Semantic-based Similiarity of Music

Michael Rentzsch and Frank Seifert

Chemnitz University of Technology, Dept. of Computer Science
09107 Chemnitz, Germany

Abstract. Existing approaches to music identification such as audio fingerprinting are generally data-driven and based on statistical information. They require a particular pattern for each individual instance of the same song. Hence, these approaches are not capable of dealing with the vast amount of music that is composed via methods of improvisation and variation. Futhermore, they are unable to measure the similarity of two pieces of music. This paper presents a different, semantic-based view on the identification and structuring of symbolic music patterns. This new method will allow us to detect different instances of the same song and acquire their degree of similarity.

1 Introduction

"A new system combining a mobile phone and a cryptographic technique will soon allow you to get the name of any music track you hear on the radio [. . .]"[1]

This idea, developed by researchers of consumer electronics giant Philips in 2001, is reality today and a service offered by many mobile phone companies. But audio identification is more than discovering the name and interpreter of a song you heard. Music (Information) Retrieval is a young and active research topic which deals with many ways to analyse and process music—symbolic or sub-symbolic—automatically.

A lot of efforts have been put in the development and improvement of Query-by-humming ([5], [10]) and audio fingerprinting systems ([1], [3]). As a consequence, they work very well for identifying music from a limited repository already. Unfortunately, they require an individual pattern for each piece that shall be identifiable. Hence, it is not possible, to use them for analysing the countless number of audio files containing live performed, covered and improvised music. Furthermore, they lack the ability to determine a fine-grained metric for the similarity of two audio documents. This insufficiency is caused by the lack of knowledge about the semantics of music that these systems show.

To develop an audio identification system that is capable of processing the vast amount of music which is created through the process of improvising and copying existing pieces, one has to give concern to human music perception: Listening to a melody, we create inner hypotheses according to our experiences related to this music theme. Our intention is to develop a system based on a semantic-based model of music identification. This model—the lead sheet model—includes the required information to simulate the process of creating and validating hypotheses.

[1] NewScientist.com, 2001-11-28

Rentzsch M. and Seifert F. (2006).
Semantic-based Similiarity of Music.
In *6th International Workshop on Pattern Recognition in Information Systems*, pages 175-180
DOI: 10.5220/0002485101750180

2 The Lead Sheet Model

The lead sheet model (LSM) has been introduced in [7]. It is designed to be a generic conceptual model for semantic-based identification and comparison of music. The name originates from a simplified form of sheet music. A lead sheet contains elementary parameters of music, i.e. melody, rhythm, chords and tempo. It is widely used for jazz and songbooks.

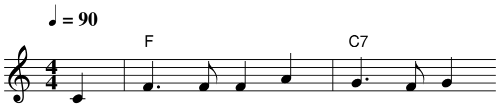

Fig. 1. Characteristic motif of the Scottish folk song "Auld Lang Syne".

Figure 1 shows the basic element of the lead sheet model: a *characteristic motif* (CM). Analysing an audio document, such motifs are the fundamental components that have to be identified. A similar approach has been published in [4]. A combination of two or more motifs that have a strong semantic connection is called a *lead sheet* (LS). Lead sheets add a structural aspect—which [4] does not consider—to the LSM. They cannot consist of characteristic motifs only but also of other (sub) lead sheets (see Figure 2). Each link from a parent LS to its sub elements is attributed with a pair of information: a number marking the temporal order of the child elements and a bit string specifying the *relations* between the two nodes. Among others, valid relations are: motif, theme, verse, refrain, intro, and many more.

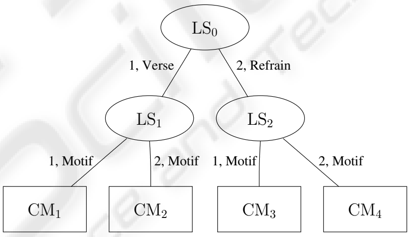

Fig. 2. Simple lead sheet graph.

The process of identifying music using the LSM can be split: Firstly, the musical parameters have to be mapped to characteristic motifs. This results in a set of temporally ordered motif instances. Secondly, each detected instance has to be correlated with a parent lead sheet, if possible. The second step might be repeated until there are no abstractions left. Using the lead sheet model, it is possible to compare and identify music in whatever instantiation it is played. Furthermore, it is not only possible to recognise simple patterns of music but also combinations of motifs and complex structures. While [7] showed that it is possible to use the LSM on symbolic documents, [8] successfully applied the above mentioned mapping process to audio data.

3 Inexact Matching of Motifs

Section 2 introduced characteristic motifs as the lead sheet model's basic elements. Thus, it is essential to develop functions that identify such motifs in audio documents. Considering music performed live and improvised, these functions have to be robust against small modifications—intended as well as unintended—of a motif. This paper presents a method for inexact matching of motifs in symbolic representation of music.

Existing approaches to identification of symbolic music are predominantly based on a representation and comparison of the shape or contour of a melody. Basically, there are three different methods for coding the shape of a melody ([2], [11]):

1. a sequence of exact intervals (semitones), e.g. +7, +3, -2, 0, -12,
2. 3-level representation, e.g. **u**p, **u**p, **d**own, **e**qual, **d**own or
3. higher-level representation, e.g. **U**p, **u**p, **d**own, **e**qual, **D**own.

Identification methods for such representations of music, e.g. calculcating the string edit distance, are mainly statistics-based and lack any musical background. They cannot determine, how *similar* two motifs are concerning the listener's overall impression. Thus, Rentzsch ([6]) introduces a different operation to compare a characteristic motif and a sequence of tones (melody) and to calculate a concrete degree of equality. This operation is based on the examination of five different features—parameters—of music: melody, rhythm, tempo, harmony and voice. Each of these features is processed separately. Thus, five individual equality values E_p are calculated.

The equality values for melody and rhythm E_M and E_R are determined using a string representation of both features. Between these representations of the motifs, the Levenshtein distance is calculated and considered in relation to the number of tones. The bigger the Levenshtein distance is, the smaller is E_M and E_R respectively. The equality value for tempo E_T is determined regarding the lengths of the first tones in the motifs that have be compared. The bigger the relation of the longer one to the shorter one of these tones is, the smaller is E_T.

To compute E_H, harmonies are compared for each tone in the considered motifs. Each comparison returns $h_i \in \{0, 0.5, 1.0\}$, depending on whether the harmonies are unequal, similar[2] or equal. Then, the overall equality value for harmony E_H is the average of all h_i. The last parameter—the value for voice—E_V is determined considering the voice that the motif appears in. If the motif is settled in the main voice, E_V is set to 1. In all other cases, E_V is 0.

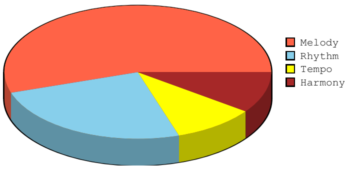

Fig. 3. Weighting of E_p (W_p).

[2] e.g. C and C7

Each comparison of the musical features resulted in an equality value E_p with $E_p = 1$ if motif and melody are completely equal. Yet, it is useful to develop a procedure to combine those values to one final degree of equality E. As some of the individual values are more important for the identification of music than others, not all E_p should have the same impact. Thus, Rentzsch defined a set of weighting values W_p as shown in Figure 3. The equality value for voice E_V is supposed to have a special impact on the final value. Thus, it affects the resulting formula as a “global” factor. Given the five equality values E_p and a set of weightings W_p, the overall degree of equality E can be calculated $E = E_V \sum_p W_p E_p$ width $p \in \{M, R, T, H\}$.

4 Comparing Template Structures

In this part of the paper, we will introduce an approach to comparing entire pieces of music based on the set of characteristic motifs they contain. Having identified all motifs—that are known to the system—in a symbolic document and considering the structure defined in the lead sheet model, every piece of music can be represented as a graph as shown in Figure 4. This graph is called *document template*. All edges are attributed with the relation—specified in the lead sheet graph—and the equality value e_i by which this CM had been identified.

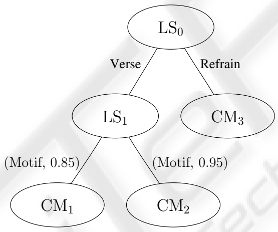

Fig. 4. Document template.

Now, the attempt to compare two pieces can be reduced to comparing these templates. Thus, it is possible to take advantage of existing approaches to graph matching. However, using one of these mainly syntax-based matching methods does not incorporate the semantic functions the modifications hold. Futhermore, the time frame covered by the template—thus, by the identified motifs—has no influence on the resulting distance. Hence, a different approach—a *semantic-based* metric—has been developed ([6]). This metric compares document templates in four levels: time level, structural level, semantic level, and motif level. Each comparison delivers a distance value D_i. The resulting overall distance is a tuple $D = (D_{\text{Time}}, D_{\text{Struct}}, D_{\text{Semant}}, D_{\text{Motif}})$.

As a first step, the common sub-template (CST), i.e. all lead sheets and motifs occurring in both pieces, is determined. On the time level, the time frame that this subtemplate covers in both documents is taken into account. The distance on time level D_{Time} is defined as the average value of the two time frames. The distance on structural level D_{Struct} is used to consider small modifications, e.g. the repetition of a motif. Each modification is rated depending on type and relation to the parent lead sheet. Then,

D_{Struct} is the average value of all such ratings. The semantic level determines and evaluates the relations between the elements in the common sub-template. Result is a set of values between 0 (motifs from different lead sheets) to 5 (same song/piece). In many cases, quite a number of different relations is detected. Thus, the distance value D_{Semant} is the average value of all relations weighted by the time frame they cover. The last level analyses the average degree of equality by which all motifs in the CST have been found. Therefore, D_{Motif} is bigger, if many motifs appear unmodified.

As mentioned before, the final distance is a tuple of the four single distance values D_{Time}, D_{Struct}, D_{Semant}, and D_{Motif}. It depends on a concrete application how these values are treated. Hence, it is possible to combine these values to one overall value as it is used for the equality value. However, this would shorten the number of (semantically) different queries that this metric can process.

5 Results and Future Work

In this paper, we have presented a semantic-based approach to comparing symbolic representations of music. This approach uses the lead sheet model as a conceptual model for music identification. Applying this method to two documents determines a tuple of distance values marking the four levels of comparison. Considering this partitioning, it is possible to respond to quite a number of different types of queries: "find pieces containing song s unmodified", "find pieces using different motifs from s_1 and s_2", "find pieces containing the refrain of song s", and many more. To put the methods described in Sections 3 and 4 in practice, we have implemented the prototype application CoSMIc (`Comparing symbolic musical instances`). This application determines the document templates of two MIDI documents and returns the resulting distance values to its user. To show the applicability of our approaches, we want to briefly describe a small experiment and its results.

Table 1. Distances to "Auld Lang Syne".

	Levels			
Piece	*Time*	*Structural*	*Semantic*	*Motif*
Original	0.91	0	5	0.95
Motif changes	0.91	0	5	0.91
Medley	0.35	0	3.1	0.95
Different song	0	0	0	0

We have compiled a set of four songs: "Auld Lang Syne" (original), "Auld Lang Syne" with modified motifs, "Oh, when the saints", and a medley of "Auld Lang Syne" and "Oh, when the saints". Each of these songs has been processed by CoSMIc and compared to the original "Auld Lang Syne". The results of this experiment—thus, the four distance values of each comparison—are presented in table 1. To make understanding these results easier, we have illustrated them as a 3D chart in Figure 5. Thus, we have omitted the values on the structural level, as they are all 0.

Apparently, it is possible to classify the first two songs as instances of "Auld Lang Syne". They are situated in one region of the chart. "Oh, when the saints" reveals no

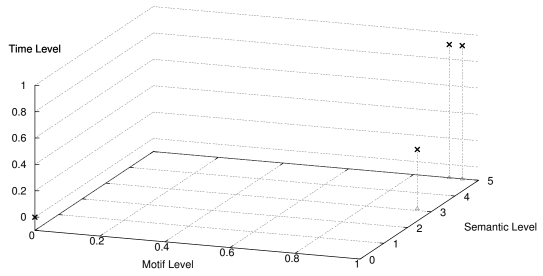

Fig. 5. Distance diagram.

similarities. The medley of different motifs from “Auld Lang Syne”, and “Oh, when the saints” is situated in between these key points. Running a number of experiments using different songs from diverse genres proved the general applicability of the techniques we have introduced. Thus, our semantic-based approach to comparing and clustering different pieces of music which can be variations of the same composition succeeds.

Future work should focus on adopting the methods presented in this paper to subsymbolic audio data using hypotheses-based recognition ([8]). Being able to detect different instances of the same song in whatever representation shall finally provide us with the required techniques to develop a music retrieval system that implements a cluster index-based access to its document repository. Hence, these techniques will assure that diverse queries considering the semantics of music will be processable efficiently.

References

1. Cano, P., Batlle, E., Kalker T., Haitsma J.: A Review of Algorithms for Audio Fingerprinting. In: International Workshop on Multimedia Signal Processing, US Virgin Islands (2002)
2. Dowling, W. J.: Scale and contour: Two components of a theory of memory for melodies. In: Psychological Review, p. 341 - 354 (1978)
3. Haitsma, J., Kalker, T.: A Highly Robust Audio Fingerprinting System. ISMIR, Paris, France (2002)
4. Neve, G., Orio, N.: Indexing and Retrieval of Music Documents trought Pattern Analysis and Data Fusion Techniques. ISMIR, Barcelona, Spain (2004)
5. Nishimura, T. et al.: Music Signal Spotting Retrieval by a Humming Query Using Start Frame Feature Dependent Continuous Dynamic Programming. In: Music Information Retrieval, p. 211 - 218, Bloomington, Indiana (2001)
6. Rentzsch, M.: Entwicklung und Implementierung von Vergleichsoperationen für symbolische Tondokumente, Diploma Thesis. Chemnitz University of Technology, Germany (2005)
7. Seifert, F.: Musikalische Datenbanken, Dissertation. Chemnitz University of Technology, Germany (2004)
8. Seifert, F.: Prediction-Driven Correlation of Audio with Generic Music Templates. EuroIMSA, Grindelwald, Switzerland (2005)
9. Shapiro, L. G., Haralick, R. M.: A metric for comparing relational descriptions. In: IEEE Trans PAMI, p. 90 - 94 (1985)
10. Song, J., Bae, S. Y., Yoon, K.: Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System. ISMIR, Paris, France (2002)
11. Youngmoo, K., Wei C., Ricardo, G., Barry, V.: Analysis of a contour-based representation for melody. Int. Symposium on Music Information Retrieval, Plymouth, Massachusetts (2000)