Content-based Music Access: an approach and its applications

Riccardo Miotto
Department of Information Engineering
University of Padova
Via Gradenigo, 6/B
35131 Padova (Italy)
miottori@dei.unipd.it

Abstract

At current time, the availability of large music repositories poses challenging research problems. Among all, content-based identification is gaining an increasing interest because it can provide new tools for easy access and retrieval. In this paper we describe an ongoing methodology for the content-based identification of unknown music recordings through a collection of music documents. Moreover, as future prospective scenario, identification is viewed in a more general similarity context, where also the perception of the users is considered.

Keywords: music identification, similarity, indexing

1. INTRODUCTION

Nowadays, the availability of large music collections poses challenging problems mainly related to the organization of documents according to some sense of similarity. To address this issue, from a research perspective, an interesting starting point seems to be the automatic identification of music recordings. In this context, given an unknown excerpt of a music work, the task is to retrieve all the recordings of a music collection sharing some content with the query and to deliver relevant metadata such as title, artist and other additional information.

An earlier approach to music identification was *audio fingerprinting* that consists in a content-based signature of a music recording to describe digital music even in presence of noise, distortion, and compression [4]. However, the fingerprint value is strictly related to a particular music performance, but the identification of a music work may be carried out also without linking the process to a particular recording. For example, identification of live performances may not benefit from the fingerprint of other performances, because most of the acoustic parameters may be different. The solution would be to collect all the possible live and cover versions of a music works, which is clearly unfeasible.

Thus, a good music identification approach has to be able to identify music works from the recording of a performance, yet independently from the particular performance, and to sort the elements of the collection according to some similarity measure with the query. Music works to identify may be live performances, cover versions, noisy registrations and so on.

In literature several approaches to establish whether two musical pieces share the same melodic or tonal progression have been proposed, such as in [7], [13] and [5]. Efficient and scalable systems for a cover music identification task were also proposed in [18] and [11].

Content-based identification approaches can have a direct implication in different fields such as musical rights management and licenses, learning about music, discovery new music and many other topics related with music perception and cognition. Furthermore, a general technique can be exploited in different tasks, where identification is just an application among other possible ones. At this purpose, the ongoing methodology that we proposed in [11] aims at being general enough to be useful in other tasks mainly related to the music similarity context.

Thus, in the following, Section 2 provides an overview of the methodology together with the new directions whereas Section 3 describes ideas and prospective approaches toward a more general music similarity system including also users perception.

2. A CONTENT-BASED MUSIC IDENTIFICATION ENGINE

The content-based identification approach described in this paper was firstly proposed in [10] and [11]. The objective is to identify each recordings of a score (including live performances and cover versions) through a collection of indexed high quality recordings. The assumption we made was that, if a performance is played according to the original score, it can be generally modeled and identified through the score information alone.

The system is mainly based on an application of hidden Markov models (HMMs) [15]. Since the identification through HMMs is linear with the size of the collection, an index of the collection have been built to extract from the collection a cluster of candidates to be re-ranked with the HMMs-based identification. Figure 1 provides the general structure of the system; a first prototype is also available on-line at [12].

Figure 1: General Structure for the Music Identification System.

2.1. Clustering the Collection

The music features exploited to extract a cluster of the collection should be both general and robust to all the variations due to differences between the query and the collection recordings (tempo variations, tonality shift, different voices, etc.).

One of the most common music content descriptor are *chroma* features. The basic idea behind chroma is that octaves play a fundamental role in music perception and composition. For instance, the perceived quality – e.g. major, minor, dominant, and so on – of a given chord depends only marginally on the actual octaves where it spans, whereas it is strictly related to the pitch classes of its notes. Following this assumption, a number of identification techniques based on chroma have been proposed, in particular in [13] and [7].

In our approach [11], chroma features are used to index the music collection. In particular, chroma are considered as pointers to the the music documents they belong to, playing the same role of words in textual documents. Each chroma feature points to a number of recordings and to a set of time positions within each recording. One major advantage of indexing in text retrieval is that the list of index terms can be accessed in logarithmic, or even constant, time. The same cannot be applied to feature vectors, because the exact match has to be replaced by a similarity search, which is less efficient. One of the technique to handle efficiently this issue is the locality sensitive hashing (LHS) [6]. Its basic idea is to apply to the feature vectors a carefully chosen hashing function with the aim of creating collisions between vectors which are close in the high dimensional feature space. The hashing function itself becomes then an effective tool to measure the similarity between two vectors.

Following this idea, we propose to represent the 12-dimensional chroma vectors with a single integer value through an hashing function, not depending on the absolute value of the chroma pitch classes, but just on their rank within the vector. In Figure 2 the whole chroma indexing process is depicted.

Figure 2: Chroma Indexing process, from the music recording to the index

Retrieval is carried out using the *bag of words* paradigm, then by counting the common chroma words between the query and the recordings of the collection. A problem that may affect retrieval effectiveness is that chroma-based representation is sensitive to transpositions. In fact, if the query and the matching recordings stored in the database are played in different tonalities, they have a totally different sets of chroma. We addressed this problem by considering that a transposition of s semitones will result in a rotation of the chroma vector of s steps. Each query is then transposed to include all the most common tonalities.

At the end, the top N retrieved recordings are considered as the cluster of potential candidates. Experimental results achieved with a collection of 1000 recordings of classical music showed that a cluster of 100 documents (1/10 of the collection) was sufficient to have a 100% of recall [11]. First evaluation with pop and rock music gave promising results even if, due to the more variety of the music, a more accurate elaboration seems to be necessary to achieve the same precision.

At this aim, experience matured in information retrieval proved that usually the combination of different features outperforms the use of just one. Then, following this assumption we believe that including other music descriptors could increase the performances of the system in terms of both precision and recall.

Considering that cover songs generally preserve not only the harmonic-melodic characteristics of the original work but also its rhythmic profile, the first idea is to combine chroma with some rhythm descriptors, such as the one proposed in [9] for a genre classification task and called *rhythm histogram* (RH). In a RH the magnitudes of each modulation frequency bin for all the critical bands of the human auditory range are summed up to form a histogram of “rhythmic energy” per modulation frequency. Similarity relationships can be measured according to the distance among the histogram representations, and the songs of the collection can be ranked following these values. A weighted final rank is then computed through a weight merge of both lists, where the greater weight is given to the chroma list.

In a similar way, another music features that could be exploited are the Mel-frequency cepstral coefficients (MFCCs). MFCCs are often used to compute music similarity especially in genre classification tasks [16, 3]. They capture the overall spectral shape of the audio signal, which carries important information about the instrumentation and its timbre, the quality of a singer's voice, and post-production effects [2]. However, they do not capture information about melody and rhythm which are the most important features for the identification task we are proposing. Anyway, following one more time the assumption that the combination of different features usually outperforms the use of just one feature, we believe that even MFCCs can be useful to increase the efficacy of the system.

MFCCs can be represented as a sequence of n -dimensional vectors, where the value of n depends on the accuracy required by the system. Thus, they can be modeled and retrieved with the same hashing approach of chroma, and can provide another rank list of potential candidates to be merged with the other rank lists. Alternatively, considering their major application in genre classification tasks, MFCCs rank list may be exploited independently to pre-filter the collection according the music genre (pop, rock, folk, etc.). Then just the subset of the collection extracted can be considered in the chroma-based identification.

2.2. Identification

The cluster extracted from the collection is re-ranked with the HMMs-based identification methodology proposed in [10]. The basic idea of the approach is that, even if two different

performances of the same music work may differ in terms of acoustic features, it is still possible to generalize their music content through a statistical models. To this aim, each recording of the collection has to pass through some modeling steps.

In a first step, a segmentation process extracts audio subsequences that have a coherent acoustic content. The aim of segmentation is to divide the music signal into subparts that are bounded by the presence of music events, where an event occurs whenever the current pattern of a music piece is modified (one or more new notes being played or stopped). Segmentation of the acoustic flow can be considered the process of highlighting audio excerpts with a stable pitch.

Coherent segments of audio are then analyzed through a second step in order to compute a set of acoustic parameters that are general enough to match different performances of the same music work. In line with the segmentation approach, also parameters extraction is based on the idea that pitch information is the most relevant information for a music identification task. Because pitch is related to the presence of peaks in the frequency representation of an audio frame, the parameter extraction step is based on the computation of local maxima in the Fourier transform of each segment, averaged over all the frames in the segment.

In a final step a HMM is automatically built to model music production as a stochastic process. The idea is that music recordings can be modeled with HMMs providing that states are labeled with events in the audio recording and their number is proportional to the number of segments, transitions model the temporal evolution of the audio recording and observations are related to the audio features previously extracted that help distinguishing different events.

At identification time, an unknown recording of a performance is preprocessed in order to extract the features modeled by the HMMs. All the models are ranked according to the probability of having generated the acoustic features of the unknown performance. Ideally the alignment of the query through the correspondent model will follow a linear trend, achieving the final higher probability. Since the simplicity of the models, coarse alignment between the events and the acoustic features could occur. This problem is handled by providing a support parameter, which measures the distance between the computed path and an estimated linear path. Such linear path can be estimated by considering the regression analysis of the computed alignment points.

A complete evaluation of the methodology with a collection composed of about 1000 recordings of classical music can be found in [11]. The complete system gave a final precision of 90% with the 84% of the analyzed query correctly ranked in the first position with an identification time of about 3 seconds. As for the clustering component, an extension for popular music is under development and initial results seems to be very promising.

3. IDENTIFICATION IN SIMILARITY ANALYSIS

A content-based music identification system may have different application scenarios, mainly towards accessing, organizing, browsing and recommending. Automatic identification of unknown recordings can be exploited as a tool for supervised manual labeling: the user is presented with a rank list of candidates, from which he can choose the matching one. Once that the unknown recording has been correctly identified, it could be indexed and added to the music collection. Identification may also be exploited to retrieve all the different versions of the query stored in the collection (live or cover).

Given the identification tool, a research prospective aims at exploiting it in a similarity measure context. At this purpose, one of the main issues concerns the understanding of the similarity concept itself. In fact, a very common consideration is about the structure of the rank list of an identification system and an immediate question may be: *"in an ideal situation where the matching documents are always ranked on the top, is the first non-matching item the most similar of the collection to the query?"*.

The answer is not an easy task. In fact, in a mathematical sense we may say that, it is the most similar because it shares the larger number of features. However, in a general sense, the similarity concept is very subjective and strictly related to the context and especially to the listeners. Basing the similarity measures just on the content is a bit reductive because music content is very various and many factors are involved. Moreover, the perception of the users could be various and likely much different from the computed one.

In the following, we propose some solutions to exploit the identification tool in a similarity context.

As it can be seen, the approach aims at computing all the content descriptors and to map them in some metric spaces in order to define a sort of distance among descriptors. Considering content distances and social tags, it would be possible to define similarity relationship for the documents of the collection. We believe that these computed similarity scores would be very descriptive since representing both music content and users perception.

A preliminary idea is based on representing the collection and the similarity relationships as an hidden Markov model where transition probabilities among states depend on the content-similarity relationships whereas the observation probabilities of each state are related to the social tags. Well known, statistical paths through the model [15] allowed an user to browse the collection according a similarity global value express in terms of probability which combines both content and user similarity relationships.

All social tags can be collected in different ways. In literature different methodologies to tag music have been proposed, mainly based on either human-annotations, web mining or auto content-based annotations [14]. A common approach is to gather human-based tags for a descriptive training set and then to exploit a content-based auto-tagger approach, such as the one proposed in [16] which uses a machine learning technique based on multivariate Gaussian mixture models to annotate the new songs.

The proposed system may be used in different tasks ranging from recommendation systems to the browsing by similarity of a collection. Especially concerning recommendation, we believe that it could be an useful tool for the "*items cold start*" issue [1], since new items will be provided with content-based descriptors that would be useful to define also likely social aspects.

4. CONCLUSIONS

This paper describes a methodology for the content-based identification of music documents. The aim is to identify each recordings of a score (including live performances and cover versions) through a collection of indexed high quality recordings. The approach is based on two steps. At first, a cluster of the collection is retrieved to highlight some potential candidates for the query, whereas a second step computes the similarity between the query and the documents of the cluster with an application of HMMs. The methodology is still under study, and prospective directions to improve the results are provided.

A content-based music identification system may have different application scenarios, where supervised manual labeling of unknown music recordings seems to be the most suitable. In this context, a larger concept of similarity, not only based on content features but also considering the cognitive perception of the listeners, can be introduced. At this aim, we proposed some different ideas to apply the music identification tool in a larger similarity context. The descriptions provided represent future approaches that we are going to investigate and we consider interesting for the community.

Bibliography

- [1] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, vol.17, no.6, June 2005.
- [2] A. Berenzweig, B. Logan, D. Ellis and B. Whitman. A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measure. *Computer Music Journal*, 28:2, pp.63–76, 2004
- [3] T. Bertin-Mahieux, D. Eck, F. Mailliet and P. Lamere. Autotagger: A Model for Predicting Social Tags from Acoustic Features on Large Music Databases *Journal of New Music Research*, vol.37(2), pp.115–135, 2008
- [4] P. Cano, E. Battle, T. Kalker, and J. Haitsma. A Review of Audio Fingerprinting. *Journal of VLSI Signal Processing*, vol.41, pp.271–284, 2005.
- [5] D. Ellis and G. Poliner. Identifying Cover Songs with Chroma Features and Dynamic Programming Beat Tracking. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.4, pp.1429–1432, 2007
- [6] A. Gionis, P. Indyk, and R. Motwani. Similarity Search in High Dimensions via Hashing. In *Proceedings of the International Conference on Very Large Databases*, pp.518–529, 1999.

- [7] P. Herrera, J. Serrá, E. Gómez and X. Serra. Chroma Binary Similarity and Local Alignment applied to Cover Song Identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151, 2008.
- [8] P. Lamere. Social Tagging and Music Information Retrieval *Journal of New Music Research*, vol.37(2), pp.101–114, 2008
- [9] T. Lidy and A. Rauber. Evaluation of Feature Extractors and Psycho-acoustic Transformations for Music Genre Classification In *Proceedings of the International Conference on Music Information Retrieval*, pp.304–310, 2005.
- [10] R. Miotto and N. Orio. A Methodology for the Segmentation and Identification of Music Works. In *Proceedings of the International Conference of Music Information Retrieval*, pp.271–284, 2007.
- [11] R. Miotto and N. Orio. A Music Identification System based on Chroma Indexing and Statistical Modeling. In *Proceedings of the International Conference on Music Information Retrieval*, pp.301–306, 2008.
- [12] R. Miotto and N. Orio. Music System for Live Performances Identification, May 2009. <http://svrims2.dei.unipd.it:8080/musicir/>.
- [13] M. Müller and F. Kurth. Efficient Index-based Audio Matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, 2008.
- [14] D. Turnbull, L. Barrington, D. Torres and G. Lanckriet. Five Approaches to Collecting Tags for Music. In *Proceedings of the International Conference on Music Information Retrieval*, pp.225–230, 2008.
- [15] L.R. Rabiner. A Tutorial on Hidden Markov Models and selected Application. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [16] D. Turnbull, L. Barrington, D. Torres and G. Lanckriet. Semantic Annotation and Retrieval of Music and Sound Effects. *IEEE Transactions on Audio, Speech, and Language Processing*, vol.16(2), pp.467–476, 2008
- [17] M. Slaney and M. Casey. Locality-Sensitive Hashing for Finding Nearest Neighbors. *IEEE Signal Processing Magazine*, vol.25(2), pp.128–131 (2008)
- [18] Y. Yu, S. Downie, F. Moerchen, L. Chen and K. Joe. Using Exact Locality Sensitive Mapping to Group and Detect Audio-Based Cover Songs. In *Proceedings of the 10th IEEE International Symposium on Multimedia*, pp.302–309, 2008.