THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.26, NO.2E 2007, 6 pp. 39-43

Automatic Music Summarization Using Similarity Measure Based on Multi-Level Vector Quantization

Sungtak Kim*, Sangho Kim* Hoirin Kim*
*The School of Engineering at Information and Communications University
(Received March 6 2007; Revised May 16 2007; Accepted Jun 16 2007)

Abstract

Music summarization refers to a technique which automatically extracts the most important and representative segments in music content. In this paper, we propose and evaluate a technique which provides the repeated part in music content as music summary. For extracting a repeated segment in music content, the proposed algorithm uses the weighted sum of similarity measures based on multi-level vector quantization for fixed-length summary or optimal-length summary. For similarity measures, count-based similarity measure and distance-based similarity measure are proposed. The number of the same codeword and the Mahalanobis distance of features which have same codeword at the same position in segments are used for count-based and distance-based similarity measure, respectively. Fixed-length music summary is evaluated by measuring the overlapping ratio between hand-made repeated parts and automatically generated ones. Optimal-length music summary is evaluated by calculating how much automatically generated music summary includes repeated parts of the music content. From experiments we observed that optimal-length summary could capture the repeated parts in music content more effectively in terms of summary length than fixed-length summary.

Keywords: *Music summarization, Similarity measure, Multi-level vector quantization, Fixed-length music summary, Optimal-length music summary, Mahalanobis distance*

I. Introduction

Recently the rapid growth of the Internet and personal storage for multimedia data has increased the need for comfortable consumption of multimedia contents. This need has resulted in greater demand for technologies that analyze and characterize multimedia data. For example, movie trailers, book reviews, and paper abstracts each play a role in providing key information on their topics. So far a number of techniques have been proposed and developed to automatically generate text or video summaries [1-3]. Music summarization provides the key content of music like text or video summaries. Basically there are two ways of music summarization discussed in former researches. The first is to provide a repeated part as music summary and the second is to provide a concatenated segment which consists of parts which have different characteristics in music content. For generating better music summary, several considerations such as feature extraction, music structure representation, and the way of summary extraction should be addressed. Of these factors, our approach focuses on music structure representation and the way of summary extraction. For representing music structure, several methods [4-12] have been proposed, but these methods represent music structure using only one fixed-resolution. In the case of using only one fixed-resolution, the presented music structure provides only limited information on music structure. In this paper, to overcome this problem, multi-level vector quantization (VQ) to represent music

Corresponding author: Sungtak Kim (stkim@icu.ac.kr)
Speech Recognition Technology Lab. in the School of Engineering at Information and Communications University, 119 Mungiro, Yuseong-gu, Daejeon, 305-732, Korea

structure with various resolutions is used. Here, multi-level VQ means that a feature vector is represented by codeword indices of various sizes of codebooks from deferent level VQs. From multi-level VQ, we can obtain more information on music structure than previous researched methods. And for similarity measure, the weighted sum of counts of same codeword indices or the weighted sum of the reciprocal of the Mahalanobis distances of same codeword indices at the same position in segments are proposed for summary extraction. In addition, we propose new techniques which provide a repeated part of music data with fixed-length or optimal-length. Optimal-length means that a proper length of music summary with which a given music content can be effectively represented.

In Section II, feature vector and the proposed method for music summarization are described. In Section III, the experimental results of the proposed methods are presented. Finally, in Section IV, we summarize and conclude the proposed methods.

II. Automatic Music Summarization

Fig. 1 conceptually illustrates a music summarization method which provides the repeated part of music content as summary. For instance, in Fig. 1, we can know that the chorus is repeated, and that the repeated chorus could be good for music summary. Our purpose is to extract the repeated parts like the chorus, and to provide these parts as music summary.

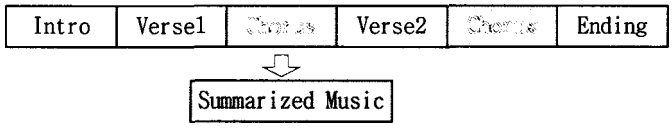

Fig. 1. Example of Music Summary.

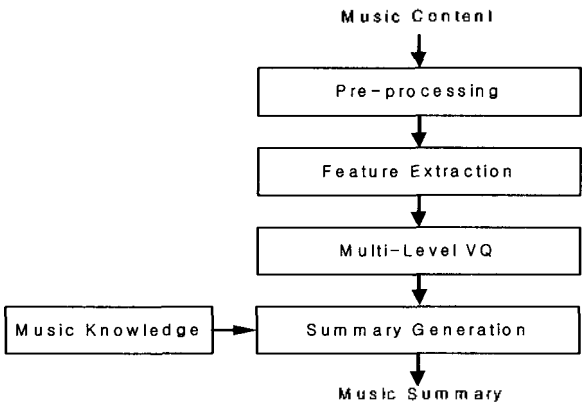

Fig. 2. Block Diagram of Music Summarization Using Multi-Level Vector Quantization.

Fig. 2 shows a block diagram of the proposed music summarization system.

2.1. Pre-Processing

Before extracting the feature vectors for music content, the music signal is segmented into fixed length and 50% overlapping frames, and the silence frame is removed by comparing its frame energy with the pre-defined threshold.

2.2. Feature Extraction

In this paper, we use MFCC (Mel-Frequency Cepstral Coefficient) as the feature vector characterizing the music content. The MFCC is the most typical feature in speech recognition [13]. MFCC is calculated as

$$c_n = \sqrt{\frac{2}{K}} \sum_{k=0}^{K} (\log S_k) \cos[n(k-0.5)\pi / K] \quad (1)$$
$$n = 1,2,\ldots,L$$

where S_k is the output of K th filter in the filter bank. 80 and 25 are used for the values of K and L, respectively.

2.3. Summary Generation

According to music knowledge, the most distinctive and representative musical theme should repetitively occur throughout an entire piece of music. In this paper, two kinds of music summaries are generated according to the length of music summary which has fixed-length or optimal-length.

A. Fixed-Length Summary Generation

We use multi-level vector quantization to present music structure. For finding the repeated part in music content, we propose two kinds of similarity measures: CSM (Count-based Similarity Measure) and DSM (Distance-based Similarity Measure). The fixed-length music summary can be generated as follows.

Find a frame, $f_{\max}$, which has the maximum $VQSM_i$ (VQ-based Similarity Measure) among all frames

$(f_1, f_2, \ldots, f_N)$ in the music content given, where N is the number of frames in the music.

$$f_{\max} = \arg\max_i VQSM_i \tag{2}$$

$$VQSM_i = \max_j \sum_{M=M_1,M_2,\ldots,M_k} w_{M_k}\left(\sum_{s=1}^{S} \cdot I_{M_k}(f_i+s, f_j+s)\right) \tag{3}$$

$$\text{CSM: } I_{M_k}(f_i, f_j) = \begin{cases} 1, & \text{if } C_{M_k}(f_i) = C_{M_k}(f_j) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\text{DSM: } I_{M_k}(f_i, f_j) = \begin{cases} (MD(f_i, f_j))^{-1}, & \text{if } C_{M_k}(f_i) = C_{M_k}(f_j) \\ 0 & \text{,otherwise} \end{cases} \tag{5}$$

where $i, j \in [1, N]$, $i + S < j$

In equations (2) – (5), M is quantization level and $C_{M_k}(f_i)$ denotes a codeword of frame f_i of M_k quantization level. w_{M_k} is a weighting factor used in the corresponding quantization level, S is the number of frames corresponding to pre-defined summary length, and $MD(f_i, f_j)$ is the Mahalanobis distance between feature vectors of frames f_i and f_j. After finding frame $f_{\max}$, which has maximum value of $VQSM_i$, $[f_{\max}, f_{\max}+S]$ is provided as a summary. The weighting factor w_{M_k} is basically proportional to the reciprocal of the average MSE (Mean Square Error) value of the quantization level M_k, and we smoothed weighting factors as in equation (6).

$$w_{M_k} = \frac{1}{MSE_{M_k}} + \alpha \qquad \alpha = \max_{M=M_k,\ldots M_k} \frac{1}{MSE_{M_k}} \tag{6}$$

B. Optimal-Length Summary Generation

In the previous section, the summary length has to be predefined before generating summary. However, it is not effective that all summaries for different music have the same length.

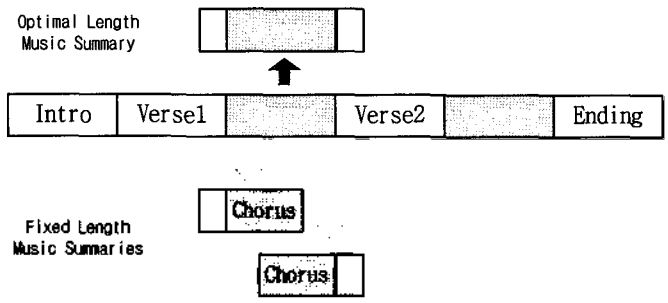

Fig. 3. Comparison of Fixed-Length and Optimal-Length Music Summary.

As shown in Fig.3, fixed-length summary could not contain the whole chorus part in some cases. However, even though optimal-length summary provide longer summary than fixed-length summary in some cases, it can provide whole chorus to listeners as much as possible. In optimal-length music summary, the length of summary can be changed by using equations (7)-(9) instead of equations (2) and (3).

$$f_{\max} = \arg\max_i VQSM_i^* \tag{7}$$

$$VQSM_i^* = \arg\max_{S_{MIN} \le S^* \le S_{MAX}} VQSM_i^{S^*} \tag{8}$$

$$VQSM_i^{S^*} = \max_j \sum_{M=M_1,M_2,\ldots,M_k} w_{M_k}\left(\sum_{s=1}^{S^*} \cdot I_{M_k}(f_i+s, f_j+s)\right) \tag{9}$$

In generating the optimal-length music summary, while the summary length is changed from S_{MIN} to S_{MAX}, find $VQSM_i^*$ and provide $[f_{\max}, f_{\max}+S^*]$ as optimal-length music summary. S^* is the number of frames corresponding to the optimal music summary length.

III. Evaluation

To evaluate the proposed methods, we use 100 Beatles songs published until now. For the feature extraction, audio signals sampled at 22.05kHz are divided into frames by 200ms window at 100ms frame rate.

3.1. Evaluation of fixed-length music summary

To evaluate the fixed-length music summary, we use Overlapping Ratio which is the degree of overlapping hand-made repeated parts and automatically generated summary. In this paper, we focus on the music summary technique which provides a repeated part as summary of music content. Therefore, Overlapping Ratio can be a reasonable evaluation for provided music summary in the case of providing a repeated part as music summary.

$$\text{Overlapping Ratio} = \frac{Length(S_{hand} \cap S_{VQ})}{Length(S_{hand})} \times 100 \tag{10}$$

where S_{hand} and S_{VQ} are segments of hand-made music

Table 1. Performance of Single-Level Vector Quantization.

Quantization Level(M)	Overlapping Ratio (OR)	
	CSM	DSM
8	61.3	64.9
16	63.7	65.2
32	59.9	64.6
64	64.4	63.9
128	64.4	64.6
256	62.7	61.9
512	61.4	62.6

Table 2. Performance of Multi-Level Vector Quantization.

Multi-level VQ (M)	Overlapping Ratio (%)	
	CSM	DSM
8,16	63.3	66.8
8,16,32	61.1	65.9
8,16,32,64	64.5	66.5
8,16,32,64,128	64.8	66.2
8,16,32,64,128,256	65.1	63.9
8,16,32,64,128,256,512	66.0	63.5

Table 3. Comparison of Former methods and The Proposed One.

	Overlapping Ratio (%)
Beth Logan' s method [4]	47.2
Chanshen Xu' s method [5]	47.5
Proposed method	66.8

summary and automatically generated music summary. Hand-made music summary is manually extracted from repeated parts by observing the musical score of music content. In this paper, the length of automatically generated summary is predefined to 20 seconds. We compare former researches [4],[5] with the proposed method. Table 1 shows the results in the case that only single quantization level was used. The results show that the case of DSM is better than CSM.

From Table 1 and 2, the performance of using multi-level VQ is better than that of using single level VQ, and the distance-based similarity measure give better performance than the count-based similarity measure. In the case of using the Mahalanobis distances of the whole frames without VQ information which is researched in previous study [7], Overlapping Ratio is 61.2% which is lower than that of DSM in the case of using single level VQ and multi-level VQ. From this result, we can see that the proposed method using VQ information is efficient to find a repeated part in music content. Table 3 shows a comparison between former researches and the proposed one.

Table 4. Performance of Fixed-Length Music Summary (M = 8, 16).

Summary Length	Optimal Overlapping Ratio (%)	
	CSM	DSM
20s	72.6	75.1
30s	80.9	80.8
40s	84.9	86.4
50s	89.4	90.2

Table 5. Optimal Overlapping Ratio (Average Summary Length over 100 songs).

Range of Length	CSM	DSM
20s ~ 40s	87.3 (37.1s)	88.2 (37.6s)
20s ~ 50s	90.3 (42.3s)	90.7 (43.3s)

3.2. Evaluation of optimal-length music summary

To evaluate the performance of optimal-length music summary, we use different performance measure modified from the Overlapping Ratio used in the previous section. The new measure called Optimal Overlapping Ratio is given in equation (11).

$$\text{Optimal Overlapping Ratio} = \max_{k=1,\ldots,K}\left[\frac{length(S_{hand,k} \cap S_{VQ})}{length(S_{hand,k})}\right] \times 100 \quad (11)$$

where K denotes the number of repeated parts in the music used. The Optimal Overlapping Ratios of fixed-length music summary and optimal-length music summary are shown in Table 4 and 5.

From the table 5, the performances of CSM and DSM are similar from Optimal Overlapping Ratio point of view. In the case of the range of summary length is 20s~50s, the average summary length and Optimal Overlapping Ratio of optimal-length music summary using DSM are 43.6s and 90.7%. However, in case of fixed-length music summary using DSM, Optimal Overlapping Ratio is 90.2% with predefined summary length, 50s. From this result, optimal-length music summary generates a repeated part more effectively with shorter length than fixed-lengthmusic summary.

IV. Conclusion

In this paper, we proposed two kinds of music summary algorithm which give fixed-length music summary or

optimal-length music summary, and also two kinds of objective test measures. To find a repeated part in music, we use weighted sum of counts or the reciprocal of the Mahalanobis distances of frames with the same codeword based on multi-level vector quantization. The results of proposed method which generates fixed-length music summary show better performance than former researches. And the algorithm for optimal-length music summary generates music summary that contains more repeated part with proper length than that for fixed-length music summary.

References

1. L. Mani, M. T. Mabury, *Advances in Automatic Text Summarization*, (MIT Press, 1999)
2. Y. Gong and X. Liu, "Summarizing video by minimizing visual content redundancies," in Proc. IEEE International Conference on Multimedia and Expo, 788-791, Tokyo, Japan, 2001.
3. I. Yahiaoui, B. Merialdo and B. Huet, "Generating summaries of muli-episode video," in Proc. IEEE International Conference on Multimedia and Expo, 792-795, Tokyo, Japan, 2001.
4. B. Logan and S. Chu, "Music summarization using key phrases," in Proc. IEEE International Conference on Audio, Speech and Signal Processing, 749-752, 2000.
5. C. Xu, Y. Zhu, and Q. Tian, "Automatic music summarization based on temporal, spectral and cepstral feature," in Proc. IEEE International Conference on Multimedia and Expo, 117-120, 2002.
6. G. Peeter, A. L. Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in Proc. International Symposium on Music Information Retrieval, 2002.
7. C. Xu, X. Shao, N. C. Maddage, M. S. Kankanhalli, and Q. Tian, "Automatically summarize musical audio using adaptive clustering," in Proc. IEEE International Conference on Multimedia and Expo, 2063-2066, 2004.
8. X. Shao, N. C. Maddage, C. Xu, and M. S. Kankanhalli, "Automatic music summarization based on music structure analysis," in Proc. ICASSP, 2, 1169-1172, 2005.
9. X. Shao, C. Xu, Y. Wang, and M. S. Kankanhalli, "Automatic music summarization in compressed domain," in Proc. ICASSP, 4,261-264, 2004.
10. X. Shao, C. Xu, and M. S. Kankanhalli, "A new approach to automatic music video summarization," in Proc. ICIP, 625-628, 2004.
11. C.Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization," IEEE Tans. Speech and Audio Process. 13 (3) May 2005.
12. Seoyoung Koh, Jeongsik Park, and Yung-hwan Oh, "Improvement of mp3-based music summarization using linear regression," in Proc. The KSPS Fall Conference, 55-58, 2005.
13. L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, (Prentice-Hall, 1993)

[Profile]

• **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, 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 robust speech recognition and speaker recognition.

• **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, Korea in 2002 and 2007, respectively. He is a researcher in Samsung electronics.

• **Hoirin Kim**

Hoirin Kimreceived 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. 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.