The song remains the same: identifying versions of the same piece using tonal descriptors

Emilia Gómez
Music Technology Group, Universitat Pompeu Fabra
Ocata, 1
08003, Barcelona
emilia.gomez@iua.upf.edu

Perfecto Herrera
Music Technology Group, Universitat Pompeu Fabra
Ocata, 1
08003, Barcelona
perfecto.herrera@iua.upf.edu

Abstract

Identifying versions of the same song by means of automatically extracted audio features is a complex task for a music information retrieval system, even though it may seem very simple for a human listener. The design of a system to perform this task gives the opportunity to analyze which features are relevant for music similarity. This paper focuses on the analysis of tonal similarity and its application to the identification of different versions of the same piece. This work formulates 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. A simple similarity measure, based on Dynamic Time Warping over transposed chroma features, yields around 55% accuracy, which exceeds by far the expected random baseline rate.

Keywords: version identification, cover versions, tonality, pitch class profile, chroma, audio description.

1. Introduction

1.1. Tonality and music similarity

The possibility of finding "similar" pieces is one of the most attractive features that a system dealing with large music collections can provide. Similarity is a ambiguous term, and music similarity is surely one of the most complex problems in the field of MIR. Music similarity may depend on different musical, cultural and personal aspects. Many studies in the MIR literature 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.

Some studies intend to compute similarity between audio files. Many approaches are based on timbre similarity using low-level features [1, 2]. Other studies focus on rhythmic similarity. Foote proposes some similarity measures based on the "beat spectrum", including Euclidean distance, a cosine metric or inner product [3]. Tempo is also used to measure similarity in [4]. The evaluation of similarity measures is a hard task, given the difficulty of gathering ground truth data for a large quantity of material. Some researchers assume that songs from the same style, by the same artist or on the same album are similar [5, 6, 7]. A direct way to measure the similarity between songs is also to gather ratings from users (see [4]), which is a difficult and time-consuming task.

Tonality has not been much applied to music similarity, as it might be not so clear for people not having a musical background. We focus here on analyzing how tonal descriptors can be used to measure similarity between pieces.

We consider that two pieces are *tonally* similar if they share a similar tonal structure, related to the evolution of chords (harmony) and key. We will assume that two pieces are similar if they share the same *tonal contour*. For song similarity, tonal contour could be as relevant as melodic contour is for melody recognition[8]. We focus then on the problem of identifying different versions of the same song, and study the use of tonal descriptors for this task.

1.2. Version identification

When dealing with huge music collections, version identification is a relevant problem, because it is common to find more than one version of the a given song. 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 [1] , which already contains around 37000 cover songs.

A song can be versioned in different ways, yielding different degree of dissimilarity between the original and the versioned tune. The musical facets that are modified can be instrumentation (e.g. leading voice or added drum track), structure (e.g. new instrumental part, intro or repetition), key (i.e. transposition) and harmony (e.g. jazz harmonization). These modifications usually happen together in versions from popular music pieces. The degree of disparity on the different aspects establishes a vague boundary between

[1] http://www.secondhandsongs.com

what is considered a *version* or what is really a different composition. This frontier is difficult to define, and it is an attractive topic of research from the perspective of intellectual property rights and plagiarism. The problem has conceptual links with the problem of analogy in human cognition, which is also an intriguing and far from being understood topic. This is the problem also when developing computational models to automatically identify these versions with absolute effectiveness.

There is few literature dealing with the problem of identifying versions of the same piece by analyzing audio. Yang proposed an algorithm based on spectral features to retrieve similar music pieces from an audio database [9]. This method considers that two pieces are similar if they are fully or partially based on the same score. A feature matrix was extracted using spectral features and dynamic programming. Yang evaluated this approach using a database of classical and modern music, with classical music being the focus of his study. 30 to 60 second clips of 120 music pieces were used. He defined five different types of "similar" music pairs, with increasing levels of difficulty. The proposed algorithm performed very well (90% accuracy) in situations where the score is the same and there are some tempo modifications, which is the worst case figure. On the same idea, Purwins et al. calculate the correlation of constant Q-profiles for different versions of the same piece played by different performers and instruments (piano and harpsichord) [10] .

2. Tonal feature extraction

The tonal features used for this study are derived from the *Harmonic Pitch Class Profile* (HPCP). The HPCP is a pitch class distribution (or chroma) feature computed in a frame basis using only the local maxima of the spectrum within a certain frequency band. It considers the presence of harmonic frequencies, as it is normalized to eliminate the influence of dynamics and instrument timbre (represented by its spectral envelope). From the instantaneous evolution of HPCP, we compute the transposed version of this profile (THPCP), which is obtained by normalizing the HPCP vector with respect to the global key. The THPCP represents a tonal profile which is invariant to transposition. For these two features, we consider both the instantaneous evolution and the global average. We refer to [11, 12] for further explanation on the procedure for feature extraction.

In order to measure similarity between global features, we use the correlation coefficient. As an example, the correlation between HPCP average vectors for two distant pieces is equal to 0.0069. This small value indicates the dissimilarity between the profiles, and can be considered as a baseline. For instantaneous features, we use a Dynamic Time Warping (DTW) algorithm. Our approach is based in [13]. The DTW algorithm estimates the minimum cost required to align one piece to the other one by using a similarity matrix.

3. Case study

We analyze here the example of four different versions of the song *Imagine*, written by John Lennon. The main differences between each of the versions and the original song is summarized in Table 2.

We first analyze how global tonal descriptors are similar for these different pieces. In order to neglect structural changes, we first consider only the first phrase of the song, which is manually detected. For the last version, performed by two different singers, we select two phrases, each one sung by one of them, so that there is a total of 6 different audio phrases. HPCP average vectors are shown in Figure 1.

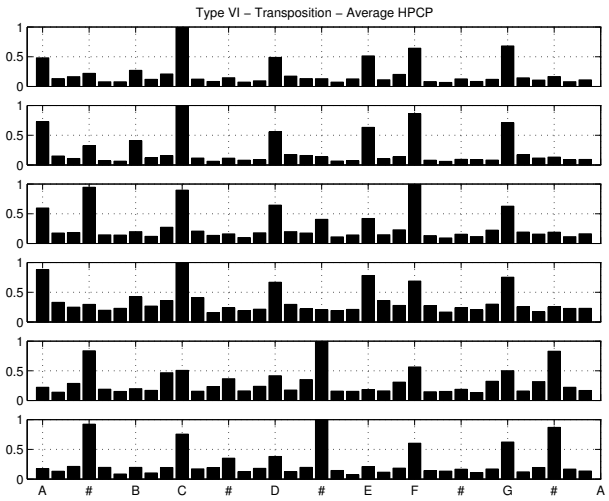

Figure 1. HPCP average for 6 different versions of the first phrase of *Imagine*. 1. John Lennon, 2. Instrumental, guitar solo, 3. Diana Ross, 4. Tania Maria, 5. Khaled and 6. Noa.

The correlation matrix R_{phrase} between the average HPCP vectors for the different versions is equal to:

$$R_{phrase} = \begin{pmatrix} 1 & 0.97 & 0.82 & 0.94 & 0.33 & 0.48 \\ 0.97 & 1 & 0.86 & 0.95 & 0.31 & 0.45 \\ 0.82 & 0.86 & 1 & 0.75 & 0.59 & 0.69 \\ 0.94 & 0.95 & 0.75 & 1 & 0.18 & 0.32 \\ 0.33 & 0.31 & 0.59 & 0.18 & 1 & 0.95 \\ 0.48 & 0.45 & 0.69 & 0.32 & 0.95 & 1 \end{pmatrix} \quad (1)$$

Table 1. Classification of tonal features used for similarity.

Feature	Pitch-class representation	Temporal scope
HPCP	Absolute	Instantaneous
THPCP	Relative	Instantaneous
Average HPCP	Absolute	Global
Average THPCP	Relative	Global

Table 2. Details on versions of the song *Imagine*.

ID	Artist	Modified musical facets	Key
1	John Lennon	Original	C Major
2	Instrumental	Instrumentation (solo guitar instead of leading voice)	C Major
3	Diana Ross	Instrumentation, tempo, key and structure	F Major
4	Tania Maria	Instrumentation, tempo, harmonization (jazz) and structure	C Major
5	Khaled and Noa	Instrumentation, tempo, key and structure	Eb Major

We can see that there are some low values of correlation between versions, mainly for the ones which are transposed to Eb major (5 and 6), as this tonality is not close to C major as F major is (3). THPCP average vectors are shown in Figure 2.

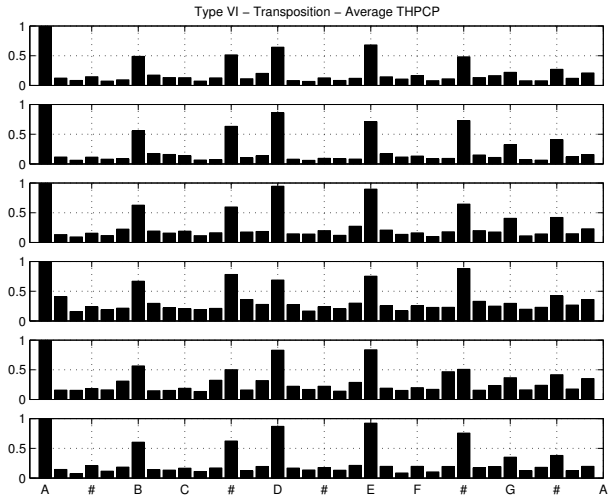

Figure 2. THPCP average for 6 different versions of the first phrase of *Imagine*. 1. John Lennon, 2. Instrumental, guitar solo, 3. Diana Ross, 4. Tania Maria, 5. Khaled and 6. Noa.

The correlation matrix $R_{t,phrase}$ between the THPCP average vectors for the different versions is equal to:

$$R_{t,phrase} = \begin{pmatrix} 1 & 0.97 & 0.97 & 0.94 & 0.94 & 0.97 \\ 0.97 & 1 & 0.98 & 0.95 & 0.91 & 0.98 \\ 0.97 & 0.98 & 1 & 0.92 & 0.95 & 0.99 \\ 0.94 & 0.95 & 0.92 & 1 & 0.86 & 0.94 \\ 0.94 & 0.91 & 0.95 & 0.86 & 1 & 0.95 \\ 0.97 & 0.98 & 0.99 & 0.94 & 0.95 & 1 \end{pmatrix} \quad (2)$$

This correlation matrix show high values for all the different versions, with a minimum correlation value of 0.86. When comparing complete songs in popular music, most of the versions have a different structure than the original piece, adding repetitions, new instrumental sections, etc. We look now at the complete 5 versions of the song *Imagine*, by John Lennon, presented in Table 2. The correlation matrix R between the average HPCP vectors for the different versions is equal to:

$$R = \begin{pmatrix} 1 & 0.99 & 0.83 & 0.96 & 0.45 \\ 0.99 & 1 & 0.86 & 0.95 & 0.45 \\ 0.83 & 0.86 & 1 & 0.79 & 0.65 \\ 0.96 & 0.96 & 0.79 & 1 & 0.35 \\ 0.45 & 0.45 & 0.65 & 0.35 & 1 \end{pmatrix} \quad (3)$$

We observe that the correlation values are lower for the piece in a distant key, which, in the case of version 5, is Eb major. We can again normalize the HPCP vector with respect to the key. THPCP average vectors are shown in Figure 3.

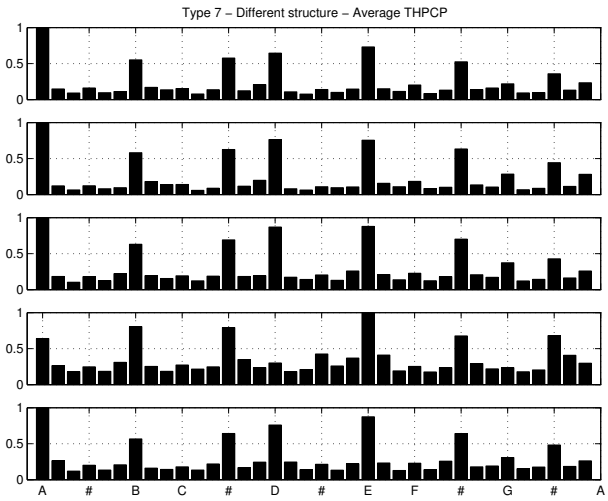

Figure 3. THPCP average for 5 different versions of *Imagine*.

The correlation matrix R_t between the average THPCP vectors for the different versions is equal to:

$$R_t = \begin{pmatrix} 1 & 0.99 & 0.98 & 0.96 & 0.98 \\ 0.99 & 1 & 0.99 & 0.95 & 0.98 \\ 0.98 & 0.99 & 1 & 0.95 & 0.99 \\ 0.96 & 0.95 & 0.95 & 1 & 0.95 \\ 0.98 & 0.98 & 0.99 & 0.95 & 1 \end{pmatrix} \quad (4)$$

We observe that the correlation values increase for version 5. In this situation, it becomes necessary to look at the structure of the piece. When the pieces under study have different structures, we study the temporal evolution of tonal features, in order to locate similar sections. Structural description is a difficult problem, and some studies have been devoted to this issue (see, for instance [14] and [15]). Foote [16] proposed the use of self-similarity matrices to visualize music. Similarity matrices were built by comparing Mel-frequency

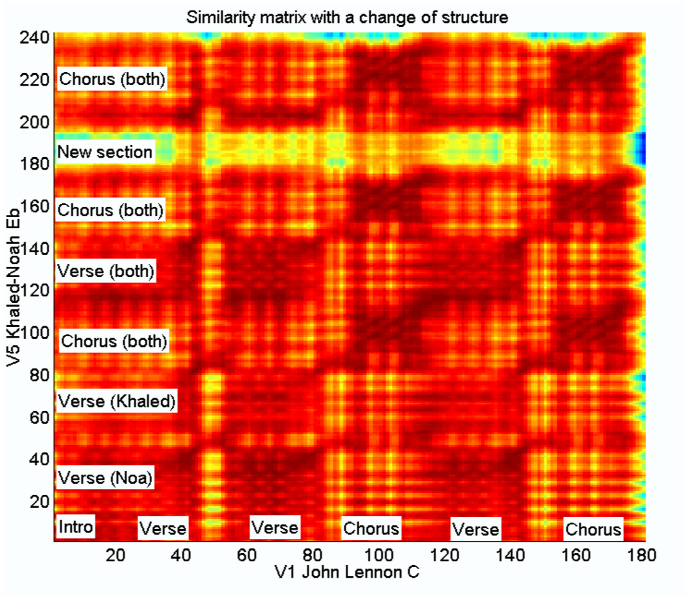

Figure 4. Similarity matrix between version 5 and the original version of *Imagine*.

cepstral coefficients (MFCCs), representing low-level timbre features. We extend this approach to the mentioned low-level tonal features. Figure 5 (at the top and left side) represents the self-similarity matrix for the original version of *Imagine*, using instantaneous THPCP. The similarity matrix is obtained using distance between THPCP profiles statistics over a sliding window.

In this self-similarity matrix we can identify the structure of the piece by locating side diagonals (*verse-verse-chorus-verse-chorus*). We also observe that there is a chord sequence which is repeating along the verse (C-F), so that there is a high self-similarity inside each verse. Instead of computing a self-similarity matrix, we compute now the similarity matrix between two different pieces. Figure 5 shows the similarity matrix between the original song (1) and the instrumental version (2).

In this figure, we also identify the same song structure as before, which is preserved in version 2. We also see that the tempo is preserved, as the diagonal is located so that the time index remains the same in x and y axis. Now, we analyze what happens if the structure is modified. Figure 4 shows the similarity matrix between the original song and version 5. Here, the original overall tempo is more or less kept, but we can identity some modifications in the structure of the piece. With respect to the original song, version 5 introduces a new instrumental section plus an additional chorus at the end of the piece. Figure 5 represents the similarity matrix for each of the 5 cover versions and the self-similarity matrix of the original song. We can see that version 4 (Tania Maria) is the most dissimilar one, so that we can not distinguish clearly a diagonal in the similarity matrix. If we listen to both pieces, we can hear some changes in harmony (jazz), as well as changes in the main melody. These changes affect the THPCP features. In this situation, it becomes difficult to decide if this is a different piece or a version of the same piece. In Figure 5, we also present the similarity matrix with a different song, *Besame Mucho* by Diana Krall, in order to illustrate that it is not possible to find a diagonal for different pieces if they do not share similar chord progressions. As a conclusion to the example presented here and to the observation of 90 versions of different pieces, we advance the hypothesis that the instantaneous tonal similarity between pieces is represented by diagonals in the similarity matrix from tonal descriptors. The slope of the diagonal represents tempo differences between pieces. In order to track these diagonals, we use a simple Dynamic Time Warping, found in [13]. This algorithm estimates the minimum cost from one piece to the other one using the similarity matrix. We study in next section how this minimum cost can be used to measure similarity between pieces.

4. Evaluation

4.1. Methodology

In this evaluation experiment, we compare the accuracy of four different similarity measures:

1. Correlation of global HPCP, computed as the average of HPCP over the whole musical piece.
2. Correlation of global THPCP, computed by shifting the global HPCP vector with respect to the key of the piece, obtained automatically as explained in [11].
3. Minimum cost computed using DTW and a similarity matrix from HPCP values.
4. Minimum cost computed using DTW and a similarity matrix from THPCP values.

The estimation accuracy is measured using average precision and recall for all songs in the database. For each one, the query is removed from the database, i.e. it does not appear in the result list. In order to establish a baseline, we compute the precision that would be obtained by randomly selecting pieces from the music collection. Let's consider that, given a query i from the collection ($i = 1 \dots N$), we randomly chose a given piece $j \neq i$ ($j = 1 \dots N$) from the evaluation collection as most similar to a query. The probability of choosing a piece with the same version Id is equal then to:

$$RandomPrecision_i = \frac{nId(i) - 1}{N - 1} \quad (5)$$

The average for all the possible queries is equal to:

$$RandomPrecision = \frac{1}{N} \cdot \sum_{i=1}^{N} RandomPrecision_i \quad (6)$$

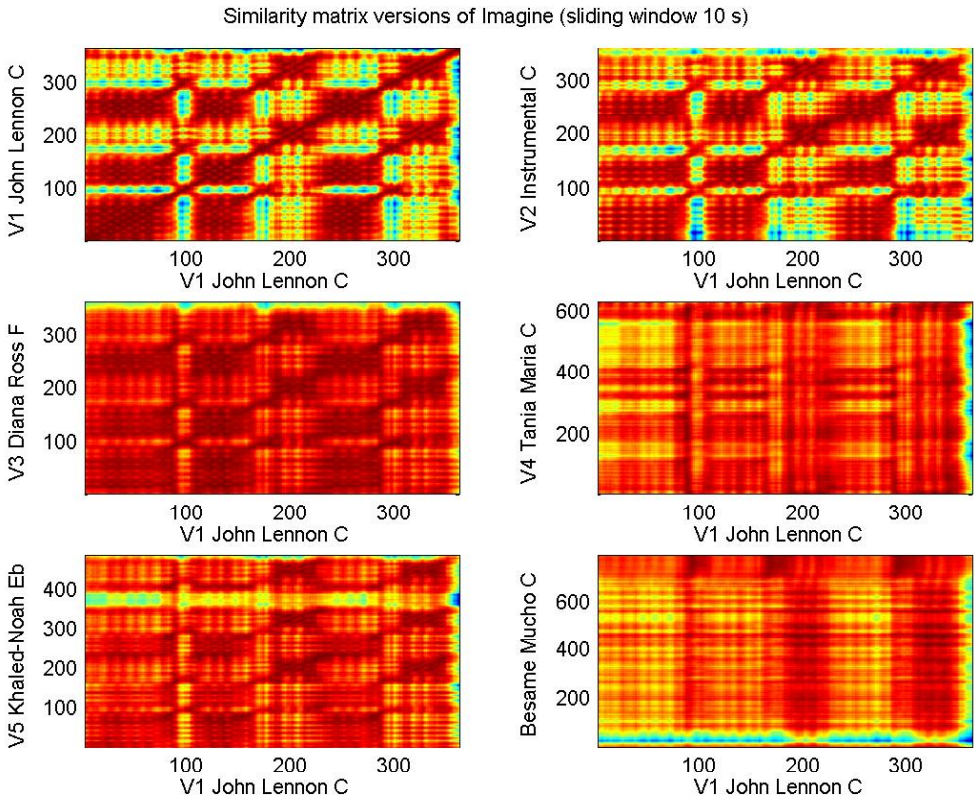

Figure 5. Similarity matrix for 5 different versions of *Imagine*.

For the considered evaluation collection, the baseline would be $RandomPrecision = 3.196\%$, with a maximum value of the F measure equal to 0.0619. This is a very low value that our proposed approach should improve.

4.2. Material

The material used in this evaluation are 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, transpositions and modifications of main melody and harmonization. The average number of versions for each song is equal to 3.07, and its variance is 2.71. Most of the versions include modifications in tempo, instrumentation, key and structure, and some of them include variations in harmonization [2] . 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.

4.3. Results

Figure 6 shows the average precision and recall for all the evaluated collection for the different configurations. When using the correlation of global average HPCP as a similarity measure between pieces, the obtained precision is very low, 20% with a recall level of 8% and a F measure of 0.145. When using global features normalized with respect to the key (THPCP), the precision increases to 35.56%, around 15% higher than using HPCP. The recall level also increases from 8% to 17.6%, and the F measure to 0.322. Using instantaneous HPCP and DTW minimum cost, the precision is equal to 23.35%, which is higher than using a global measure of HPCP. The recall level is slightly higher, equal to 10.37% and the F value is equal to 0.159. Finally, if we use DTW minimum cost computed from instantaneous THPCP as similarity measure, we observe that the maximum precision increases up to 54.5%, and the recall level is equal to 30.8%, obtaining a F measure of 0.393. This evaluation shows that relative descriptors (THPCP) seem to perform better than absolute chroma 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, which is sometimes neglected. The best accuracy is then obtained when using a simple DTW minimum cost computed from THPCP descriptors, and it is around 55% precision (recall level of 30%, F measure equal to 0.393).

[2] The list of songs in the music collection and some additional material to this work is presented in http://www.iua.upf.edu/~egomez/versionid

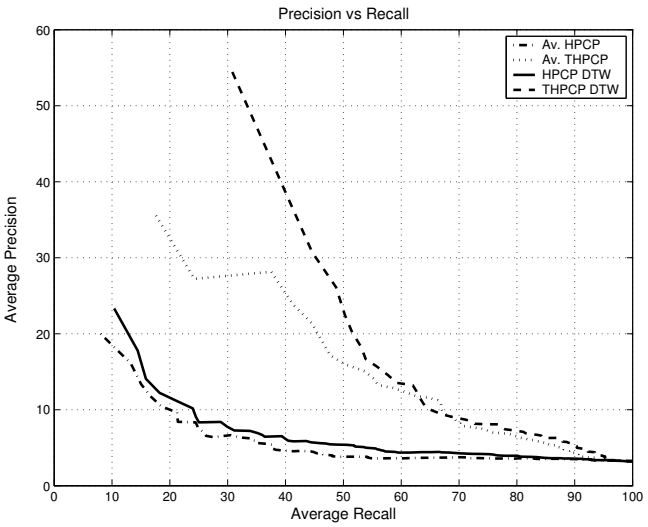

Figure 6. Precision vs recall values for the different configurations.

5. Conclusions and future work

We have focused in this paper on the analysis of tonal similarity and its application to the identification of different versions of the same piece. We have presented a small experiment showing that tonal descriptors by itself can be helpful for this task.

There are some conclusions to this study. First, it is necessary to consider invariance to transposition when computing tonal descriptors for similarity tasks. Second, we should look at the structure of the piece to yield relevant results. Looking at the tonal structure of the piece yields very good results that may probably exceed those attainable using other types of descriptors (i.e. timbre or rhythm).

Version identification is a difficult problem requiring a multifaceted and multilevel description. As we mentioned before, our evaluation database represents a real situation of a database including cover versions, where even the harmony and the main melody is modified. This fact affects the pitch class distribution descriptors. Even in this situation, we see that only using low-level tonal descriptors and a very simple similarity measure, we can detect until 55% of the versions with a recall level of 30% (F measure of 0.393). These results overcome the baseline (F measure of 0.0619) and show that tonal descriptors are relevant for music similarity.

Further experiments will be devoted to include higher level structural analysis (determining the most representative segments), to improve the similarity measure, and to include other relevant aspects as rhythmic description (extracting characteristics rhythmic patterns) and predominant melody estimation.

6. Acknowledgments

This research has been partially supported by EU-FP6-IST-507142 project SIMAC [3] and e-Content HARMOS [4] project, funded by the European Comission. The authors would like to thank Anssi Klapuri, Flavio Lazzareto and people from MTG rooms 316-324 for their help and suggestions.

References

[1] Elias Pampalk. A matlab toolbox to compute music similarity from audio. In *ISMIR*, Barcelona, Spain, 2004.

[2] Jean-Julien Aucouturier and François Pachet. Tools and architecture for the evaluation of similarity measures: case study of timbre similarity. In *ISMIR*, Barcelona, Spain, 2004.

[3] Jonathan T. Foote, Matthew Cooper, and Unjung Nam. Audio retrieval by rhythmic similarity. In *ISMIR*, Paris, France, 2002.

[4] Fabio Vignoli and Steffen Pauws. A music retrieval system based on user-driven similarity and its evaluation. In *ISMIR*, London, UK, 2005.

[5] Beth Logan and Ariel Salomon. A music similarity function based on signal analysis. In *International Conference on Multimedia and Expo*, Tokyo, Japan, 2001.

[6] Elias Pampalk, Simon Dixon, and Gerhard Widmer. On the evaluation of perceptual similarity measures for music. In *International Conference on Digital Audio Effects*, London, UK, 2003.

[7] Adam Berenzweig, Beth Logan, Daniel P.W. Ellis, and Brian Whitman. A large-scale evalutation of acoustic and subjective music similarity measures. In *International Conference on Music Information Retrieval*, Baltimore, USA, 2003.

[8] W. Jay Dowling. Scale and contour: two components of a theory of memory for melodies. *Psychological Review*, 85(4):341–354, 1978.

[9] Cheng Yang. Music database retrieval based on spectral similarity. In *ISMIR*, 2001.

[10] Hendrik Purwins, Benjamin Blankertz, and Klaus Obermayer. A new method for tracking modulations in tonal music in audio data format. *Neural Networks - IJCNN, IEEE Computer Society*, 6:270–275, 2000.

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

[12] Emilia Gómez. *Tonal description of music audio signals*. Phd dissertation, Universitat Pompeu Fabra, July 2006. http://www.iua.upf.es/~egomez/thesis.

[13] Dan Ellis. Dynamic Time Warp (DTW) in Matlab. Online resource, last accessed on May 2006. http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw.

[14] Wei Chai. *Automated analysis of musical structure*. Phd thesis, MIT, August 2005.

[15] Beesuan Ong and Perfecto Herrera. Semantic segmentation of music audio contents. In *ICMC*, Barcelona, 2005.

[16] Jonathan T. Foote. Visualizing music and audio using self-similarity. In *ACM Multimedia*, pages 77–84, Orlando, Florida, USA, 1999.

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

[4] http://www.harmosproject.com