a representative portion of the music by signal processing technology. Our proposed method uses 2-dimensional similarity matrix, tempo tracking, and clustering techniques to extract several segments which have different moods or dissimilar semantic structure in the music. The segments extracted are combined to generate a complete music summary. The three main techniques used in this paper are well-known and widely used for extracting music summary. However, we use them in a different way, and experiments show the proposed method captures the main theme of the music more effectively than conventional methods. The experimental results also show that one of the proposed methods could be used for real-time application since the processing time in generating music summary is much faster than other methods.**

Keywords—Music summarization; tempo tracking; clustering

Topic area—Multimedia databases

I. INTRODUCTION

Recently, digital music is moving into the mainstream of consumer life. Sales of single track downloads in the US in 2004 rose to 142.6 million from 19.2 million in the second half of 2003 [1]. The digital music market is rapidly growing. As such, there has been great importance placed on efficient management of numerous digital music databases. However, locating or browsing through thousands of tracks has a considerable data management problem [2]. Therefore, automatic music summarization is very helpful and important for music indexing, content-based music retrieval, and on-line music distribution [3]. Typical methods for music summarization use 2-dimensional (2D) similarity matrix [2], [4], [5], [6]. The method segments music signals into uniform length, extracts features from the frames, and finds the frame-to-frame similarity. Then, the matrix is used for pattern matching. If some part of the music is repeated after a time in the music, the distribution of similarity values of the latter part is similar to the previous one. So, we can find the best matching music phrase, and the phrase could be an optimal summary of the music. Some methods apply singular value decomposition to the similarity matrix to find repeated or substantially similar groups of segments [2]. Other methods compute a summary score by simply summing columns of the similarity matrix. Then, the most representative contiguous portions of the piece are extracted [4]. In 2000, Logan used a clustering technique and hidden Markov model (HMM) to extract the key phrases in the music [7]. The method extracts features from music signals and labels them. Then, it segments to analyze music structure and uses some heuristics to find the key phrase. These methods basically extract just a main theme or the most important and representative part of the music, making it difficult to capture various information of the music. To solve this problem, a few methods have been proposed to extract several parts of the music after analyzing the music structure [8], [9], [10]. Some of these use melody-based metrics to analyze the music structure from the similarity matrix. One of them uses both a k-means algorithm and HMM to analyze the music structure. Although the experimental results in some of the previous works have shown good performances, it seems necessary to devise a method that can reduce processing time and find more advanced music summary based on music psychology. In the aspect of music psychology, rhythm is the most fundamental factor in classifying the mood of music with other factors such as timbre and tonality [11]. If we use these characteristics when we generate a music summary that includes several parts of music, we can more easily extract dissimilar segments. But most of the methods use the rhythm of music just as a supporting factor in summarizing music. Therefore, if we track the rhythm or tempo of music along time at the first stage, the output of the summarization algorithm will be enhanced in the aspect of music psychology. In addition, to make the algorithm more robust, we use a different clustering method using variable threshold. Finally, we use an objective measure to evaluate the proposed methods. The results show that one of the proposed methods is relatively good in capturing the main theme of music and the other is applicable to real-time application.

II. FUNDAMENTALS

A. Feature extraction

We extract features from acoustic music signals. The process of feature extraction is illustrated in Fig. 1. Mel-Frequency Cepstral Coefficient (MFCC) is well-known for speech and audio signal processing. Other features such as spectral contrast and shape features [12], [13] are also used for music signal processing. The spectral contrast feature may be more suitable for music signal processing than MFCC and the

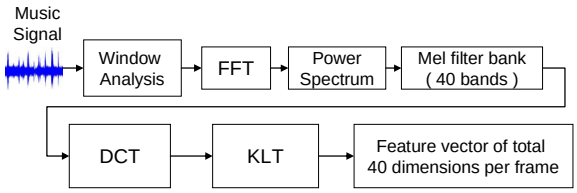

Fig. 1. Basic process of the feature extraction

octave scale filter bank is also frequently used. But it depends on the application. In our experiments, MFCC was good enough to analyze the similarity and dissimilarity between signals. And, linear predictive cepstral coefficient (LPCC) was better in discriminating delicate differences of timbre. However, our focus is not on the differences of timbre but on the explicit differences of music. Thus, MFCC is more reasonable. So, we use only MFCC features in this paper and the K-L transform is performed to map it onto an orthogonal space and remove correlation among dimensions of features.

B. 2D similarity matrix

2D similarity or self-similarity matrix is used since it effectively shows visual information for music summarization. To visualize the similarity between frames, a similarity measure S(i,j) is calculated for all combinations of frame indices i and j. Then an image is constructed so that each pixel at location i, j of the image is represented by a grayscale value proportional to the similarity measure, which is scaled so that the maximum similarity is given by the maximum brightness [5]. So, a 2D similarity matrix is obtained as follows. First, an input music signal is segmented with uniform length. Then, a feature vector is extracted at each frame. Finally, when V_i and V_j are feature vectors of i-th and j-th frames, frame-to-frame similarity $S_c(i,j)$ is computed using cosine distance measure as follows.

$$S_c(i,j) = \frac{\nu_i \bullet \nu_j}{\| \nu_i \| \cdot \| \nu_j \|} \quad (1)$$

The equation is the cosine distance measure which is the dot product of the feature vectors and normalized by its magnitudes to remove the dependence on magnitude [4]. Thus, this is exactly the cosine of the angle between two vectors. Another similarity metric is the Euclidean distance measure. It is defined as follows.

$$S_e(i,j) = \| \nu_i - \nu_j \|^2 \quad (2)$$

We used both the cosine distance measure and the Euclidean distance measure. The difference between the two methods was not considerable, but the Euclidean distance measure was better than the cosine distance measure when visualizing the similarity between feature vectors at the first stage. Fig. 2 shows an example of the matrix based on the Euclidean distance. We used an MIDI synthesizer to generate the same phrase of music played by three different instruments. The duration played by each instrument is 20 seconds. Thus, total

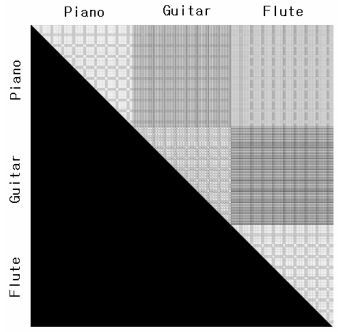

Fig. 2. 2D similarity matrix of music signals which have same phrase played by piano, electric guitars and flute, sequentially.

duration is 60 seconds. It is shown that the lower triangular of the matrix is black because the similarity matrix is symmetric so we did not process the lower triangular part. We can also find that the self similarity between phrases played by the same instrument is high and the cross similarity between phrases played by different instruments is relatively low.

C. Tempo tracking

We use Alonso's tempo estimation algorithm which shows very good performance [14]. The algorithm consists of three main modules which are onset detection, periodicity estimation, and beat location estimation. To track tempo in Beats Per Minute (BPM), we use a 20 second-analysis window and 1 second-step size. Thus, the transition of tempo along time can be plotted. But it is difficult to exactly track the tempo in BPM because there may be tempo doubling or the halvening problem like pitch doubling or halvening. The problem is caused by the rhythm pattern of music. We are not able to exactly track tempo if the interval between strong beats of music is twice (or half) of the interval of beats calculated by real BPM. However, what we want is not exact tempo in BPM value along time but rough transition information of tempo. So, we can use the structure of tempo transition along time although the tracked value is not correct. Then, the BPM value along time is quantized and a rough estimation of tempo is plotted as shown in Fig. 3.

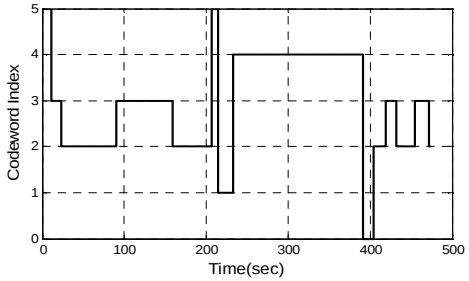

Fig. 3. Tempo transition along time ("Orion" by Metallica)

III. PROPOSED METHOD

The proposed method consists of two components which are pre-clustering and main clustering as shown in Fig. 4. We assume that music segments which have similar tempo could

be regarded as a same cluster at the first stage. So, in the pre-clustering stage, tempo tracking is performed along time and its BPM value is quantized with 10 BPM step size ranging from 60 to 220 BPM. Then, each section, which is longer than some threshold, is clustered after feature extraction. In our experiments, the threshold was set to 100 seconds. The section shorter than 100 seconds and longer than 8 seconds is regarded as an independent segment to be combined at the last stage. The sub-clustering procedure can be explained as follows. First, frames are segmented using novelty score [15]. Then, a mean of each segment is calculated. If the threshold is fixed, the segmentation technique will not give enough segments to merge segments, especially when the content change of the music signal is small. So the initial threshold is set to almost maximum similarity value and is decreased by feedback. Secondly, the segment similarity is calculated using the cosine distance measure as given in (1) and the Dynamic Time Warping (DTW) method is used to measure the distance or dissimilarity between segments. Previous research has shown that DTW could be used effectively for music signal processing [16]. Thus, the total similarity using the cosine distance measure and DTW can be obtained as

$$S_T(i,j) = (1-\lambda)\times S_C(i,j) + \lambda \times S_{DTW}(i,j) \qquad (3)$$

where $S_C(i,j)$ and $S_{DTW}(i,j)$ are similarities using cosine measure and DTW, respectively, and the lambda is a weighting factor set to 0.2 which was experimentally chosen in this work. Thirdly, similar segments are merged using a feedback scheme until the number of segments after the merging process reaches a predefined range, and each mean of the final segments is calculated. After that, the mean vector of feature vectors of each merged segment is used as an initial codeword of k-means algorithm. Then, k-means algorithm is applied to get final codewords. State transitions of frames along time are obtained by using the final codewords. Finally, several post-processing techniques such as deleting fluctuation of state transition, adding fade-in and fade-out effects to each segment, and combining each segment are performed. Deleting or filtering fluctuations of state transitions is very important because general pop music has various short-time sound effects as in cymbal sound or electronic sound which cause the feature vectors of adjacent frames to be very dissimilar even though the frames are actually included in the same cluster. The sub-clustering algorithm is summarized in Fig. 5. In Fig. 5, N_{seg} is the number of segments generated, Θ_N is a predefined number of segments, Θ_{S1} is a threshold of similarity at the segmentation, $\Theta_{lower,\ up}$ is a predefined number of segments, and Θ_{S2} is a threshold of similarity used in the merging process. Thus, we can get several sub-segments on each segment obtained from the pre-clustering. The sub-segments which are longer than 8 seconds are used for one of the final segments. The length of each segment is restricted to 16 seconds. In summary, we segment music by tempo transition and cluster each segment again using features extracted. Thus, we use tempo transition exhaustively for finding more dissimilar and representative parts of music.

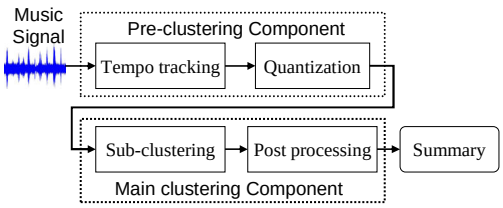

Fig. 4. Architecture of the proposed algorithm

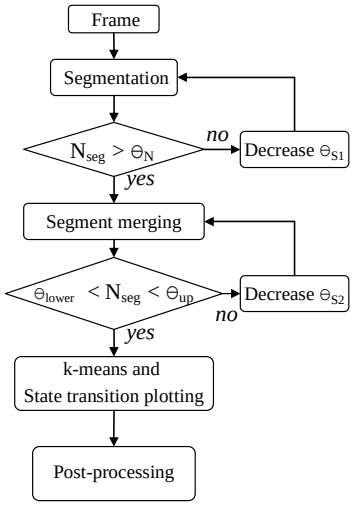

Fig. 5. Flowchart of main clustering

IV. Experimental Results and Discussion

In order to evaluate the methods, four criterions are used. The first is how well the methods grasp the main theme of music. It is related to the accuracy in Table 1. The accuracy is an average of the percentile representation of the ratio between the number of summaries which contain the chorus of the original song, and the number of total songs. The second is how much the method compresses the original music. It is related to the compression ratio in Table 1. The compression ratio is an average of the percentile representation of the ratio between the length of the summarized song and the length of the original song. The third is how much the final music summary contains dissimilar segments of original music. It is related to the total segments and the total NSS in Table 1. The total segments are the total number of segments the method generated automatically, and the total NSS is the total summation of the NSS which is the number of similar segments within a summary automatically generated. The last is how fast the method extracts the summary. It is related to the processing time. It is shown in Fig. 6. We used 10 songs (Avril Lavigne, Michael Jackson, etc) for evaluation. The test songs were manually annotated to evaluate the accuracy and the NSS. The annotation process was conducted by an author who has knowledge of the test songs and music psychology. So it was not widely important to include subjective tests for performance comparison. And, all the songs are sampled at 16 kHz with 16bits per sample and mono format. The comparison results are shown in Table 1 and Fig. 6. CL using TT stands for the proposed method using tempo tracking, and CL refers

TABLE I.THE RESULTS OF PERFORMANCE COMPARISON

Measure \ Method	HMM	CL	CL using TT
Accuracy (%)	50	20	90
Compression ratio (%)	13.16	14.75	17.13
Total segments	25	22	29
Total NSS	2	2	4

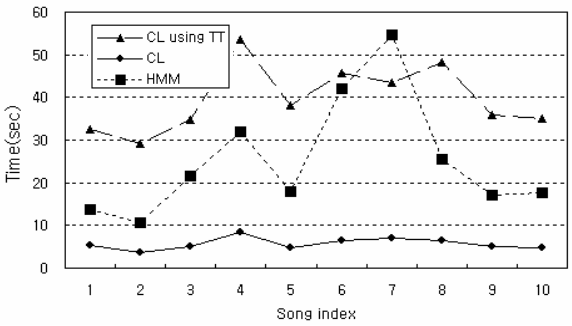

Fig. 6. Processing time comparison among proposed methods with(CL using TT) or without(CL) tempo tracking, and Peeters method using HMM

to the proposed method using only clustering technique. And, the Peeters method is denoted by HMM [8]. The results show that CL using TT is better at capturing the main theme of the music over other methods. This is because the highlight of music, a hook or a chorus, has a different or special rhythm pattern which is directly related to the tempo extracted. So, tempo tracking is beneficial to catch the hook of the music as well as dissimilar parts. This method among three methods also includes the largest number of segments. In addition, there was no big difference in NSS and compression ratio. Thus, we can think the proposed method using tempo tracking is the best for off-line processing. However, much time was needed to track the tempo along time. Thus, the method is not proper for use in real-time applications which aim to provide customized summaries based on user query. The proposed method using only clustering is the fastest among three methods. So, the method may be one of the choices for real-time applications although its performance is not the best. However, the performance of the method can be controlled by adjusting thresholds such as the predefined number of segments at the segmentation and the merging processes. If more segments are extracted, the final summary could include more various information of music although the compactness of the summary deteriorates as the length of the summary increases.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed new methods for automatic music summarization which attempt to find several segments within a single music piece based on music psychology. In general, almost all the existing summarization methods use the tempo estimation technique just for finding exact boundaries or aligning musical phrases. However, we used the tempo tracking more exhaustively for finding more meaningful summaries. The experimental results show that the proposed methods have good performance. The proposed method using tempo tracking could catch the main theme of music very well as well as dissimilar parts within a single music piece. In the future, we will use more various characteristics related to the rhythm pattern. We can also consider pitch transitions, chord progressions and tonality of music in later works. It is also needed to use more valuable knowledge on music psychology. In addition, we need to test more various songs for performance measure. We also hope that this approach, which uses a kind of rhythm pattern in an effort to utilize knowledge on music psychology, inspires several researches related to music signal processing.

REFERENCES

[1] International Federation of the Phonographic Industry (IFPI), *2005 Digital Music Report*

[2] Matthew Cooper and Jonathan Foote, "Summarizing popular music via structural similarity analysis," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October, 2003

[3] Xi Shao, Naumunu C. Maddage, Changsheng Xu and Mohan S Kankanhalli, "Automatic Music Summarization Based on Music Structure Analysis," In *IEEE International Conf on Acoustics, Speech and Signal Processing (ICASSP05)*, Philadelphia, USA, 2005.

[4] Matthew Cooper and Jonathan Foote, "Automatic Music Summarization via Similarity Analysis," *Proc. IRCAM*, pp. 81-85, Oct. 2002

[5] Jonathan Foote, "Visualizing Music and Audio using Self-Similarity," *Proc. ACM Multimedia Conference*, pp. 77~80, Orlando, Florida, November 1999.

[6] J. Foote, "Automatic Audio Segmentation using a Measure of Audio Novelty," In *Proceedings of IEEE International Conference on Multimedia and Expo*, vol. I, pp. 452-455.

[7] Logan, B., and Chu, S. "Music summarization using key phrases," In *Proc. IEEE ICASSP*. 2000.

[8] Geoffroy Peeters, Amaury La Burthe and Xavier Rodet, "Toward Automatic Music Audio Summary Generation from Signal Analysis," *Proc. ISMIR*, Paris, 2002

[9] N. Maddage, X. Changsheng, M. Kankanhalli, and X. Shao. "Content-based music structure analysis with applications to music semantics understanding," In *6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, October 2004.

[10] Lie Lu, Muyuan Wang and Hong-Jiang Zhang, "Repeating Pattern Discovery and Structure Analysis from Acoustic Music Data," *Workshop on Multimedia Information Retrieval 2004*, in conjuction with ACM Multimedia 2004, Oct 15-16,2004, New York, NY,2004.

[11] Rudolf E. Radocy and J. David Boyle, *Psychological foundations of musical behavior*. Charles C. Thomas Publisher, Ltd.

[12] Dan-Ning. Jiang, Lie. Lu, Hong-Jiang Zhang, Jian-Hua Tao and Lian-Hong Cai. "Music type classification by spectral contrast feature," In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Lausanne (Switzerland), August 2002

[13] Lie Lu, Dan Liu and Hong-Jiang Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE transactions on audio, speech, and language processing*, Vol. 14, No.1, January 2006.

[14] Miguel Alonso, Bertrand David and Gael Richard, "Tempo and beat estimation of musical signals," In *Proceedings of ISMIR 2004*, Barcelona, Spain, October 2004

[15] Lie Lu, Liu Wenyin and Hong-Jiang Zhang, "Audio textures: Theory and Applications," *IEEE transactions on speech and audio processing*, Vol. 12, No.2, March 2004

[16] C. Yang, "MACS: Music audio characteristic sequence indexing for similarity retrieval," In *Proc. Workshop Applications of Signal Processing to Audio and Acoustics*, 2001.