2007－MUS－72　(6)
2007／10／11

音響信号の主要部分抜粋による楽曲のバージョン同定

Bee Suan Ong[1], Emilia Gómez, Perfecto Herrera
Universitat Pompeu Fabra

音響信号の楽曲構造記述に基づいて楽曲の主要部分の抜粋を構成し，これを用いて一楽曲の異なるバージョンを同定する新しい手法を提案する．楽曲バージョン同定の用途に適した抜粋箇所の選択という点に関し，我々の提案する選択基準(反復等価性重視)を，楽曲要約に一般的に用いられる選択基準(反復性重視)と比較した．さらに，[1]と同一のデータベースを用いて提案手法の評価を行なった．実験の結果，我々の提案する抜粋手法は，楽曲全体を用いる手法にくらべ，楽曲バージョン同定において有意にすぐれた性能を示した．

Extracting Representative Audio Excerpts for Song Version Identification

Bee Suan Ong[1], Emilia Gómez, Perfecto Herrera
Universitat Pompeu Fabra

This paper propose a novel approach towards retrieving different versions of the same song by exploiting the representative audio excerpts from audio signals, based on its music structural descriptions. In detecting such excerpts for song version retrieval purposes, we compared our selection criterion (repetitive-equivalence emphasized) with the one that is commonly used in music summarization (repetitiveness emphasized). Additionally, we evaluated our method on the same database as presented in [1]. The experimental results show that our proposed excerpt approach has achieved a significant better performance in song version identification compared with using the whole length of the piece.

1. INTRODUCTION

Current literature in audio-based music structural analysis and discovery mainly point towards audio browsing and music summarization or thumbnailing contexts. So far there exists no exploration with regards to the practical usability of music structural descriptions in other contexts besides the above mentioned area. In this paper, we investigate the applicability of music structural description to the song version identification problem. Automatic audio cover song or version identification, which is closely related to song similarity and retrieval, has been receiving much attention from the MIR community lately. Various approaches have been proposed for the song version identification task, such as methods based on melodic similarity [2], beat synchronous chroma features [3], statistical spectrum descriptors [4] and so forth. A common property of these existing methods is that the whole length of music is to be considered for retrieval. The main shortcoming of this common property is that different versions of the same piece of music may vary in its musical structure. Thus, by comparing the whole length of the root query and its version, it may show a low similarity between the two pieces. Considering this issue, we propose the use of short representative excerpts extracted or summaries from audio signals, based on its music structural description to identify song versions in music collections. Another advantage of using short excerpts instead of full tracks lies on the computational efficiency when working on large collections (depending on the actual type of similarity computation).

This paper is organized as follows: In section 2, we describe how to compute structural descriptions directly from audio signals. Short-summary approach for song version identification is presented in section 3. Section 4 includes quantitative evaluation and comparison results with an existing approach. Finally, overall conclusion and future work are discussed in section 5.

2. MUSIC STRUCTURAL DISCOVERY

Music structural discovery is the first step towards generating structural descriptions directly from music signals. Our structural description system [5] is developed based on further improvement upon an existing method for

[1] The author is currently employed by YAMAHA Corporation, Japan

detecting chorus sections in music [6], to produce a compact representation of music structure through labelling and time-stamping marking (dis)similar sections that appear in the music signals (i.e. verse, chorus, bridge, etc.).

Figure 1 gives the overview framework of our automatic structural description system. We first segment the input signal into overlapped frames (4096-sample window length) with the hop size of 512 samples. Then we extract octave equivalence pitch class distribution features, HPCP [5], of each of these frames to obtain short-term description of the input audio signal. In order to avoid the system from having high computational load by processing the complete set of HPCP feature vectors, the system computes the average of each 10 extracted feature frames to represent the tonal distributions of the original input signal of every 116 ms, approximately. With the computed mean feature values, we measure the (dis)similarity distance between each 116ms of the tonal descriptors using the cosine distance measure. For easing the processing of identifying repetitive segments in music, we compute the time-lag matrix of the similarity representation, by orientating the diagonal of the computed similarity matrix towards the vertical axis. We then apply matrix binarization and morphological filtering operations to get rid of redundancies and remove line segments that are too short to contain any significant repetition in music.

In detecting repetitive segments in music, we adopted Goto's approach [6] by calculating the possibility of containing line segments of each lag. In the line segments integration process, we first organize the detected repetition pairs into groups. Apparently, line segments that share a common line segment are the repetitions of one another and should be given the same labelling. Based on this observation, we integrate those line segments, which share a common line, into one group with the same label. It is then followed by computing distance measures through selecting the first line segment of each group and correlating it along the pre-processed features. This is for the purpose of recovering undetected repetitions that we have missed in the previous detection process. Based on the computed distances, we use an adaptive threshold, defined by the summation of the lowest occurring distance value with a fixed tolerance margin, to determine significant repetitions appearing in the music. Then all local minima falling below the threshold are considered to be relevant to the occurrence of repetition.

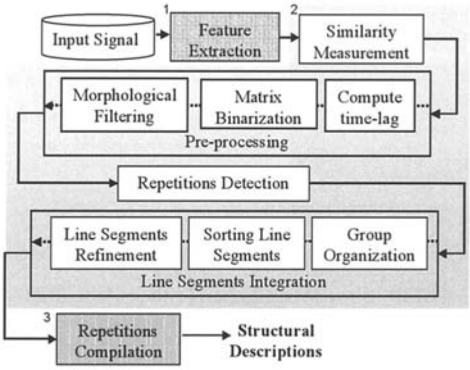

Figure 1. Overview framework of the automatic structural descriptions system.

For generating the music structural description, we select the three most repetitive groups. We compile the repetitive segments by lining up all the line segments of these repetitive groups according to their labels. If there exists an overlap between two particular labels, all the overlapped sections of these two labels will be given a new label, whereas the non-overlapped sections will be given another label. Unlabelled sections between all the labelled segments will be given a new label respectively as a new repetition group by itself. We then select one line segment of each label and perform another repetition detection procedure by correlating it with the pre-processed feature. The repetition detection process terminates when all labels have been examined. Finally, based on the assumption that structural sections in music (i.e. intro, verse, chorus, etc.) are less than 25 sec in length, we generate the final structural description of music by combining all the repeated labels, with the length of less then 25 sec to become a single label.

3. SHORT SUMMARY APPROACH

In identifying representation excerpts or summaries of music, most literature pays great attention to the significance of repetitions in music. In the existing literature [7] [8], the most repetitive segments are considered as the most significant excerpts to represent a piece of music. Considering its application context in version retrieval, we explore the potential of some other factors that could be useful to retrieve songs with its different versions. In our short-summary approach, we investigate two ways of identifying representative excerpts of music for version identification purposes with the use of Harmonic Pitch Class Profiles (HPCP) features [1]. Following the commonly used criteria, the first approach emphasizes more on the significance of the most repetitive excerpts in music. The second approach considers all repetitions as equivalent. Thus, the total duration of all identical repeated patterns are taken as the highest priority factor in selecting the best suitable audio excerpts to represent a piece of music. Based on the structural description obtained via music structural discovery, we categorize all the repeated segments into groups according to their labels.

3.1. Repetitiveness Emphasized

The number of elements in a group denotes the occurrence frequency of a repeated pattern. Thus, the first approach, which emphasizes the significance of the most repetitive excerpts in music, selects the group containing the most its repeated segments and extracts a fixed duration, l seconds, from the starting-time information of the group's first segment.

3.2. Repetitiveness-Equivalence Emphasized

As mentioned earlier, different versions of the same piece of music may vary in its musical structure. The most repetitive segments of the query song may not appear to be the most repetitive segments in its song versions. Considering this issue, we generate two short summaries or segments from a song in order to overcome instances which have variances in its musical structure between the root songs and its versions. Music summaries are generated based on the following two criteria:

(i) The selected segments are repeated at least once in the whole song.

(ii) The selected repeated groups should hold the majority of the song duration compared with other repeated groups.

Since all the repeated segments within the same group have approximately the same length, we calculate the total length that each label subsumes in a piece by multiplying the length of its one segment with its total number of segments, n. With the above mentioned selection criteria, we select one segment from each of the first two groups, which holds the longest duration of the song, to compute music summaries. Finally, we extract a fixed duration, l seconds, from each selected segment based on their starting-time information.

4. DYNAMIC TIME WARPING

Research work by Gómez [1] provides a successful example of version identification by means of analyzing the similarity of tonal features between music pieces. Thus, following previous research work [1], we compute the instantaneous evolution of HPCP for both short summaries extracted from each song query and all the songs in the database. In order to measure similarity between two pieces, we apply the Dynamic Time Warping (DTW) algorithm, which estimates the minimum cost required to align one piece to the other one, on short summaries belonging to both pieces alternately as shown in Figure 2. Here, we can see that there appear four similar measures for each pair of comparisons. Finally, we choose the highest similarity among the four values to represent the similarity estimation between two pieces.

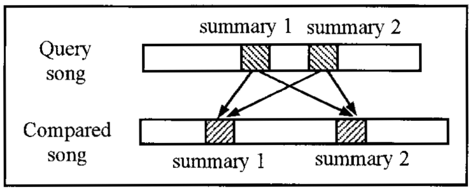

Figure 2. The comparison of summaries between two songs.

Since DTW actually performs a direct comparison between summaries from both pieces and considers that versions of the same piece do not necessarily maintain the same key (key change) as the original, we need to transpose the compared summaries to the same key as the query before computing similarity. One of the advantages of HPCP, being an octave equivalence tonal descriptors, is that ring shifting of the feature vectors correspond to the transposition in music perception. Thus, we generated 12 different sets of the shifted feature vectors for each compared summary to evaluate the similarity between the query summaries and the 12 semitone transpositions of the compared summaries. Following that, we apply the DTW algorithm to query summaries and each of the 12 transposed compared summaries alternately to estimate the minimum cost between two summaries. Finally, the lowest estimated minimum cost is selected to represent the similarity between two songs.

5. EVALUATION

5.1. Dataset

The goal of this study is to evaluate the applicability of structural descriptions in identifying different versions of a piece of music. Thus, we reuse the dataset described in [1], which consists of 90 versions from 30 different songs (root query) of popular music as our test set. For this evaluation, we will compute a similarity measure between two different pieces based on low-level tonal descriptors, i.e. HPCP values. We will compare the efficiency of version identification obtained through the full length of the song with the one obtained through the song summaries.

5.2. Quantitative Measurements

Version identification, which involves song query and retrieval, is a type of information retrieval system. Thus, for evaluation purposes, we use IR standard measures, such as recall and precision, to rate effectiveness of the retrieval. The recall rate is defined as the ratio of the number of relevant returned documents to the total number of relevant documents for the user query in the collection. Whereas the precision rate is the ratio of the number of relevant returned documents to the total number of documents for a given user query.

To investigate the influence of the length, l, of the short summaries on the performance of version identification, we extract various durations from the range of 15 seconds to 25 seconds with an interval of 5 seconds from the audio signal. To estimate the optimal or upper bound performance of using summaries in version identification with our test set, we manually select two short segments (approximately 25 seconds depending on the tempo of the music), which are repeated in all the versions of the same songs, according to their time-varying harmonic contour in the segments. We substitute the manually selected segments for short summaries extracted based on music structural descriptions to represent the song itself. Whereas for estimating the lower bound of performance with the use of the short-summary approach, we randomly select two 25-seconds short segments for each song in the test set to represent the music itself. Finally, we compute similarity measures using the randomly selected or manually selected short segments. As explained above, we then select the highest similarity among the four values to represent the similarity estimation of the root query and the compared song.

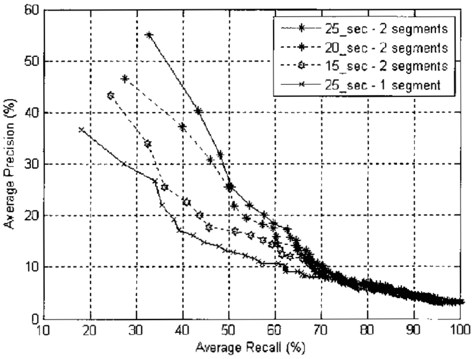

Figure 3. The performances of song version identification using various numbers of short summaries of different lengths.

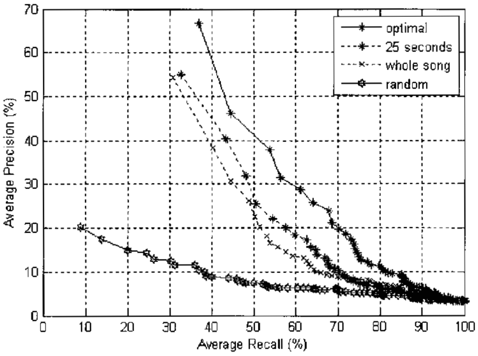

Figure 4. The performances of song version identification: whole-song approach vs. short-summary approach.

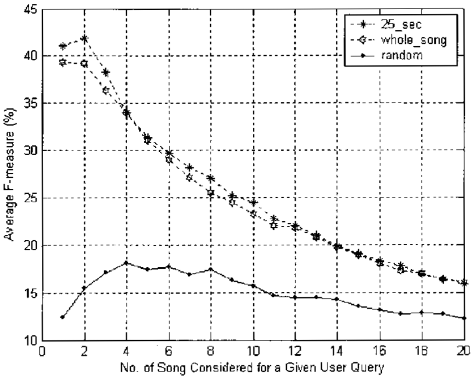

Figure 5. Average F-measures of both approaches (short-summary and whole-song) in version identification according to the number of songs considered for a given query.

5.3. Experimental Results

Figure 3 shows the performance of version identification using various members of short summaries extracted from the songs in different segments' lengths. From our results, we observe that the best performance is in the case of 25-seconds with two segments, which achieves a high precision and recall rates of 55.1% and 32.8%, respectively. As expected, the performances become impaired when the extracted summaries from the audio signals are decreased in length. For the case of 20-seconds, the performance achieves the precision and recall rates of 46.7% and 27.6%, respectively, whereas for the case of 15-seconds, the performance only scores 43.3% and 24.4% in its precision and recall measures. For the case where repetitiveness emphasis is applied on the short-summary approach, where only one 25-second summary is extracted from the songs, the achieved precision rate is the lowest, 36.7% with a recall level of 18.1%.

Figure 4 shows the performance of version identification using the whole-song approach versus the short-summary approach based on its average precision and recall measures. From the precision-recall graph, we observe that by using two extracted short summaries (with the length of 25 seconds each) from the songs, we can achieve a slightly better performance in version identification compared with using the whole length of the piece. By only considering the first retrieved song for a given user query, the short-summary approach exceeds 0.6% and 2% in its precision and recall rates respectively compared with the whole-song approach. The estimated upper bound results for identifying different versions of the same song reaches the precision and recall rates of 66.6% and 36.8%, respectively. Whereas by using randomly extracted short summaries from songs, the achieved precision rate is very low, 22.2% with a recall level of 9.0%. The short-summary approach, besides its better accuracy compared with the whole song approach, also consumes less time in performing version identification tasks. For our test set, which consisted of 90 audio data with an average audio length of 3 minutes and 45 seconds, the short summary approach accomplishes the identification task at least 33% faster than the other approach. Figure 5 plots the average F-measures obtained from both approaches considering various numbers of songs for a given query. The statistical t-test shows that the obtained average F-measures from the short-summary approach is significantly higher than those from the whole-song approach with the test result of t(19)=3.966, p<0.01 beyond the 99% confidence level.

Through analyzing the low performance of a few query songs, we have realized that there occurs an issue with regards to the transitivity relationship between songs due to our two extracted short summaries comparison approach (see Figure 6). For instance, if Song-A has two summaries with each appearing in Song-B and Song-C, by querying Song-A, we will be able to find both Song-B and Song-C as its versions. However if Song-C happens to have summaries which appear one in Song-A but none in Song-B, by querying Song-C, we will only find Song-A but miss Song-B since we do not infer any relationship between songs. Nevertheless, the failure in this aspect could be exploited or considered interesting for generating an additional source of metadata that is not directly stored in the database. Seeing that cover songs tend to imitate the original song, by inferring the transitivity relationships among different versions of the same song, it would provide clues to defining the original song among its different versions. For instance, in the above given example, the present of version relationships of Song-B and Song-C with Song-A respectively but not within themselves (Song-B and Song-C) may imply that Song-A could be considered the canonical song (the original song version).

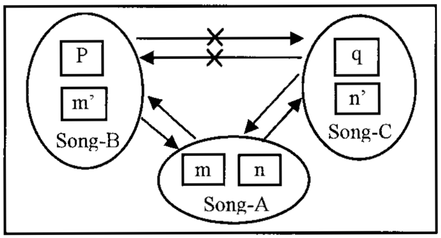

Figure 6. Transitivity relationship between songs.

6. CONCLUSION AND FUTURE WORK

In the experiment results, perhaps the most notable result from this chapter's experiment is the distinctive dichotomy in performance between the two distinct selection criteria (repetitiveness emphasis vs. repetitive equivalence emphasis) in identifying representative excerpts of music for version identification applications. For the segment selections that make use of a complementary musical knowledge (i.e. repetitive equivalence emphasis), we see generally good performance. This dichotomy generally supports the notion that repetitiveness of music segments is important in identifying representative excerpts of music. However it is not the only assumption that we should rely on, depending on the application context. Incorporating musical knowledge related to the processing data (e.g. structural differences within the song versions) somehow improves performance.

Finally, as a conclusion of this small-scale of evaluation, we can also see that the short-summary approach seems to perform better than the whole-song approach in both retrieval accuracy and computational efficiency. From this study, we have observed a few advantages and disadvantages of using the short-summary approach in version identification compared with the whole-song approach. The advantages are:

(i) Time consumption factor – less time consuming and higher identification performance for the database, which consists of songs with long durations;

(ii) Modulation within piece – since only two short segments are extracted from the song itself, the performance accuracy is not to be affected by modulation within the pieces;

(iii) Different music structural descriptions in song versions – flexible to structural changes since only the core segments are extracted from the music itself;

Whereas, the disadvantages of using such an approach include:

(i) Identifying a song and its versions with large tempo variances – since short and fixed time constraints are applied in extracting summaries from the song, false negatives may occur for the query and its versions which have large differences in tempo;

(ii) Songs with short duration – applying such an approach to songs with durations shorter than double the extracted summaries length is more time consuming than the whole-song approach;

7. ACKNOWLEDGEMENT

This research has been partially funded by the EU-FP6-IST-507142 project SIMAC [2]. The author would like to thank Emilia Gómez and members of MTG in room 316 for their help and suggestions.

8. REFERENCES

[1] Gómez, E. *Tonal Description of Music Audio Signals.* PhD thesis. UPF, Barcelona, 2006.

[2] Sailer, C. and Dressler, K. "Finding Cover Songs by Melodic Similarity", *Third Music Information Retrieval Evaluation eXchange (MIREX), 2006.*

[3] Ellis, D. "Identifying Cover Song with Beat-Synchronous Chroma Features", *Third Music Information Retrieval Evaluation eXchange (MIREX), 2006.*

[4] Lidy, T. and Rauber, A. "Computing Statistical Spectrum Descriptors for Audio Music Similarity and Retrieval", *Third Music Information Retrieval Evaluation eXchange (MIREX), 2006.*

[5] Ong, B., Gómez, E. and Streich, S. "Automatic Extraction of Musical Structure Using Pitch Class Distribution Features", *1st Workshop on Learning the Semantics of Audio Signals (LSAS)*, 2006.

[6] Goto, M. "A Chorus-Section Detection Method for Musical Audio Signals", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. V-437-440, 2003.

[7] Logan, B. and Chu, S. "Music Summarization Using Key Phrase", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000.

[8] Bartsch, M. and Wakefield, G. "To Catch A Chorus: Using Chroma-Based Representations for Audio Thumbnailing", *IEEE WASPAA*, New Paltz, USA, 2001.

[2] http://www.semanticaudio.org