MULTIMODAL SIMILARITY BETWEEN MUSICAL STREAMS FOR COVER VERSION DETECTION

Rémi Foucard, Jean-Louis Durrieu, Mathieu Lagrange, and Gaël Richard *

Institut TELECOM, TELECOM ParisTech, CNRS-LTCI
37, rue Dareau
75014 Paris, France
remi.foucard@telecom-paristech.fr

ABSTRACT

Expressing the similarity between musical streams is a challenging task as it involves the understanding of many factors which are most often blended into one information channel: the audio stream. Consequently, separating the musical audio stream into its main melody and its accompaniment may prove as being useful to root the similarity computation on a more robust and expressive representation.

In this paper, we show that considering the mixture, an estimation of its main melody and its accompaniment as modalities allows us to propose new ways of defining the similarity between musical streams. In the context of the detection of cover version, we show that highest performance is achieved by jointly considering the mixture and the estimated accompaniment. As demonstrated by the experiments carried out using two different evaluation databases, this scheme allows the scoring system to focus more on the chord progression by considering the accompaniment while being robust to the potential separation errors by also considering the mixture.

Index Terms— Cover Song Identification, Music Similarity, Main melody extraction, Signal Processing, Music Information Retrieval

1. INTRODUCTION

Music similarity is receiving a continuously growing interest due to the number of potential applications that can be derived in the field of Music Information Retrieval (MIR). Indeed, music similarity is essential to provide users the ability to search large music databases using high level semantic descriptors. Music similarity has however many facets and may refer to a large variety of common attributes between two songs such as genre, tonality, chord progression, rhythmic structure or timbral content.

Cover version detection, which consists in retrieving different interpretations or recordings of a pre-recorded music piece, is a specific aspect of music similarity. It addresses different applications such as musicology or copyright control but also provides users an efficient tool to build personal collections of cover songs which is quite popular amongst jazz and rock fans. In addition, the problem of cover version identification is well defined with a clear ground-truth annotation which allows for objective evaluation. As mentioned in [1], two cover versions can differ in a number of musical dimensions including timbre, tempo, tonality, rhythm or lyrics language. But it is commonly agreed that tonal (or chord) sequence and melody are, in most cases, largely preserved between two cover songs. Not surprisingly, previous studies in cover song detection are either based on the similarity of the harmonic progression (or tonal sequence) or on the similarity of the melody. In most studies [1] [2] [3], the tonal sequence is described using chromagrams, which represent the instantaneous distribution of the spectral energy of the signal across a predefined number of intervals within one octave (a choice of 12 intervals corresponds to one interval per semitone). The approaches based on melodic similarity are often tackled directly in the symbolic domain, e.g. by comparing the MIDI-like representation of the melodies obtained by predominant melody extraction algorithms (see for example [4] [5] for sub-melody comparisons or [6] for the entire melody). These latter approaches have evident links with query-by-humming systems [7].

At a first glance, it seems reasonable to combine these different approaches into a single multimodal system. However, the different fusion schemes that jointly consider the melody and the accompaniment did not prove effective in our experiments. Many reasons can be stated to explain this fact, from the quality of the separation to the heterogeneity of the representations involved in respectively describing the melody and the accompaniment. Nevertheless, we show that the use of a leading voice separation algorithm [8] can help identifying an invariant aspect among the two separate components as well as the mixture. Promising results are obtained and it is, in particular, shown that early and late fusion strategies that combine information from the mixture and the accompaniment lead to enhanced performances on a public database (e.g. Covers80) compared to the reference method [1] as well as on a larger dataset.

The paper is organized as follows: the proposed method is detailed in section 2. Our approach is rooted on the separation of the original song (the mixture) in two components: the melody and the accompaniment using a leading voice separation method that is described first. We next present the pair-wise similarity computation method that is considered to compute the similarity between two elements of the same modality (melody, accompaniment or mixture). Lastly, the different fusion schemes are detailed and evaluated. Those experiments as well as the results are summarized in section 3 and some conclusions are suggested in section 4.

2. PROPOSED METHOD

2.1. System outline

Our work is focused on the problem of cover version detection. This problem is usually stated as an audio similarity task: given a query

*This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation

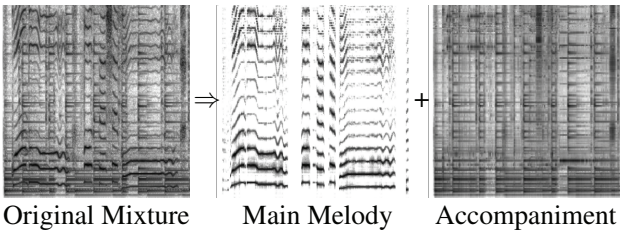

Fig. 1. Example of a separation of a given song into its main melody and its accompaniment displayed as spectrograms.

song and a song collection, we would like to know which song from the reference collection is the most similar to the query song. The similarity in this application is defined in a binary way: two songs that are cover versions or original versions of one same "root" song are similar. They are not considered similar otherwise. In practice, we first compare the query song with all the entries of the reference collection. The algorithm then returns the list of songs ranked by decreasing similarity with the query. The evaluation metrics are described in section 3.

In the approach proposed in this paper, each song is decomposed into 3 different "modalities": its original mixture, the main melody and accompaniment as estimated using [8]. Each modality is then compared using the method proposed in [1] which obtained the best results at the international evaluation MIREX 2007 campaign. The different fusion strategies used to combine matching information issued from the different modalities are next described.

2.2. Leading Melody/Accompaniment Separation

We assume that a cover song and its original root song share at least the same melody line or the same chord progression. However it is not necessary that they share both cues at the same time. We therefore propose to separately process these two elements of the songs.

To execute separated analysis on the melody and the accompaniment, we first use a main melody extraction technique [8]. Each song is assumed to be the mixture of two contributions: the leading voice, usually a singer, which is modeled by a source/filter model, and the accompaniment, which is modeled using Non-negative Matrix Factorization (NMF) formalism. The leading voice is assumed to be harmonic and monophonic. The separation system mainly tracks the leading voice following two cues: first its energy, and second the smoothness of the melody line. Therefore, the resulting separated leading voice is usually the instrument that is the most salient in the mixture, over certain durations of the signal, see Figure 1.

For each song of our collection, this technique provides us with two separated signals, one for the melody and one for the accompaniment. Since they ideally represent two distinct aspects of the songs, these signals can be analyzed, either separately or jointly, in order to compare two songs.

2.3. Pair-Wise Similarity Computation

The similarity between two songs using a given modality is based on [1]. Its aim is to provide a similarity measure that takes into account potential transposition and which gives the best alignment score over all the subsequences of two songs: A (the "reference") and B (the "query"). As in many previous works on cover version detection [2, 3, 4], the features chosen to represent each song are pitch class distribution features. In [1], the authors chose the harmonic pitch class profile (HPCP), proposed in [9]. First, a sequence of I-bin HPCP is constructed from both songs A and B: $\{\mathbf{h}_{A,n}, n \in [1,N]\}$ and $\{\mathbf{h}_{B,m}, m \in [1,M]\}$, where N and M represent the number of analysis frames in each song. For our study, we set $I = 36$, as suggested in [1].

The chroma sequence of the query is then transposed into the key of the reference. A global HPCP is computed as the average over all the frames for each song: $\mathbf{h}_A$ and $\mathbf{h}_B$. Then the *Optimal Transposition Index* (OTI) is computed as follows:

$$OTI\left(\mathbf{h}_A, \mathbf{h}_B\right) = \underset{0 \leq i \leq I-1}{\operatorname{argmax}} \left\{\mathbf{h}_A \cdot \operatorname{circshift}(\mathbf{h}_B, i)\right\} \qquad (1)$$

where "$\cdot$" indicates a dot product and $\operatorname{circshift}(\mathbf{h}, n)$ is a function rotating vector $\mathbf{h}$, n bins "downwards". Once the OTI is calculated, the sequence $\mathbf{h}_{B,1:M}$ can be transposed into the key of song A: $\forall n, \mathbf{h}^{Tr}_{B,m} \leftarrow \operatorname{circshift}(\mathbf{h}_{B,m}, OTI)$.

The following step consists in building a similarity matrix $\mathbf{S}$ between the two sequences. This matrix is binary: for frame n of song A and frame m of song B, if $OTI(\mathbf{h}_{A,n}, \mathbf{h}^{Tr}_{B,m}) \in \{0, 1, 35\}$), then the two instants are considered "*similar*", $S_{n,m} = \mu_+$. Otherwise, the two instants are considered "*not similar*", $S_{n,m} = \mu_-$. In this work, we kept the same values as in [1]: $\mu_+ = +1$ and $\mu_- = -0.9$.

From $\mathbf{S}$, an alignment matrix $\mathbf{H}$ is then computed. The alignment method, called the Dynamic Programming Local Alignment (DPLA), is inspired by Dynamic Time Warping [11] but designed to detect and align similar subsequences of the two compared signals. Each value $H_{n,m}$ of the alignment matrix represents the optimal cumulative similarity of two subsequences ending at frame n for song A and frame m for song B. These scores increase along similar subsequences, and decrease otherwise.

Finally, the similarity score between A and B is set to the maximum value of the alignment matrix. Indeed, the values of this matrix represent cumulative similarity, so we can consider that the maximum value shows the best aligned subsequences.

2.4. Fusion Schemes

While in [1], the authors propose a method using only the mixture of each song, we propose to use the three different separated signals: the mixture ("Mix."), the melody ("Mel.") and the accompaniment ("Acc."). Using these three modalities, we desire a single similarity score. We have evaluated several strategies by multiplexing the output of the separation process for both late and early fusion schemes, namely merging similarity scores or merging similarity matrices (see Figure 2).

Late Fusion: The most straight-forward way to merge all the analysis is by computing similarity scores on each of the modality, and then, merging them. There are many ways to execute the late fusion: e.g. taking the maximum or minimum of the intermediate scores. We will show in section 3.2 that the maximum appears to be the best operator for our application. One can also choose which modality to include in the decision. We have tested 4 configurations: considering only 2 modalities, {Mix., Acc.}, {Acc., Mel.}, {Mix., Mel.} and all of them, {Mix., Acc., Mel.}.

Early Fusion: Late fusion is very simple to set up, but its drawback is that the multimodal information is taken into account only at the end of the process. As a matter of fact, late fusion boils down to analyzing separately the considered components of the song, and then returning a score considering the results of these separate analysis.

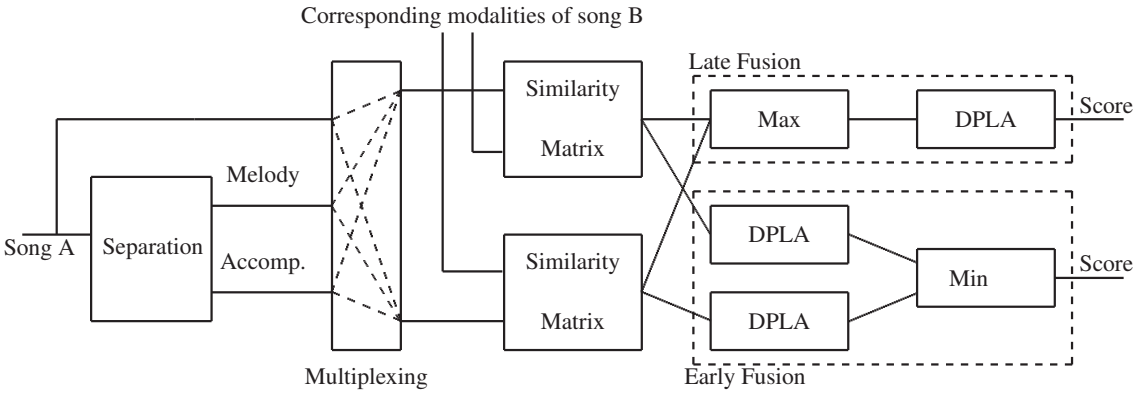

Fig. 2. Block-diagram of the proposed similarity computation between two songs A and B. After separation, the different modalities are combined using two different fusion schemes: at an early stage (top) and at a late stage (bottom).

In order to use the multimodal information provided by the source separation earlier in the process, we also considered merging similarity matrices. Thus, the similar subsequences located by the DPLA could be more relevant.

We first compute similarity matrices corresponding to the mixture signal, the separated melody and the separated accompaniment: $\mathbf{S}^{\text{mix}}$, $\mathbf{S}^{\text{mel}}$ and $\mathbf{S}^{\text{acc}}$. Taking the maximum value of all 3 similarity matrices yielded the best results. The entries of the final similarity matrix thus verify:

$$S_{n,m} = \max(S^{\text{mix}}_{n,m}, S^{\text{mel}}_{n,m}, S^{\text{acc}}_{n,m}). \quad (2)$$

The alignment is then achieved through a DPLA procedure, see Figure 2.

3. EXPERIMENTS AND RESULTS

3.1. Evaluation framework

We developed our system using the Covers80 database[1] [12], which is made up of 80 original pop songs, with about one cover for each (which amounts to 164 songs). The songs are sampled at 16 kHz in mono, with 16 bits samples. For our experiments, we successively took every song of the set as a request, and then we ordered the list of the remaining songs by similarity with this request. When the request song was an original song, we wanted this request to return the cover in first position. And when the request was a cover song, we searched for the original one.

As the Covers80 database only provided us one cover per original song, the precision, recall and F-measure values (recommended by [13]) did not seem sufficient to represent the relevance of a returned list of songs. Indeed, if we just look for one particular song, a request should only return the most similar song. In that case, precision, recall and F-measure are equivalent; their value is 1 if the returned song is the searched one, and 0 otherwise. Thus, in addition to precision, we observed the rank of the first (and only) actual cover in the returned list. Obtaining $\text{rank}_1 = n$ means that the first cover in the list is ranked n^{th}. For all the requests, we watched the mean precision, as well as the mean and median rank of the relevant song.

For the sake of completeness, we also evaluated the proposed approaches using another cover database developed at Telecom-ParisTech. This database is composed of 20 root songs (mainly pop, rock and jazz), and approximately 20 covers per root song. For this database, the evaluation metrics are those used in [1].

[1] http://labrosa.ee.columbia.edu/projects/coversongs/covers80/

3.2. Results

	Precision	Mean rank_1	Median rank_1
Mix.	**0.35**	**32.02**	**7**
Acc.	0.34	33.18	**7**
Mel.	0.21	43.98	21
Mix., Acc.	**0.37**	31.41	**5**
Acc., Mel.	0.34	32.28	6.5
Mix., Mel.	0.35	**31.18**	7
Mix., Acc., Mel.	**0.37**	31.24	**5**
Early fusion	0.34	**30.49**	5.5

Table 1. Mean precision, mean and median rank of the cover for the three modalities, several late fusion schemes and one early fusion scheme on the Covers80 database.

First, let us study the performances of the DPLA presented separately for the three components. No fusion was realized for these preliminary experiments. The results are presented on top of Table 1, where the best result in each column is indicated in bold. A strong difference is clearly set forth between Mean rank_1 and Median rank_1, which can be interpreted as follows: many results are very good, but the queries that provide bad results, although less frequent, often provide very bad results.

We can also notice that the DPLA analysis, performed on the accompaniments, presents roughly the same performances as the state of the art. However, the DPLA does not seem appropriate for being applied on melodies only.

When considering several modalities at the same time, late fusion yields best results by taking the maximum of the considered scores. This can be explained by the fact that analyzing several components increases the chance to discover the least varying aspect between an original song and its cover. Thus, keeping the maximum similarity obtained from several modalities appears to be the better approach. The results are presented at the bottom of Table 1. We can see that every combination presents at least one measure outperforming the simple analysis. But the combinations involving the solo analysis appear to suffer from the poor results obtained by DPLA when applied to this component. Furthermore, we observed

	F-measure	Precision	Recall	Mean rank$_1$	Median rank$_1$
Mix.	0.20	0.20	0.23	9.35	**2**
Acc.	0.20	0.19	0.22	8.70	**2**
Solo.	0.02	0.16	0.01	15.93	8
Late fusion	**0.22**	**0.21**	**0.24**	**8.02**	**2**

Table 2. Mean F-measure, mean and median rank of the cover for several systems performing on the 362 songs of the Telecom-ParisTech database.

that the performance gain is negligible when adding information on solos to the fused analysis on original mixtures and accompaniments (see the small difference between the first and last rows). As we have done for early fusion, similarity matrices of different components are merged with several operators. We have seen in section 2.4 that the best results are obtained by considering the fusion scheme of Eq. 2.

The gain of performance brought by multimodal analysis is actual, since almost every measure of the multimodal systems is better than DPLA alone (except the mean precision obtained by early fusion). The same conclusion can be drawn by the results obtained with the Telecom-ParisTech database, see Table 2. The discrepancy in terms of representation between the musical information carried by the melody and the accompaniment seems to be the determinant factor for the results achieved by the evaluated systems. However promising, the estimated main melody did not prove as informative. Several reasons can be stated. First, the separation is far from being perfect and in particular, melodic components can be extracted even though the main instrument is not active. Secondly, representing the main melody as a series of chroma vectors is convenient as far as early fusion scheme is concerned, but it may not be the most effective option. We tried comparing solos using other algorithms [4]. Unfortunately, none of them integrated well in the proposed system.

On the contrary, jointly considering the accompaniment and the original mixture, and representing both of them as a series of chroma vectors appears consistent. Indeed, both of them focus on the chord progression. This progression is enhanced in the accompaniment modality when the separation is successful and still captured in the original mixture when the separation induces too many artifacts.

4. CONCLUSION

Considering different modalities of a musical stream as its main melody and accompaniment is a promising research direction in order to offer new ways of comparing musical streams.

We have seen that a state-of-the-art source separation algorithm can provide useful information for enhancing the performances of a cover song identification system. However, the study presented in this paper also showed the intrinsic limitation of simple fusion schemes whose capabilities seem to be limited to merging modalities that carry more or less the same type of information.

Future work will first focus on new algorithms of melody matching that are adapted to our purpose, e.g. [5], [6] or [7]. Secondly, we will have to develop merging algorithms that are specific to musical streams. Indeed, those streams usually exhibit at high level of time redundancy, for example the main melody is very often repeated many times. Taking those specificities into account could lead to an increase of robustness and expressivity which are both determinant factors when defining similarities between music streams.

5. REFERENCES

[1] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1138–1151, Aug. 2008.

[2] E. Gómez, B. S. Ong, and P. Herrera, "Automatic tonal analysis from music summaries for version identification," in *Proc. Audio Engineering Society Convention (AES)*, Oct. 2006.

[3] D. P. W. Ellis and G. E. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," in *Proc. IEEE International Conference on Acoustic, Speech, Signal Processing (ICASSP)*, Apr. 2007, vol. 4, pp. 1429–1432.

[4] M. Marolt, "A mid-level melody-based representation for calculating audio similarity," in *Proc. International Conference for Music Information Retrieval (ISMIR)*, 2006, pp. 280–285.

[5] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R. Van Oostrum, "Using transportation distances for measuring melodic similarity," in *Proc. International Conference for Music Information Retrieval (ISMIR)*, 2003, pp. 107–114.

[6] W-H. Tsai, H-M. Yu, and H-M. Wang, "A query-by-example technique for retrieving cover versions of popular songs with similar melodies," in *Proc. International Conference for Music Information Retrieval (ISMIR)*, 2005, pp. 183–190.

[7] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the musart testbed," *Journal of the American Society of Information Science Technololgy*, vol. 58, no. 5, pp. 687–701, 2007.

[8] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *Proc. IEEE International Conference Acoustic, Speech, Signal Processing (ICASSP)*, 2009.

[9] E. Gómez, *Tonal description of music audio signals*, Ph.D. thesis, Music Technol. Group, Univ. Pompeu Fabra, Barcelona, Spain, 2006, Available: http://www.iua.upf.es/˜egomez/thesis/.

[10] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proc. International Computer Music Conference (ICMC)*, 1999, pp. 464–467.

[11] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall PTR, April 1993.

[12] Daniel P. W. Ellis and Courtenay V. Cotton, "The 2007 labrosa cover song detection system," 2007, Available: http://www.ee.columbia.edu/˜dpwe/pubs/EllisC07-covers.pdf.

[13] J. Serrà, "A qualitative assessment of measures for the evaluation of a cover song identification system," in *Proc. International Conference for Music Information Retrieval (ISMIR)*, Viena, Austria, 2007, pp. 319–322.