A MUSIC IDENTIFICATION SYSTEM BASED ON CHROMA INDEXING AND STATISTICAL MODELING

Riccardo Miotto and Nicola Orio
Department of Information Engineering, University of Padova
{miottori,orio}@dei.unipd.it

ABSTRACT

A methodology is described for the automatic identification of classical music works. It can be considered an extension of fingerprinting techniques because the identification is carried out also when the query is a different performance of the work stored in the database, possibly played by different instruments and with background noise. The proposed methodology integrates an already existing approach based on hidden Markov models with an additional component that aims at improving scalability. The general idea is to carry out a clustering of the collection to highlight a limited number of candidates to be used for the HMM-based identification. Clustering is computed using the chroma features of the music works, hashed in a single value and retrieved using a bag of terms approach. Evaluation results are provided to show the validity of the combined approaches.

1 INTRODUCTION

This paper presents a methodology and its Web-based implementation for the automatic identification of music works. The approach is an extension of *audio fingerprinting* techniques. As it is well known, an audio fingerprint is a unique set of features that allows to identify digital copies of a given audio file because it is robust to the presence of noise, distortion, lossy compression and re-sampling. Identification is carried out by comparing the fingerprint of the unknown audio file with the fingerprints stored in a database, normally using indexing techniques. A comprehensive tutorial on methods and techniques for audio fingerprinting can be found in [3].

The main difference between the proposed approach and typical fingerprinting techniques is that the goal is to identify any possible recording of a given music work through statistical modeling. This representation is automatically computed from audio recordings, which are stored in a database. To this end, the methodology is strictly related to research work on *cover identification*. Yet, the envisaged application scenario is the automatic identification of baroque, classical, and romantic music or of any music genre where a written score is assumed to be the starting point of performances. In all these cases, the term "cover" is misleading because it assumes the existence of a prototype recording from which all the others derive. For this reason, the more general term *music identification* is used in this paper; for simplicity, the term "classical music" is used to address all the aforementioned repertories.

An automatic system able to identify performances of classical music may be helpful in a variety of situations. For instance, most theaters and concert halls routinely record all the rehearsals. This valuable material may totally lack metadata, with the possible exception of the date of the rehearsals, and the reconstruction of how a particular production has been created may become a difficult task. Similar considerations apply to the recordings of concerts that have been broadcasted and archived by radio and television companies. Most of the times, metadata are uninformative about the content of each single audio track, if not totally missing. From the end user point of view, automatic labeling of classical music can be helpful because many compositions have been recorded in a number of different CDs that, apart from the recordings made by well known artists, may not be listed in online services such as Gracenote [9], in particular for the typical case of compilations containing selected movements of recordings already released.

Automatic identification of classical music can be considered also a viable alternative to fingerprinting. In order to effectively identify a recording, fingerprinting techniques need to access the original audio files. For classical music this means that all the different recordings of a given work should be present, with an increase in computational costs, not mentioning the economical aspects in creating such a large reference collection.

The proposed methodology integrates an approach for the identification of orchestral live music [14] based on hidden Markov models (HMMs). A HMM is generated for each music work, starting from an audio recording. The approach showed to be effective but not scalable, because the recording has to be compared to all the recordings in the database. The identification becomes clearly unfeasible when the database contains a large number of music works.

To this end, an additional component has been designed with the aim of improving scalability. The idea is to carry out a clustering of the collection to highlight a limited number of candidates to be used for the HMM based identifi-

cation. The main requirement of clustering is efficiency, while effectiveness can be modulated by carefully selecting the size of the cluster.

2 CHROMA-BASED DESCRIPTION

Chroma features have been extensively used in a number of music retrieval applications. The concept behind chroma is that octaves play a fundamental role in music perception and composition [1]. For instance, the perceived quality – e.g. major, minor, diminished – of a given chord depends only marginally on the actual octaves where it spans, while is strictly related by the pitch classes of its notes. Following this assumption, a number of techniques have been proposed based on chroma features, for chord estimation [7, 18], detection of harmonic changes [11], and the extraction of repeating patterns in pop songs [8]. The application of chroma features to an identification task has been already been proposed for classical [15] and for pop [12] music.

2.1 Indexing

We propose to use chroma features as *indexes* for a collection of classical music recordings. These recordings are stored in a database system, and are used to create statistical models aimed at identification as described in Section 3. The basic idea is that information retrieval techniques can be generally employed outside the textual domain, because the underlying models are likely to be shared by different media [13]. Chroma features are considered as pointers to the recordings they belong to, playing the same role of words in textual documents. The information on the time position of chroma features is used to directly access to relevant audio excerpts.

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 apply to feature vectors, because exact match has to be replaced by similarity search, which is less efficient. A number of research works has been carried out to effectively carry out similarity search in high dimensional spaces, achieving good results in different tasks such as K-nearest neighbors search, clustering, and range searches [20]. In particular, a promising technique is Locality Sensitive Hashing [6]. This techniques aims at applying to feature vectors a carefully chosen hashing function with the aim of creating collisions between vectors that are close in the high dimensional space. The hashing function itself becomes then an effective tool to measure the similarity between two vectors.

It can be noted that fingerprint techniques, such as the one described in [10], can be considered variants of Locality Sensitive Hashing. Following this idea we propose to represent the 12-dimensional chroma vector with a single integer value, obtained through an hashing function. In particular, the actual hash depends on the ranks of the chroma pitch classes, and on on their absolute values.

More formally, a chroma vector $c(i)$ is an array of pitch classes, where $i \in \{1 \dots 12\}$, which are computed from the Fourier transform $S(f)$ of a windowed signal according to equation

$$c(i) = \sum_f B_i(f) \cdot S(f) \tag{1}$$

where $B_i(f)$ is a bank of bandpass filters, each filter centered on the semitones belonging to pitch class i and with a bandwidth of a semitone. The filterbanks may range over all the audible octaves, even if it is common practise to limit its support. In our case the minimum and maximum frequencies are respectively 20 and 5000 hertz.

Once vector $c(i)$ is computed for a window of the signal, we proposed to compute its rank-based quantization $q(i)$ through the following rule

$$q(i) = \begin{cases} rank[c(i)] & if\ \ rank[c(i)] \leq k \\ 0 & elsewhere \end{cases} \tag{2}$$

where k is the number of quantization levels that are taken into account and $rank[c(i)]$ is a function that outputs the rank of the value of the energy in pitch class i over the values of the whole array, that is $rank[c(j)] = 1$ if pitch class j has the highest energy and so on.

Finally, a non negative integer hash h is computed from $q(i)$ according to equation

$$h = \sum_{i=1}^{12} k^{i-1} q(i) \tag{3}$$

where common hashing techniques can be additionally applied to store the values h in one array, which can be accessed in constant time. In particular, we simply compute the reminder of h divided by a carefully chosen prime number.

Indexing is applied both to the recordings in the database and to the queries. One of the problems that may affect retrieval effectiveness is that chroma-based representation is sensible to transpositions. If the recording used as a query and the recording stored in the database are played in different tonalities, they have totally different sets of chroma. We addressed this problem by considering that a transposition of s semitones will result in a rotation of vector $c(i)$ of s steps (the direction of the rotation depending on the direction of the transposition). The value of h^s of quantized chroma h transposed by s ascending semitones, where $s \in \{1 \dots 11\}$ can be computed through equation

$$h^s = k^{12-s}(h \ \% \ k^s) + \lfloor \frac{h}{k^s} \rfloor \tag{4}$$

where $\%$ gives the reminder of a division and $\lfloor \cdot \rfloor$ is the floor function. Thus a query is represented by twelve sets

of hashes H^s, that is the representation H^0 in the original tonality and eleven additional ones that take into account all the possibile transpositions. This representation can be considered as an application of query extension techniques.

2.2 Retrieval

With the main goal of efficiency, continuing the parallel with textual information retrieval techniques, retrieval is carried out using the *bag of words* paradigm. Thus the similarity between the query and the recordings in the collection is carried out by simply counting the number of hashes of quantized chroma they have in common. No attempt to align, not even to consider the relative ordering of the quantized chroma, is made. The only additional processing is to highlight the excerpts of each recordings that potentially correspond to the query. This is simply made by using the timing information contained in the indexes to divide the recordings in overlapping frames.

The similarity M_q between the query and the recordings is computed by simply counting the number of hashes they have in common, that is by performing an intersection between the set H^0 representing the query – or one of its transposed versions H^s with $s \in \{1 \dots 11\}$ – with the set H_{jf} representing the frame f of the recording j in the database. This can be expressed through equation

$$M_q(s, j, f) = \|H^s \cap H_{jf}\| \tag{5}$$

where $||\cdot||$ returns the cardinality of a set, that is the number of its elements. The approach allows us to sort the recordings in the dataset according to the value of M and, at the same time, to highlight the frames f that more likely correspond to the query and the difference in semitones between the two tonalities.

The proposed similarity value is not very common in information retrieval, where more complex measures are exploited, such as the cosine of the angle between the vectors representing the query and the items in the database. Moreover, the similarity does not take into account the number of occurrences of a given hash value within the query and the recordings, which is normally exploited through the $tf \cdot idf$ weighting scheme. The choice of this particular similarity measure has been motivated by a number of tests showing that this simple measure outperformed more complex ones. It has to be considered that the distribution of a given hash value inside the database, which is taken into account by the *inverse document frequency* element idf, does not have a clear effect on the similarity between two recordings of the same music work. At the same time, the distribution of a given hash within a recording, the *term frequency* element tf, can be affected by the rate at which chroma vectors are computed and give different similarity values depending on the difference in tempo between two recording.

The top A retrieved recordings are considered as the cluster of candidates that are used in the second step, as discussed in the next section. It is important to note that quantized chroma are used only in the clustering step, where recall is much more important than precision. It is likely that more sophisticated techniques will increase precision, by assigning a higher rank to the recording that corresponds to a query, but will have a minor impact on recall.

2.3 Chroma Evaluation

The indexing methodology has been evaluated with a testbed composed by a database of 1000 recordings and by a query set of 50 different performances of a subset of the works in the database. The audio files were all polyphonic recordings of orchestral music, with a sampling rate of 44.1 kHz. We segmented the recordings in frames, as discussed in the previous section, with a length of 10 seconds. Subsequent frames had an overlap of 50%. Queries were made of audio excerpts of about the half of the length of the frames, taken at random within the audio query files. Testing was carried out by extracting random frames of the 50 queries, matching them with all the frames of the recordings in the database and measuring the rank of the recording that corresponded to the query. This experimental setup aims at simulating a real situation, where the user can query the system using any part of an unknown recording. The purpose of the evaluation was to prove the validity of hashes of the quantized chroma features as index terms, to highlight the best configuration of the parameters, and to select the size of the cluster.

We tested the effectiveness of the approach depending on the size of the window used to compute the chroma features. According to this value, the precision of chroma features would increase or decrease, affecting the clustering rates. We tested the system with different window sizes, while the hopsizes between windows consistently set to $1/4$ of the window size. Results showed that the optimal window length, with this experimental setup, is 1024 points (about 23 milliseconds). Moreover, we tested different sizes of the hashing map by varying the quantization level k, that is the number of pitch classes that are used to compute the hashing function, as explained in Section 2.1. The results highlighted the optimal number of levels used for quantization, which is $k = 6$. With this value, and with this experimental setup, a cluster of the first 100 recordings is enough to assure that the correct document will be processed in the following HMM-based identification step. A cluster size of only 50 elements contains the correct document in 96% of the cases.

These results are particularly encouraging, because they are obtained with a compact representation of audio recordings – which allows us to compute an exact match between features – and a retrieval based on a simple bag of terms approach – which does not require any alignment technique.

Table 1 reports the percentage of times the correct recording has been ranked within predefined thresholds.

method	$=$ **1**	$\leq$ **10**	$\leq$ **50**	$\leq$ **100**	$\leq$ **200**
chroma	69	88	96	100	100

Table 1. Identification rates using the quantized chroma

3 A COMPLETE IDENTIFICATION SYSTEM

The first system architecture was proposed in [14], where the results of a set of experimental tests have been presented. Identification is based on a *audio to audio* matching process, which goal is to retrieve all the audio recordings that represents the same music content as the audio query. The audio recordings used for the identification are stored in a relational database, together with a set of relevant metadata. Audio to audio matching has been also proposed in [15], where a chroma representation was used to identify classical music. Another application of chroma features to an identification task, this time on pop music, has been proposed in [12] where chroma features of the query song and the songs in the database are aligned using Dynamic Time Warping (DTW). Global similarity, obtained through DTW, between audio excerpts has been used in [5] to align in real time two different performances of the same music work.

The current retrieval architecture, which is depicted in Figure 1, is based on two main components: a clustering block (on the left) that selects efficiently a number of candidate compositions and an identification block (on the right) that computes the probability that the audio query correspond to each of the candidates.

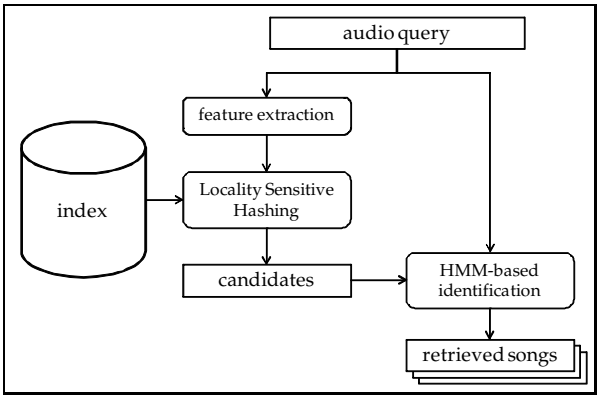

Figure 1. Structure for a music identification system

In this section, for the sake of completeness, the main elements of the identification block are briefly summarized. A more detailed description can be found in [14]. In a first step, the segmentation process extracts audio subsequences that have a coherent acoustic content. Segmentation exploits a technique that has been proposed for the segmentation of textual documents in coherent parts, a task also called topic segmentation [4]. The aim of segmentation is to divide the music signal into subsequences that are bounded by the presence of music events, where an event occurs whenever the current pattern of a music piece is modified, typically one or more new notes being played or stopped.

Coherent segments of audio are 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. To this aim each segment needs to be described by a compact set of features that are automatically extracted. Considering pitch as the most relevant parameter for music identification and 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 from segmentation and parameterization to model music production as a stochastic process. HMMs are stochastic finite-state automata, where transitions between states are ruled by probability functions. At each transition the new state emits a random vector with a given probability density function [19]. The basic 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 pre-processed 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 performance to be identified.

3.1 Evaluation of the Complete System

We evaluate the effectiveness and the efficiency of a system implementing the methodology by using the same experimental collection described in Section 2.3. Clustering has been carried out by selecting, for each query, 100 frames of the recordings in the database according to their similarity with the query or its transposed versions. Hashes have been computed using an analysis window of 1024 points and using $k = 6$ levels of quantization. The queries had a length of 7 seconds, while recordings in the database have been segmented in overlapping frames of 10 seconds.

Results are shown in Table 2, which compares the effectiveness of the complete approach with the one based on HMM. As it can be seen, clustering improves effectiveness because the complete method gave a Mean Reciprocal Rank

(MRR) of 90.2, which is significantly higher than MRR for the pure HMM-based approach. Moreover, for the complete approach, every query gives the correct work ranked among the first 10 positions, which is an important result for many of the applications of supervised music identification that have been described in Section 1.

method	$= 1$	≤ 3	≤ 5	≤ 10	≤ 20	MRR
only HMM	84	88	90	94	94	86.7
complete	86	94	94	100	100	90.2

Table 2. Identification rates using the HMM-based approach and the complete methodology

The complete approach gives clear improvements also on efficiency. Because chroma-based clustering can be carried out efficiently, its computational cost is negligible compared to HMM-based identification. Thus the reduction of the number of candidates gives an improvement of orders of magnitude, because the cluster contains a fixed number of excerpts of recordings while, without clustering, the query should compared against all the experts of all the recordings. In our experiments, identification of an audio excerpts of about 7 seconds is carried out in 2.5 seconds.

It is interesting to note that most of the queries have a different tempo than the corresponding recordings in the database. The approach seems to be robust to tempo differences, even if commercial recordings, especially of orchestral music, do not present large deviations from the tempo written in the score. In order to test robustness to tempo, queries have been stretched from 0.8 to 1.2 of their original bpms (larger deviations were totally unrealistic with this particular dataset). Results with this artificial dataset did not showed changes in the identification rate, yet more experiments should be carried out with real data to effectively measure the robustness to tempo deviations

3.2 A Prototype for Music Identification

The system in Section 3 has been implemented in a Web-oriented prototype available on-line [16], which works as a demonstrator of the proposed approach. The core engine of the application, which include indexing, modeling and identification steps, has been implemented in C++. The music collection is composed by the indexed music recordings and is stored in a PostgreSQL database, including both the metadata and the music content information. A number of queries are available for the user to test the system, including some recordings where different noisy sources have been added to the original queries, such as environmental noises, and typical noises of analog equipments (i.e., clicks, hisses). Additional queries are available with piano solo version of orchestral music and dramatic changes in tempo, both faster and slower than the original, in order to show the effect of changes in orchestration and in tempo, although the approach has not been tested extensively with changes in orchestration.

All the examples are coded with lossy compression, in order to be easily listened also through the Web. The user can access to the metadata associated to the recordings in the database and listen to the rank list of music works returned by the prototype. For this reason, only copyright free recordings are stored in the database, which thus has a limited size of about 1000 music works.

At the current state the prototype allows to test the system with the supplied examples, by visualizing the first 20 results of the whole rank list. In order to find the right match, the user can listen to the resulting elements to find the most similar one. Considering all the experimental results reported in the previous sections, the user has good probability of listening just the first retrieved music works to find the relevant one. The approach is thus to perform a supervised identification, because the final choice of which is the correct music work is left to the user.

4 CONCLUSIONS

This paper describes a methodology for the identification of works of classical music. It is based on two steps. At first it carries out a clustering of a database of recordings in order to select a set of candidates music works. Clustering is based on the introduction of an hashing function that maps the chroma features on integer numbers. The second step computes the similarity between the audio query and the recordings in the database by exploiting an application of HMMs. Thus identification is carried out using two different strategies, the first to achieve efficiency, the second to refine effectiveness.

The methodology has been implemented in a Web-based prototype that allows the user to identify even short audio excerpts of classical music. The identification rate, which has been computed using a database of 1000 recordings of orchestral music, is encouraging. Already using the first step, the system was able to correctly identify the query in 69% of the times. The second step introduces an additional refinement of the identification rate that reaches, in our experiments, 86%.

It is important to note that the proposed approach assumes that different performances are based on a similar music score, even if played with different instrumentation, tempo, or key signature. The system has not been tested on music genres where improvisation is a predominant feature, such as jazz music, or where the music score can be substantially modified through arrangements. Yet, some results on the effectiveness of the individual techniques show that the general approach could be extended also, at least, to pop and rock music. For instance, many works on cover song

identification are based on the use of chroma vectors [18] while HMM-based identification has been applied also to pop music with encouraging results [17].

It could be argued that the proposed methodology can hardly be extended to the identification of recordings of all the music genres. One of the lessons learnt by the information retrieval community is that some of the techniques for text retrieval (i.e., stemming), are not equally effective for all the languages and all the retrieval tasks. Nevertheless, we aim at extending the proposed approach to other music genres, by introducing different descriptors that are related to the main features of each music genre. For instance, the main melody in the case of pop and rock or the chord progression in the case of jazz.

5 ACKNOWLEDGMENTS

This work was partially supported by the SAPIR project, funded by the European Commission under IST FP6 (Sixth Framework Programme, Contract no. 45128), and by the project "Analysis, design, and development of novel methodologies for the study and the dissemination of music works", funded by the University of Padova.

6 REFERENCES

[1] M.A. Bartsch and G.H. Wakefield "Audio Thumbnailing of Popular Music Using Chroma-based Representations", *IEEE Transactions on Multimedia*, 1996.

[2] L. Boney and A. Tewfik and K. Hamdy "Digital Watermarks for Audio Signals", *IEEE Proceedings Multimedia*, 1996.

[3] P. Cano and E. Batlle and T. Kalker and J. Haitsma "A Review of Audio Fingerprinting", *Journal of VLSI Signal Processing*, 2005.

[4] F.Y.Y. Choi "Advances in domain independent linear text segmentation", *Proceedings of the Conference on North American chapter of the Association for Computational Linguistics*, 2000.

[5] S. Dixon and G. Widmer "MATCH: a Music Alignment Tool Chest", *Proceedings of the International Conference of Music Information Retrieval*, London, UK, 2005.

[6] A. Gionis and P. Indyk and R. Motwani "Similarity Search in High Dimensions via Hashing", *Proceedings of the International Conference on Very Large Data Bases*, Edinburgh, UK, 1999.

[7] E. Gómez and P. Herrera "Automatic Extraction of Tonal Metadata from Polyphonic Audio Recordings", *Proceedings of the Audio Engineering Society*, London, UK, 2004.

[8] M. Goto "A Chorus Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station", *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[9] Gracenote `http://www.gracenote.com/`, June 2008.

[10] J. Haitsma and T. Kalker and J. Oostveen "Robust audio hashing for content identification", *Proceedings of the Content-Based Multimedia Indexing Conference*, Firenze, IT, 2001.

[11] C.H. Harte and M. Sandler and M. Gasser "Detecting Harmonic Changes in Musical Audio", *Proceedings of the ACM Multimedia Conference*, Santa Barbara, USA, 2006.

[12] N. Hu and R.B. Dannenberg and G. Tzanetakis "Polyphonic Audio Matching and Alignment for Music Retrieval", *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[13] M.S. Lew and N. Sebe and C. Djeraba and R. Jain "Content-based Multimedia Information Retrieval: State of the Art and Challenges", *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2006.

[14] R. Miotto and N. Orio "A Methodology for the Segmentation and Identification of music Works", *Proceedings of the International Conference of Music Information Retrieval*, Vienna, A, 2007.

[15] M. Müller and F. Kurth and M. Clausen "Audio Matching Via Chroma-based Statistical Features", *Proceedings of the International Conference of Music Information Retrieval*, London, UK, 2005.

[16] Music Identification Prototype `http://svrims2.dei.unipd.it:8080/musicir/`, June 2008.

[17] N. Orio and C. Zen "Song Identification through HMM-based Modeling of the Main Melody", *Proceedings of the International Computer Music Conference*, Copenhagen, DK, 2007.

[18] G. Peeters "Chroma-based Estimation of Musical Key from Audio-signal Analysis", *Proceedings of the International Conference of Music Information Retrieval*, Victoria, CA, 2006.

[19] L.R. Rabiner "A Tutorial on Hidden Markov Models and Selected Application", *Proceedings of the IEEE*, 1989.

[20] H. Samet "Foundations of Multidimensional and Metric Data Structures.", Morgan Kaufmann Publishers Inc., San Francisco, USA, 2006.