Audio Engineering Society

Convention Paper

Presented at the 121st Convention
2006 October 5–8 San Francisco, CA, USA

Automatic tonal analysis from music summaries for version identification

Emilia Gómez, Beesuan Ong and Perfecto Herrera

Music Technology Group, Universitat Pompeu Fabra, Barcelona 08003, Spain
{egomez,beesuan,pherrera}@iua.upf.edu

ABSTRACT

Identifying versions of the same song by means of automatically extracted audio features is a complex task to achieve using computers, even though it may seem very simple for a human listener. The design of a system to perform this job gives the opportunity to analyze which features are relevant for music similarity. This paper focuses on the analysis of tonal and structural similarity and its application to the identification of different versions of the same piece. This work describes the situations where a song is versioned and several musical aspects are transformed with respect to the canonical version. A quantitative evaluation is made using tonal descriptors, including chroma representations and tonality, combined with the automatic extraction of a summary of a piece through music structural analysis.

1. INTRODUCTION

The possibility of finding "similar" music pieces is one of the most attractive features that a system dealing with large music collections can provide. Similarity is an ambiguous term, and may depend on different musical, cultural and personal aspects. Many studies try to define and evaluate the concept of similarity, i.e., when two pieces are similar. There are many factors involved in this problem, and some of them (maybe the most relevant ones) are difficult to measure.

We focus here on the problem of identifying different versions of the same song, as working with versions seems to open a way for shedding some light on the factors involved in music similarity. When dealing with huge music collections, version identification is a relevant problem, because it is common to find more than one version of a given song (i.e. cover songs. We can identify different situations for this in mainstream popular music, as for example re-mastered, recorded live, acoustic, extended or disco tracks, karaoke versions, covers (played by different artists) or remixes. One example of the relevance of cover songs is found in the Second Hand Songs database (http://www.secondhandsongs.com), which contains around 37000 cover songs.

We have analyzed how tonal descriptors are useful to locate versions of the same song [2]. In these studies, we consider that two pieces are tonally similar if they share a similar tonal contour, related to the evolution of chords (harmony) and key. This research has revealed the significance of musical structure when comparing two pieces of music.

In the present paper, we investigate how the analysis of music structure and the extraction of a summary for each analyzed piece can improve previous results for version identification using tonal features. After describing the current system, we present some evaluation results and discuss how the identification of versions by means of tonal analysis can benefit from structural description.

2. FEATURE EXTRACTION

The block diagram for feature extraction is presented in Figure 1.

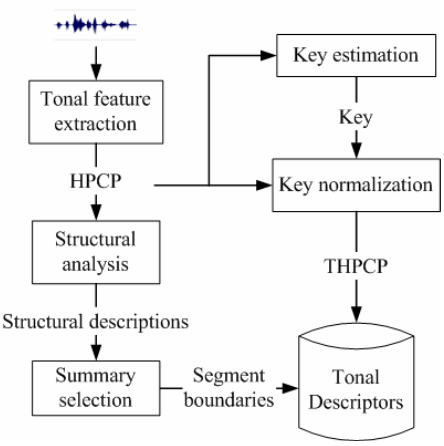

Figure 1: Block diagram for feature extraction

First, the audio signal is analyzed in order to extract the Harmonic Pitch Class Profile (HPCP) vector, which is obtained for each analysis frame. Then, these features are used to investigate the structure of the song in terms of pitch distributions, and then to obtain a short summary of the piece. For each analyzed song, we also perform an estimation of its key and normalize the HPCP values with respect to this key, in order to obtain the transposed version of the features (THPCP). We finally store the THPCP values for the obtained music summary. The different steps are further described in the following sections.

2.1. Tonal feature extraction

The tonal features used for this study are derived from the Harmonic Pitch Class Profile (HPCP), a pitch class distribution vector computed for each analysis frame. This profile (considering *size=12)* represents the relative intensity of each of the 12 semitones of the equal-tempered scale. It is extended to work with higher interval resolution, so that its size can be set to any multiple of 12 (*size=24, 36...*). In this study, we have used *size=120*, providing an interval resolution of 10 cents. The HPCP is computed following the schema represented in Figure 2.

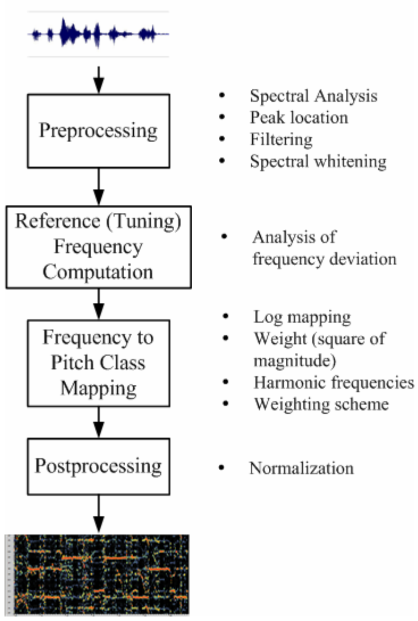

Figure 2: Block diagram for HPCP computation

First, there is a preprocessing stage including a set of procedures: spectral analysis by means of the Fourier Transform, spectral whitening, peak estimation and frequency filtering. Then, the obtained frequency values from spectral peaks are mapped into HPCP bins using a logarithmic scale, where the contribution of each frequency is proportional to the square of its spectral magnitude. In order to improve the accuracy and resolution of the mapping procedure, we include a weighting process based on considering the contribution of sub-harmonic and neighbor frequency values. For a more detailed explanation of the method, we refer to [3].

From the HPCP, we compute the transposed version of this profile (THPCP), which is obtained by normalizing the HPC with respect to the global key, which can be automatically extracted [3]. The THPCP represents a tonal profile which is invariant to transposition, being its first element always related to the tonic.

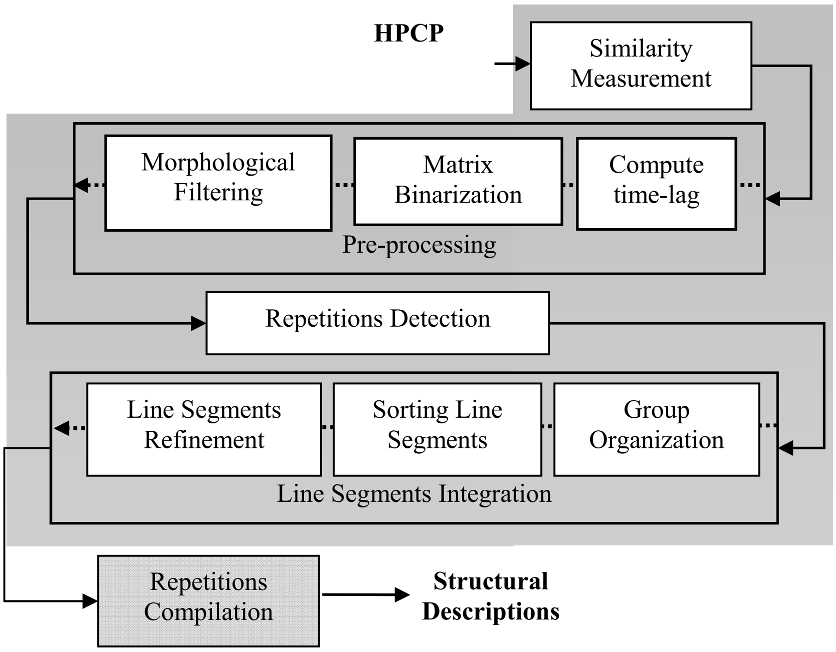

Figure 3: Overview of the framework for automatic structural description

For these two vectors, we consider in this study both the instantaneous evolution and the global average, computed for the whole piece and for just the obtained music summaries.

2.2. Structural analysis

In this step, the music structural description is obtained from automatic analyzing of audio signals. Our structural description system presented in this paper is based on Goto's method [4] for detecting chorus sections in music. We have introduced further improvements on this method to offer a more complete music structural description. It is done by marking (dis)similar sections that appear in the music signal (i.e. verse, chorus, bridge, etc.) through labelling and time-stamping of them. There are three main steps in our system, which are illustrated in Figure 3.

An explanation in detail regarding the methodology of extracting music structural description from audio signal is documented in [5].

2.3. Summary selection

Once the audio stream is segmented into different segments, this step is devoted to extract the most representative one, which is considered as the summary of the piece. Based on the structural descriptions obtained from the previous step, we categorize all the repeated segments into groups according to their labels, as follows:

$$Group_{A\_label} = \{Segment_1, Segment_2, ..., Segment_m\}$$

Here, *m* represents the number of repeated segments with label A.

For version identification purposes, we hypothesize that different versions of the same piece may have differences in its musical structure. In order to overcome this problem, we generate two music summaries from a song. These music summaries are generated based on the following two criteria:

1. The selected segments are repeated at least once within the song.

2. The selected repeated segment 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 of each group within a piece by multiplying the length of one segment by the total number of segments. 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. We use a segment duration from 15 to 25 seconds with an interval of 5 seconds based on the begin time of each selected segment as the final summaries.

2.4. Similarity measures

In order to measure similarity between global HPCP and THPCP vectors, computed as averages over the considered segment, we use the Pearson's correlation coefficient. Figure 2 shows the THPCP global profile for 6 different versions of the first phrase of the song *Imagine*, by John Lennon.

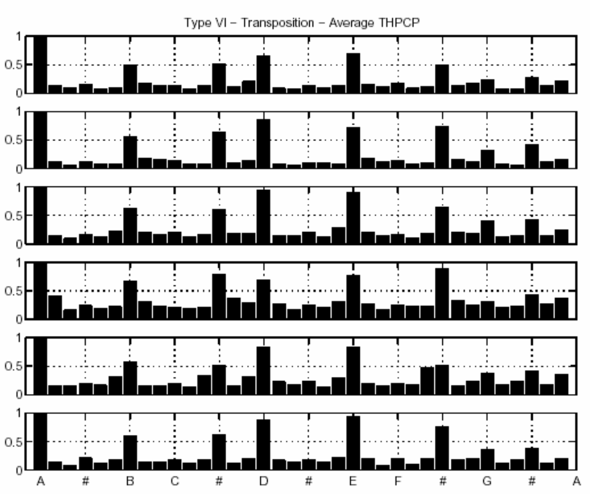

Figure 4: Global THPCP for 6 different versions of the first phrase of the song *Imagine,* by John Lennon

For instantaneous HCPC and THPCP features, we use a Dynamic Time Warping (DTW) algorithm, which estimates the minimum cost required to align one piece to the other one, and is based on [1].

Figure 5 shows the similarity matrix built upon instantaneous THPCP between a version of the song *Imagine* by John Lennon, performed by Diana Ross, and the original piece.

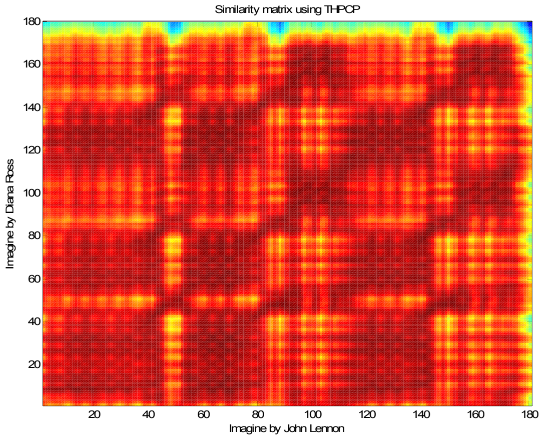

Figure 5: Similarity matrix, built from THPCP values, and DTW trajectory for two versions of the song *Imagine,* by John Lennon (original, x axis) and by Diana Ross (y axis)

As mentioned before, these similarity measures are computed both for the entire piece and for the two extracted music summaries. Figure 6 illustrates how we compute the final similarity measure based on the summaries extracted from the pieces.

As two summaries are extracted, there appear four similarity measures for every two compared pieces. We choose the highest value to represent the similarity between two pieces.

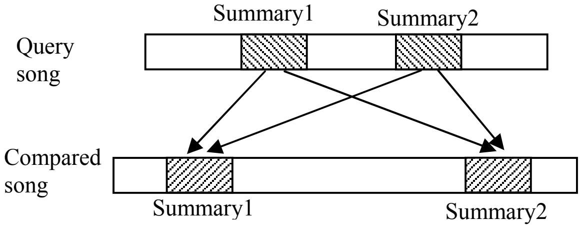

Figure 6: Comparison between the extracted summaries from two pieces

3. EVALUATION

3.1. Evaluation material

The evaluation collection used is composed by 90 versions from 30 different songs taken from a music collection of popular music. The versions include different levels of similarity to the original piece, which are found in popular music: noise, modifications of tempo, instrumentation, transposition and modifications of the main melody and the harmonization. The average number of versions for each song is equal to 3.07, and its standard deviation is 1.65. We are then dealing with the most difficult examples, so that the evaluation can be representative of a real situation when organizing digital music collections of mainstream popular music. We refer to http://www.mtg.upf.edu/~egomez/versionid for a more detailed description of the collection.

3.2. Evaluation measures

In order to evaluate the performance of different similarity measures, we use precision and recall values. For each query, the query song is deleted from the retrieval database. The precision value obtained by randomly selecting songs from the collection is equal to 3.196%, with a maximum F measure equal to 0.0619%. This value will be considered as a baseline for evaluation. Two different strategies have been used: automatically extracted summaries and manually selected summaries. The later is used to estimate the optimal or upper bound results of our version identification process using music summaries. In this case, we manually select two segments, which are repeated in all the versions of the same song, according to their time-varying harmonic contour and some musical knowledge. They substitute the two automatically extracted summaries as explained in section 2.3. The selected segments are roughly 25 seconds in length depending on the tempo of the music. We have also computed the similarity measures using these manually selected summaries.

In order to estimate the lower bound results, we randomly extract two segments for each piece, with approximate duration of 25 seconds. When computing the similarity between two songs, we choose the highest similarity among the four values as explained in Figure 6, representing the similarity between the two pieces.

3.3. Results

Tonal similarity measures over the whole song

Figure 7 shows the precision and recall vales for version identification considering all the song. The maximum precision obtained using the whole song is equal to 54.5%, with a recall level of 30.8%, obtaining an F-measure of 39.3%. The comparison of different similarity measures shows that relative descriptors (THPCP) seem to perform better than absolute pitch class distribution features, which is coherent with the invariability of melodic and harmonic perception to transposition. Also, it seems that it is important to consider the temporal evolution of tonality (captured with DTW) which is sometimes neglected.

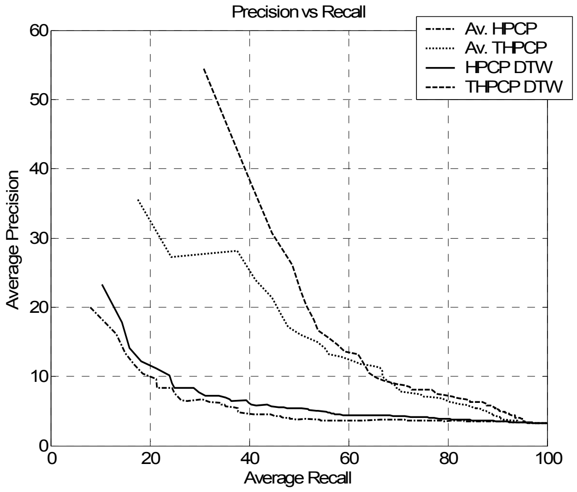

Figure 7: Precision and recall for version identification considering all the song.

Tonal similarity measures over music summaries

Figure 8 shows precision and recall measures for version identification considering music summaries. We consider the use of the tonal descriptors which provided better result in the previous experiment, THPCP and a Dynamic Time Warping algorithm to compute the similarity measure. We present the accuracy for the analysis of the whole song, manually annotated summaries (optimal), automatically extracted summaries (with different durations: 15, 20 and 25 seconds), and randomly selected summaries of 25 seconds.

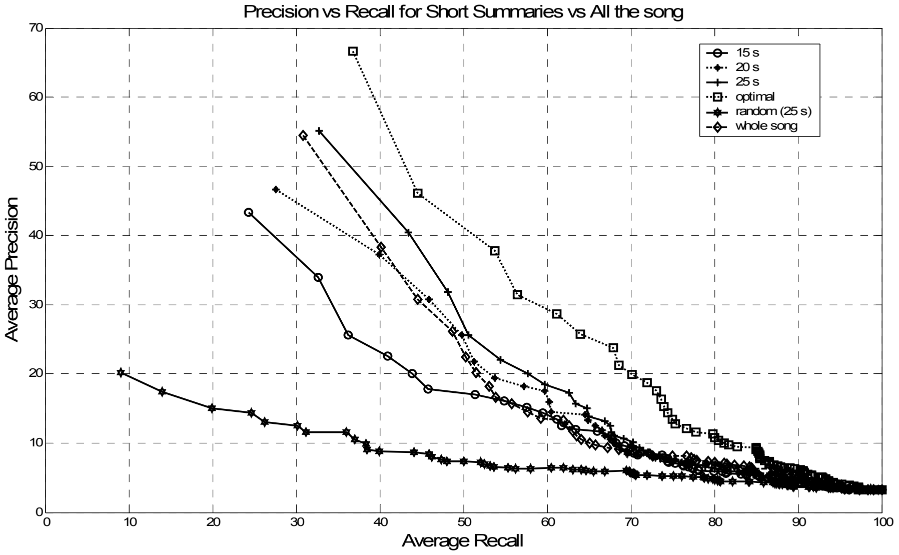

Figure 8: Precision and recall for version identification considering music summaries.

We observe that the best results are obtained when using manually labeled summaries (*optimal)*, equal to a precision of 66.67%, a recall level of 36.82% and F measure equal to 47.44%. The use of manually annotated summaries provides an improvement around 10% in the F measure. These results are followed by automatically computed summaries of 25 seconds duration, obtaining a precision of 55.06%, a recall level of 32.77% and F measure equal to 41,08%, which provides a modest improvement from analyzing the whole piece. Finally, the worse results are obtained when using randomly selected summaries from the songs, equal to 20.22% precision, recall level of 8.99% and F measure equal to 12.44%. All these results surpass the baseline of precision (3.196%) that would be obtained by randomly selecting pieces of the collections. We then verify that the use of music summaries provides some improvements with respect to the analysis of the whole piece but still leaves room for additional enhancement.

4. CONCLUSIONS

We have presented a system to identify versions of the same song by using tonal descriptors and music summaries. We observe that the use of structural analysis improves the accuracy of the system, obtaining a maximum precision of 66.67%. More work is needed in order to improve the similarity measures, the procedure for summary selection and increase the evaluation collection.

5. REFERENCES

[1] Ellis, D. last accessed July 2006. *Dynamic Time Warp (DTW) in Matlab*, resource available on-line http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/

[2] Gómez, E. and Herrera, P. 2006. *The song remains the same: identifying versions of the same song using tonal descriptors*, International Conference on Music Information Retrieval, Victoria, Canada.

[3] Gómez, E. 2006. *Tonal description of polyphonic audio for music content processing*. INFORMS Journal on Computing, Special Cluster on Computation in Music Vol.18 .3

[4] Goto, M. 2003. *A Chorus-Section Detecting Method for Musical Audio Signals*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICCASP), pp. V-437-440.

[5] Ong B., Gómez E., and Streich S. 2006. *Automatic Extraction of Musical Structure Using Pitch Class Distribution Features*, in preparation.