THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.27, No.2E 2008. 6. pp. 51~56

Automatic Music Summarization Using Vector Quantization and Segment Similarity

Sangho Kim*, Sungtak Kim*, Hoirin Kim*
*The School of Engineering at Information and Communications University, Korea
(Received March 31, 2008; accepted May 29, 2008)

Abstract

In this paper, we propose an effective method for music summarization which automatically extracts a representative part of the music by using signal processing technology. Proposed method uses a vector quantization technique to extract several segments which can be regarded as the most important contents in the music. In general, there is a repetitive pattern in music, and human usually recognizes the most important or catchy tune from the repetitive pattern. Thus the repetition which is extracted using segment similarity is considered to express a music summary. The segments extracted are again combined to generate a complete music summary. Experiments show the proposed method captures the main theme of the music more effectively than conventional methods. The experimental results also show that the proposed method could be used for real-time application since the processing time in generating music summary is much faster than other methods.

Keywords: *Music summarization, Repetitive pattern, Segment similarity, Vector quantization*

I. Introduction

Recently, digital music is moving into the mainstream of consumer life. Sales of single track download in the US in 2004 rose to 142.6 millions from 19.2 millions in the second half of 2003 [1]. As the digital music market rapidly grows, there has been a 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 methods segment music signals into uniform length, extract features from frames, and find the frame-to-frame similarity. Then the similarity 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 similarto the previous one. So we can find the best matching music phrase, and the phrase could be a good summary of the music. Some methods apply singular value decomposition to the similarity matrix to find similar or substantially repetitive groups of segments [2]. Other methods compute a summary score by simply summing columns of the similarity matrix. Then the most representative contiguous pieces of the part 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. On the other hand, a few methods have been

Corresponding author: Hoirin Kim (hrkim@icu.ac.kr)
School of Engineering at Information and Communications University (ICU), 119 Mungiro, Yuseong-gu, Daejeon, 305-732, Korea

proposed to extract several parts of the music after analyzing the music structure [8], [9], [10]. Some of those 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 effective music summary for more general patterns such as pop music. The repetition of musical phrase is one of the important factors to recognize the catchy tune which is the most important and representative part of pop music because listeners easily can remember this repeated part and recognize original music when listening this. We can know that the chorus is repeated, and that the repeated chorus could be good for music summary. If we can find the exact repetition of a phrase with high similarity between them, the phrase could be regarded as one of candidates for music summary. Here the number of times that segments are repeated is not considered in this scheme, because there could be some possibility that meaningless segments are frequently repeated. Instead we consider only the degree of similarity between two segments. Tocalculate the segment similarity, distance between codeword indices after vector quantization (VQ) is used. As a result, we could greatly reduce processing time. For evaluation, we use objective and subjective measures. The results show that the proposed method is effective in capturing the main theme of music and applicable to real-time application because the processing time is very fast.

The rest of this paper is organized as follows. In Section II, the feature extraction is explained using this paper. The proposed music summarization technique by using vector quantization and segment similarity is presented in Section III. Finally, in Section IV and V, the experimental results of the proposed methods and our conclusion are given.

II. Feature Extraction

In this scheme, chromagram, also called the pitch class profile (PCP) feature, is used because human tends to recognize the sameness of music segments by melody metric. Chromagram represents energy pattern of musical notes. The chromagram combines the frequency components in short-time Fourier transform (STFT) belonging to the same pitch class and results in a 12-dimensional representation, corresponding to C, C#, D, D#, E, F, F#, G, G#, A, A#, and B in music, respectively. For the representation, let $X_{STFT}[K,n]$ denote the magnitude spectrogram of signal $x[n]$, where $0 \le K \le N_{FFT}/2$, K is the frequency index, and N_{FFT} is the FFT length [11]. The chromagram of $x[n]$ can be defined as

$$X_{PCP}[\tilde{K},n] = \sum_{K:P(K)=\tilde{K}} X_{STFT}[K,n] \tag{1}$$

The warping between the frequency index K in STFT and the index K in PCP is

$$P(K) = [D \cdot \log_2(\frac{K}{N_{FFT}} \bullet \frac{f_s}{f_1})] \bmod D \tag{2}$$

where f_s is the sampling rate and f_1 is the reference frequency. The reference frequency can be set to the C3 note. In addition, dimensions of feature vector can be varied by varying the numerical value, D in (2). As a result, we can get D-dimensional feature vector in which each elements present the energies of pitch classes. In Experiments of this paper, D is set to 12.

III. Proposed Method

When some conventional summarization methods for extracting representative segmentsof music focus on the repetitiveness ofmusical phrases, they consider

on the re-occurrence frequencies of the segments as well. In addition, they usually need much processing time. And, the approaches donot consider the human perception process of recognizing important parts of music. In general, popular music has a typicalpattern as shown in Fig. 1.

In the beginning, there is an intro section. The section usually does not have any vocal sound. The general aim of the section is to increase listener's interest. After the section, there could be a vocal section. The section, called verse,has relatively smooth mood or melody. And then, the chorus section comes out. The section is usually regarded as the most important, catchy,and representative part of music. However, average people compared to an expert need to listen to the music more repetitively to recognize or analyze the structure of music. But the process of recognizing the structure is not very different. Almost people can recognize similar melody, phrase, lyrics, and mood of music because theycan remember the repetition of phrases although the melody or lyrics isslightly varied at the later occurrences. Thus, it is not difficult to understand or analyze the structure of music although individual's musical ability is different. Thus, considering similarity between two segments will be very useful when we find the structure of music. The most important parts of the structure are the verse and chorus sections, and the verse and chorus are several times repeated with slight modification within a piece of music. Thus, the property of the structure is the most important factor in the segment similarity method for extracting music summary. The algorithm of the segment similarity method is shown in the Fig. 2.

Firstly, frame analysis is performed. The frame size is 1 sec with no overlapping. Secondly, feature vector of each frame is extracted. Thirdly, LBG clustering is conducted. Of course, other VQ methods likek-means can be used. The codebook size was set to 128 in this work. After training, the codebook is used for encoding the whole frames of music. Thus, each frame has the corresponding codeword index. Then calculation of the segment similarity is performed. The segment similarity (SS) is calculated by using the equations defined as

Fig. 1. General pattern of pop music.

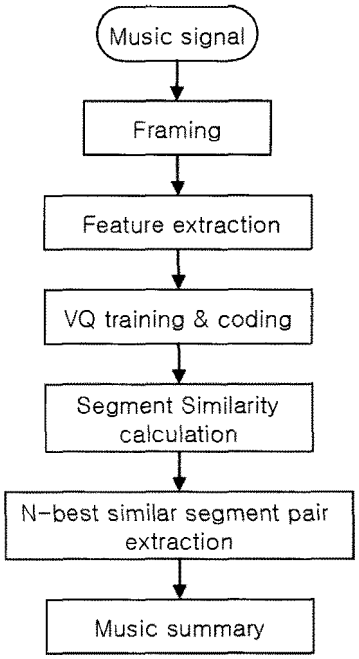

Fig. 2. Flow chart of the algorithm.

$$S_{frame}(x,y)=\begin{cases}1, & if\ C(x)=C(y)\\ 0, & otherwise\end{cases} \tag{3}$$

$$SS_{x,y}(k_1,k_2)=\frac{1}{k_2-k_1}\sum_{t=0}^{k_2-k_1-1}S_{frame}(x+t,k_1+t) \tag{4}$$

$$(k_1^*,k_2^*)=ArgMax\ SS_{x,y}(k_1,k_2) \tag{5}$$

where $0\le x\le N-L$, $y=x+L$, $y\le k_1$, $k_2=k_1+L$, and $L_{\min}\le L\le L_{\max}$. N is the number of total frames in the music. L is the number of frames in a segment, $L_{\min}$ is the number of frames of predefined minimum segment length, $L_{\max}$ is the number of frames of predefined maximum segment length, S_{frame} is a similarity measurement between frames, $SS_{x,y}(k_1,k_2)$ is the segment similarity between a segment from x-th frame to y-th frame and a segment from k_1-th

frame to k_2-th frame, and $C(n)$ is a codeword index of the n-th frame. Thus equation (3) finds the sameness between frames. The function returns 1 if the codeword indices of two frames are same; otherwise it returns 0. Equation (4) calculatesthe similarity between two segments which are separated apart. A segment is composed of several frames of music. So the similarity between segments can be evaluated by considering the sameness of whole music. Then one segment pair which has the largest similarity value is found presented in equation (5). And the process of calculating the segment similarity is repeated until it reaches a predefined number. Here the pairs found previously are excluded in the next process. That is, the range of the segment pair which has the largest similarity value in the present process should not overlap with the ranges of pairs found previously. It is possible to use some techniques to select just one segment among the segments searched. It depends on user preference or user query. If we want to construct a system which has functionality that user could select type of music summary such as long or short version, we need to add other algorithm to select one segment among several segments in the segment similarity scheme. One possible choice is to use the energy of each segment.

IV. Experimental Result and Discussion

In order to evaluate the method, four criterions which were defined in our previous paper [12] are used again. The first is how well the method grasps the chorus of music. It is related to the accuracy in Table 1. The meaning of the accuracy of table 1 is the average probabilityof catching chorus successfully for total 10 songs. If automatically extracted summary includes any one of hand-made choruses of original music, we regard this automatically extracted summary as succeed.The second is how much the method compresses the original music. It is related to the compression ratio in the table. 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 one. 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. 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 generated automatically. The last is how fast the method extracts the summary. It is related to the processing time. The comparison results are shown in Table 1 and Fig. 3.

The Peeters' method [8] is denoted by HMM. In this Peeter's method, a HMM is estimated by using the Bau-Welch algorithmfor given music, and then music summary is extracted using state sequences which are outputs of Viterbi decoding. Two-stage stands for the method developed previously [12]. In the first state of this two-stage method, frames are classified by k-means algorithm after computing the BPM (Bits Per Minute). The classes of frames are grouped into several groups, and then music su-

Table 1. The results of performance comparison among the Peeter's method using HMM (HMM), two-stage clustering (TWO-STAGE), and the segment similarity method (SS)

Symbol	HMM	Two-stage	SS
Accuracy (%)	50	90	80
Compression ratio (%)	13.16	17.13	17.49
Total segments	25	29	30
Total NSS	2	4	6

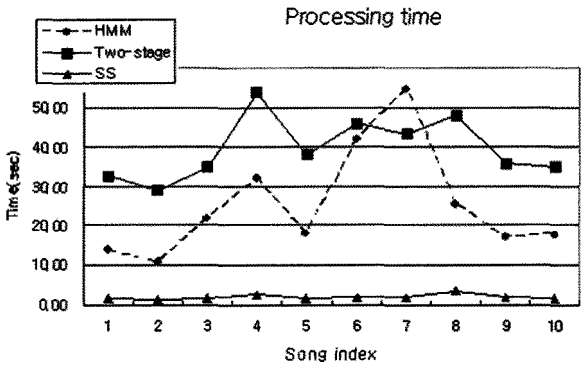

Fig. 3. Processing time of the summarization methods.

mmary is extracted by using frames included in these groups. SS stands for the segment similarity method. We used 10 songs (Avril Lavigne, Michael Jackson, etc) for evaluation. The test songs were manually annotated to evaluate the accuracy and the NSS. All the songs are sampled at 16 kHz with 16 bits per sample and mono format. The results show that two-stage is better in capturing the main theme of the music than other methods. The segment similarity method catches the hook of the music well although the accuracy is less than the two-stage method. This method among three methods also includes the largest number of segments, which is bounded to a fixed number in the scheme. However, the compression ratio is not much bigger than other methods. So we can say that the performance of the segment similarity method for catching hooks and generating several segments is reasonable. But the method is not good in the aspect of NSS factor. It means that the numberof similar segments within a summary is greater than other methods because similar pairs could be found at the second or third process if a segment is frequently occurred more than 4 times with slight variation. To prevent this situation, it is possible to delete duplicate segments within a summary. But it was not considered in this scheme because it is difficult to set an exact threshold to know duplicate segments. We could use melody metrics or devise other approaches in later works. In the aspect of processing time, however, the segment similarity method is much faster than other methods. To find repetitive sections, we just compare the identity of codeword indices after VQ. Thus the proposed method could be the best choice for a real-time application which requires fast processing time and comparable performance.

V. Conclusion

In this paper a method for automatic music summarization, which attemptsto find several segments within a single music piece, was proposed. The experimental results show that the proposed method has good performance. The method uses VQ, PCP feature, and equationsfor calculation of the segment similarity value. The experimental results show that the method was very good in the aspect of processing time and the accuracy was also good. Nowadays, music indexing, retrieval, and browsing technologies are becoming more and more important. In addition, users' taste could be different. Some users like a short music summary which includes just a chorus part of music, and some users like a long music summary which includes various segments. However, there will be storage problem if we generate whole different types of music summaries in advance. Thus it will be very helpful if a summarization method has very fast processing time. By using themethod, we do not need to generate music summaries in advance. Service provider could generate a music summary based on user query in real -time. In this aspect, the proposed method could have a merit. In the next step, we will consider melody metrics to find duplicate segments within a summary. In addition, we need to test various codeword sizes and clustering algorithms. The results of this paper were presented through the average performances of four evaluation criteria, but if these performances are given according to genres, the performance of the proposed method is more valuable for researchers in the music summarization.

References

1. *International Federation of the Phonographic Industry* (IFPI), 2005 Digital Music Report.
2. M. Cooper and J. 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. X. Shao, N. C. Maddage, C. Xu and M. 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. M. Cooper and J. Foote, "Automatic Music Summarization via

Similarity Analysis," Proc. IRCAM, 81-85, Oct. 2002.
5. J. Foote, "Visualizing Music and Audio using Self- Similarity," Proc. ACM Multimedia Conference, 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, I, 452-455.
7. B. Logan, and S. Chu, "Music summarization using key phrases," In Proc. IEEE ICASSP, 2000.
8. G. Peeters, A. L. Burthe and X. 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. L. Lu, M. Wang and H. J. Zhang, "Repeating Pattern Discovery and Structure Analysis from Acoustic Music Data," Workshop on Multimedia InformationRetrieval 2004, in conjuction with ACM Multimedia 2004, Oct 15-16,2004, New York, NY,2004.
11. W. Chai, "Semantic segmentation and summarization of music," IEEE signal processing magazine, March 2006.
12. Sangho Kim, Sungtak Kim, Suk-bong Kwon, and Hoirin Kim, "A Music summarization scheme using tempo tracking and two stage clustering," IEEE workshop on Multimedia Signal Processing 2006, Vol. 1, pp 225-228, Victoria, Canada, October 4, 2006.

[Profile]

• Sangho Kim

Sangho Kim received the B.S. degree in electronics engineering from the Sejong University and the M.S. degree in multimedia communications and processing from the Information and Communications University (ICU), Korea in 2002 and 2007, respectively. His research interests are music information retrieval and automatic music summarization.

• Sungtak Kim

Sungtak Kim received the B.S. degree in electronics engineering from the Ulsan University and the M.S. degree in multimedia communications and processing from the Information and Communications University (ICU), Korea in 2000 and 2003, respectively. He is currently pursuing the Ph. D degree in multimedia communications and processing at the Information and Communications University. His research interests are automatic music summarization, robust speech recognition and speaker recognition.

• Hoirin Kim

Hoirin was born in Seoul, Korea in 1961. He received the M.S. and Ph.D. degrees from the Dept. of Electrical and Electronics Engineering, KAIST, Korea, in 1987 and 1992, respectively. From October 1987 to December 1999, he has been a Senior Researcher in the Spoken Language Processing Lab. at the Electronics and Telecommunications Research Institute (ETRI). From June 1994 to May 1995, he was on leave to the ATR-ITL, Kyoto, Japan. From July 2006 to July 2007, he was in the Institute of Neural Computation, UCSD, USA as a visiting researcher. Since January 2000, he is an Associative Professor at Information and Communications University (ICU), Korea. His research interests are signal processing for speech & speaker recognition, audio indexing & retrieval, and spoken language processing.