Automatic Identification of Music Works through Audio Matching

Riccardo Miotto and Nicola Orio

Department of Information Engineering, University of Padua, Italy
{miottori, orio}@dei.unipd.it

Abstract. The availability of large music repositories poses challenging research problems, which are also related to the identification of different performances of music scores. This paper presents a methodology for music identification based on hidden Markov models. In particular, a statistical model of the possible performances of a given score is built from the recording of a single performance. To this end, the audio recording undergoes a segmentation process, followed by the extraction of the most relevant features of each segment. The model is built associating a state for each segment and by modeling its emissions according to the computed features. The approach has been tested with a collection of orchestral music, showing good results in the identification and tagging of acoustic performances.

1 Introduction

Automatic identification of music works is gaining increasing interest because it can provide new tools for music accessing and distribution. Manual identification of music works is a difficult task that, ideally, should be carried out by trained users who remember by heart hundreds, or even thousands, of hours of music. Non expert users, instead, are usually able to recognize only well known works, and they may require the aid of an automatic tool for labeling the recordings of performances of unknown works. Automatic tools are particularly useful with instrumental music, when lyrics are not available for recognizing a particular work. Metadata about music works are needed also during the creation of a music digital library. For instance, theaters, concert halls, radio and television companies have usually hundreds of hours of almost unlabeled analog recordings, which witness the activities over the years of the institution and that need to be digitized and catalogued for preservation and dissemination. Moreover, music is extensively used as the background of commercials, television shows, and news stories. The automatic identification of music works employed as audio background may be useful for users, that can access for new interesting material.

A common approach to music identification is to extract, directly from a recording in digital format, its *audio fingerprint*, which is a unique set of features that allows for the identification of digital copies even in presence of noise, distortion, and compression. It can be seen as a content-based signature that summarizes an audio recording. Applications of audio fingerprinting include

Web-based services that, given a sample of recording, provide the users with metadata about authors, performers, recording labels, of given unknown digital recordings. A comprehensive tutorial about audio fingerprinting techniques and applications can be found in [1]. Audio fingerprinting systems are designed to identify a particular performance of a given music work. This assumption is valid for many applications. For instance, users are interested to particular recordings of a given music work, e.g. the one of a renown group rather than of a garage band. Moreover, digital rights management systems have to deal also with the rights of the performers. For these reasons, the audio fingerprint is computed from recordings, and usually it is not able to generalize the features and to identify different performances of the same music work. On the other hand, the identification of a music work may be carried out also without linking the process to a particular performance. Music identification of broadcasted live performances may not benefit from the fingerprints of other performances, because most of the acoustic parameters may be different. In the case of classical music, the same works may have hundreds of different recordings, and it is not feasible to collect all of them in order to create a different fingerprint for each recording. To this end, a methodology that allows the user to recognize the different instances of a given music work, without requiring the prior acquisition of all the available recordings, could be a viable alternative to audio fingerprinting.

An alternative approach to music identification is *audio watermarking*. In this case, research on psychoacoustics is exploited in order to embed in a digital recording an arbitrary message, the watermark, without altering the human perception of the sound [2]. The message can provide metadata about the recording (such as title, author, performers), the copyright owner, and the user that purchases the digital item. The latter information can be useful to track the responsible of an illegal distribution of digital material. Similarly to fingerprints, audio watermarks should be robust to distortions, additional noise, A/D and D/A conversions, and compressions. On the other hand, watermarking techniques require that the message is embedded in the recording before its distribution and it is almost impossible to watermark the millions of digital recordings already available on the Internet. Moreover, watermarks can be made unreadable using audio processing techniques.

This paper reports a novel methodology for automatic identification of music works from the recording of a performance, yet independently from the particular performance. Unknown music works are identified through a collection of indexed audio recordings, ideally stored in a music digital library. The approach can be considered a generalization of audio fingerprinting, because the relevant features used for identification are not linked to a particular performance of a music work. Clearly the approach allows the user to identify the metadata related to a musical work and not to the particular performance used for the identification. The limitation of not identifying the performers can be balanced by the fact that only a single instance of a given work needs to be stored in the database. Moreover, as already mentioned, audio fingerprinting techniques are not able to identify live performances. The methodology reported in this paper extends

previous work on music identification based on audio to score matching [3], where performances were modeled starting from the corresponding music scores. Also in this case, identification is based on hidden Markov models (HMMs). The application scenario is the automatic labeling of performances of tonal Western music through a match with pre-labeled recordings that are already part of an incremental music collection. Audio to audio matching has been proposed in [4,5] for classical music audio to audio matching and audio to audio alignment respectively, and in [6] for pop music.

2 Automatic Identification of Music Performances

The automatic identification of music performances is based on a *audio to audio* matching process, which goal is to retrieve all the audio recordings from a database or a digital library that, in some sense, represent the same musical content as the audio query. This is typically the case when the same piece of music is available in several interpretations and arrangements.

The basic idea of the proposed approach is that, even if two different performances of the same music work may dramatically differ in terms of acoustic features, it is nevertheless possible to generalize the music content of a recording in order to model the acoustic features of other, alternative, performances of the same music work. A recording can thus be used to statistically model other recordings, providing that they are all performed from the same score. It has to be noted that the proposed methodology is particularly suitable for tonal Western music, and other music genres where performers strictly adhere to a give music score. This may not be the case of jazz music, where musicians may change the melodic and rhythmic structure of a given song. To cope with this genre, other dimensions may be more suitable, for instance the harmonic structure. Applications to rock and pop music are under current development, generalizing the concept of music score with a representation similar to the lead-sheet model proposed in [7].

With the aim of creating a statistical model of the score directly from the analysis of a performance, the proposed methodology is based on a number of different steps, as depicted in Figure 1. In a first step, *segmentation* extracts audio subsequences that have a coherent acoustic content. Audio segments are likely to be correlated to stable parts in a music score, where there is no change in the number of different voices in a polyphony. Coherent segments of audio are analyzed through a second step, called *parameter extraction*, which aims at computing a set of acoustic parameters that are general enough to match different performances of the same music work. In a final step, *modeling*, a HMM is automatically built from segmentation and parametrization to model music production as a stochastic process. At matching 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.

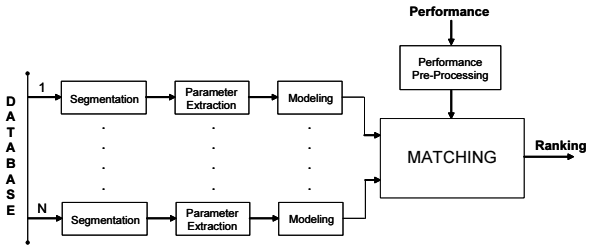


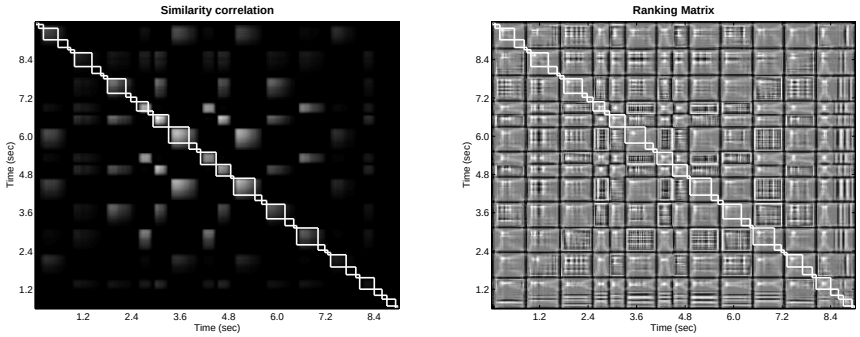
Fig. 1. Audio to audio matching process.

2.1 Segmentation

The audio recording of a performance is a continuous flow of acoustic features, which depends on the characteristics of the music notes – pitch, amplitude, and timbre – that vary with time according to the music score and to the choices of the musicians. In order to be structured, the audio information has to undergo a *segmentation* process. According to [8], the word segmentation in the musical world can have two different meanings: one is related to musicology and is normally used in symbolic music processing, whereas the other one follows the signal processing point of view and it is used when dealing with acoustic signals. In the latter case, the aim of segmentation is to divide a musical signal into subsequences that are bounded by the presence of music events. An event, in this context, occurs whenever the current pattern of a musical piece is modified. Such modifications can be due to one or more notes being played, possibly by different instruments, to active notes being stopped, or to a change in pitch of one or more active notes. This approach to segmentation is motivated by the central role that pitch plays in music language. In fact the segmentation of the acoustic flow can be considered the process of highlighting audio excerpts with a stable pitch.

The representation of a complete performance can then be carried out through the concatenation of its segments. In the proposed approach, segmentation is carried out by computing the spectrogram of the signal, and then taking the correlation of different frames represented in the frequency domain. Frames were computed using windows of 2048 samples – approximately 46 msecs – with an hopsize of 1024 samples. High correlation is expected between frames where the same notes are playing, while a drop in correlation between two subsequent frames is related to a change in the active notes. Thus correlation has been used as a similarity measure between audio frames. Similarity between different parts of an audio recording can be represented as in the left part of Figure 2, that is with a symmetric matrix where high similarity values are represented by bright pixels, top-left and bottom-right pixel show the self-similarity for the first and last frame and bright square regions along the diagonal represent the potential similar regions.

Segmentation has been carried out according to the methodology proposed in [9], which has been developed in the context of text segmentation. In partic-



(a) (b)

Fig. 2. Similarity (a) and rank (b) matrices with highlighted segments.

ular, hierarchical clustering on the similarity matrix is exploited to segment a sequence of features – being either textual elements or acoustic parameters – in coherent passages. According to [9], segmentation effectiveness can be improved if clustering is performed on a ranking matrix, which is computed by replacing each value in the similarity matrix with its rank in a local region, where the local region size can vary according to the context. The rank parameter is defined as the number of neighbors with a lower similarity value and it is expressed as a ratio between the number of elements with a lower value and the number of elements examined to circumvent normalization problems along the matrix bounds. Figure 2 shows the two different matrixes depicting the segments along the main diagonal.

The clustering step computes the location of boundaries using Reynar's maximization algorithm [10], a method to find the segmentation that maximizes the inside density of the segments. A preliminary analysis of the segmentation step allowed us to set a threshold for the optimal termination of the hierarchical clustering. It is interesting to note that it is possible to tune the termination of hierarchical clustering, in order to obtain different levels of cluster granularity, for instance at note level or according to different sources or audio classes. In our experiments, audio samples have been normalized to obtain similar levels of segmentation granularity between the performances.

Figure 3 depicts the segments computed with the proposed techniques, superimposed to the energy trend of an audio recording, the same that have been used to represent matrixes of Figure 2. It is important to note that these figures depicts the results of segmentation applied to a quite simple audio excerpt – monophonic plain audio – and they are shown in order to have a clearer visual representation than polyphonic audio segmentation, which is the scope of our approach. In this case, it can be seen that the spectral based segmentation is highly correlated with the energy envelope.

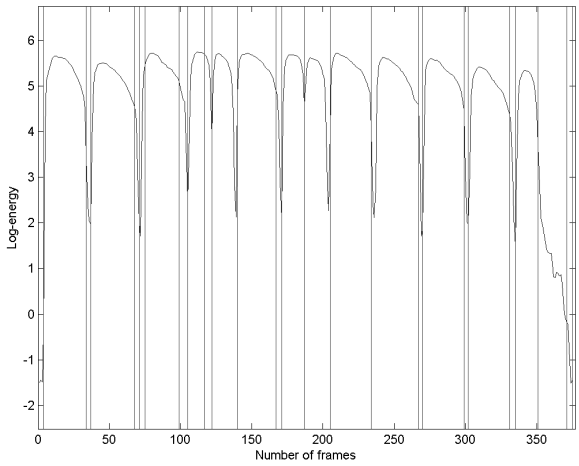


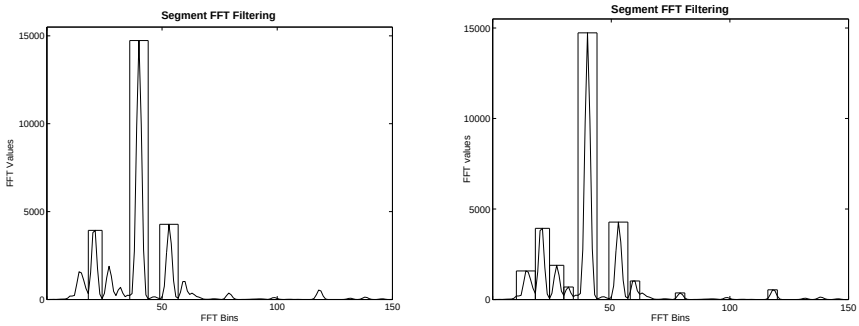
Fig. 3. Example of segmentation of a recording represented by its energy envelope.

2.2 Parameter Extraction

In order to obtain a general representation of an acoustic performance, each segment needs to be described by a compact set of features that are automatically extracted. In line with the approach to segmentation, also parameter extraction is based on the idea that pitch information is the most relevant 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 general, the spectra of different real performances of the same music work may vary because of differences in performing styles, timbre, room acoustics, recording equipment, and audio post processing. Yet, for all the performances the positions of local maxima are likely to be related to the position along the frequency axis of fundamental frequency and the first harmonics of the notes that are played in each frame. Thus a reasonable assumption is that alternative performances will have at least similar local maxima in the frequency representations, that is the dominant pitches will be in close positions.

When comparing the local maxima of the frequency representation, it has to be considered that Fourier analysis is biased by the windowing of a signal, which depends on the type and of the length of the window. These effects are expected both on the reference performances and on the performance to be recognized. Moreover, small variances on the peaks positions are likely to appear between different performances of the same music work, because of imprecise tuning and different reference frequency. For these reasons, instead of selecting only the peaks in the Fourier transform, each audio segment has been described



(a) (b)

Fig. 4. Parameters extraction considering 70% (a) and 95% (b) of the overall energy.

by a set of bin intervals, centered around the local maxima and with the size of a quarter tone. The number of intervals is computed automatically, by requiring that the sum of the energy components within the overall intervals is above a given threshold. Figure 4 depicts two possible sets of relevant intervals, depending on the percentage of the overall energy required: 70% for case (a) and 95% for case (b). It can be noted that a small threshold may exclude some of the peaks, which are thus not used as content descriptors.

2.3 Modeling

Each music work is modeled by a hidden Markov model, which parameters are computed from an indexed performance. 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. A HMM λ is completely defined by:

- a set of N states $Q = \{q_1, \ldots, q_N\}$, in which the initial and final states are identified;
- a probability distribution for state transitions, that is the probability to go from state q_i to state q_j;
- a probability distribution for observations, that is the probability to observe the features r when in state q_j.

Music works can be modeled with a HMM providing that: states are labeled with events in the audio recording, transitions model the temporal evolution of the audio recording, and observations are related to the audio features previously extracted that help distinguishing different events. The model is hidden because only the audio features can be observed and it is Markovian because transitions and observations are assumed to depend only on the actual state.

The number of states in the model is proportional to the number of segments in the performance. In particular, experiments have been carried out using n

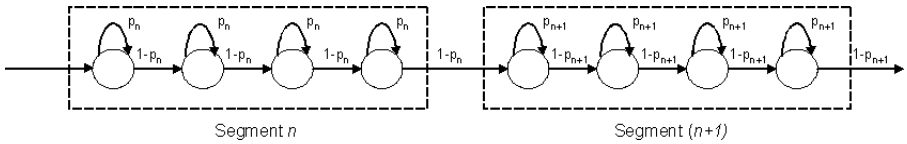


Fig. 5. Graphical representation of HMM corresponding to two general segments.

states for each segment. Figure 5 shows the HMM topology corresponding to two segments of four states each. It is proposed that states can either perform a self-transition, which models segments duration, or forward-transitions, which model the change from a segment to the following one. All the states in a given segment have the same probability p of performing a self-transition. Given this limitation, the probability of having a given segment duration is a negative binomial:

$$P(d) = \binom{d-1}{n-1} p^{d-n}(1-p)^n$$

The values n and p can be computed on the basis of the expected duration of the segments and transition probabilities, information that can be extracted from the actual duration of each segment. Durations need to be statistically modeled because different performances of the same music work may remarkably differ in timing. Each state in the HMM is labeled to a given segment and, accordingly with the parameter extraction step, emits the probability that a relevant fraction of the overall energy is carried by the frequency intervals computed at the previous step.

The modeling approach is similar to the one presented in [3], and, in fact, one of the goals of this work was to create a common framework where an unknown performance could be recognized from either its score or an alternative performance.

2.4 Identification

Recognition, or identification, is probably the application of HMMs that is most often described in the literature. The identification problem may be stated as follows:

- given an unknown audio recording, described by a sequence of audio features $R = \{r(1), \cdots, r(T)\}$,
- given a set of competing models λ_i,
- find the model that more likely generates R.

The definition does not impose a particular evolution of models λ_i, that is the path across the N states that corresponds to the generation of R. This allows us to define a number of criteria for solving the identification problem, depending on different constraints applied to the evolutions of the states of λ_i. Three different approaches are proposed, whose names are derived from the notation proposed in a classical tutorial on HMMs [11].

Approach α The most common approach to HMM-based identification, is to compute the probability that λ_i generates R regardless of the state sequence. This can be expressed by equation

$$\lambda_\alpha = argmax_i P(R|\lambda_i)$$

where the conditional probability is computed over all the possible state sequences of a model. The probability can be computed efficiently using the *forward probabilities*, also known as alpha probabilities. Even if approach α is the common practise for speech and gesture recognition, it may be argued that also paths that have no relationship with the actual performance give a positive contribution to the final probability. For instance, a possible path, which contributes to the overall computation of the forward probabilities, may consist in the first state of the HMM that continuously performs self-transitions. These considerations motivate the testing of two additional approaches.

Approaches δ and γ Apart from recognition, another typical HMM problem is finding the most probable path across the states, given a model and a sequence of observations. A widely used algorithm is Viterbi decoding, which computes a path that is *globally* optimal according to equation

$$q^\delta = argmax_q P(q|R, \lambda_i)$$

Alternatively, a *locally* optimal path [12] can be computed, according to equation

$$\overline{q}(t) = argmax_q P(q(t)|R, \lambda_i)$$
$$q^\gamma = \{\overline{q}(1), \overline{q}(2), \ldots, \overline{q}(T)\}$$

Both global and local optimal paths can be used to carry out an identification task, for finding the model that more likely generates R while state evolution is constrained. This approach leads to equations

$$\lambda_\delta = argmax_i P(R|q^\delta, \lambda_i)$$
$$\lambda_\gamma = argmax_i P(R|q^\gamma, \lambda_i)$$

that show how the probability of R is conditioned both by the model λ_i and by the state sequence of the global or optimal paths.

2.5 Computational Complexity

All the presented approaches allow the computation of the probabilities using a dynamic programming approach. In particular, it is known in the literature that each of the approaches requires $\mathbf{O}(DTN^2)$ time, where D is the number of competing models, T is the duration of the audio sequence in analysis frames, and N is the average number of states of the competing HMMs. Considering that, as described in Section 2.3, each state may perform a maximum of two

transitions, it can be shown that complexity becomes $\mathbf{O}(DTN)$. In order to increase efficiency, the length of the unknown sequence should be small, that is the method should give good results also with short audio excerpts.

An important parameter for computational complexity is the number of states N. A first approach to reduce N is to compute a coarse segmentations, which corresponds to a smaller number of group of states. On the other hand, a coarse segmentation may give poor results in terms of emission probabilities, because a single segment could represent parts of the performance with a low internal coherence. Another approach to reduce the computational complexity is to use a small number of states n for each segment, and model the durations with higher values of the self-transition probabilities p. As previously mentioned, in our experiments we found that setting $n = 4$ for each segment gave a good compromise.

3 Experimental Evaluation

The methodology has been evaluated with real acoustic data from original recordings taken from the personal collection of the authors. Tonal Western music repertoire has been used as a test-bed because it is a common practice that musicians interpret a music work without altering pitch information, which is the main feature used for identification.

The audio performances used to create the models were 100 incipits of orchestral works of well known composers of Baroque, Classical, and Romantic periods. All the incipits used for the modeling had a fixed length of 15 seconds. The audio files were all polyphonic recordings, with a sampling rate of 44.1 kHz, and they have been divided in frames of 2048 samples, applying a hamming window, with an overlap of 1024 samples. With these parameters, a new observation is computed every 23.2 milliseconds. The recordings to be recognized were 50 different performances of the same music works used to build the models. Also in this case they were an incipit of the music works, with a length of 10 seconds. The shorter lengths guaranteed that the performances to be recognized were shorter than the ones used to build the models, even in the case the two performances had a different tempo. The actual requirement is that the performance to be recognized is at least as long as the performance used for the recognition.

The 50 audio excerpts have been considered as unknown sequences to be identified, using the alternative approaches presented in Section 2.4. Table 1 reports the identification rates for the three approaches in terms of mean Average Precision, which is a well known measure in information retrieval and, for an identification task, it is equal to the mean of the reciprocal of the rank of the musical work to be identified.

With this experimental setup the average identification rate was quite different among the approaches, with α outperforming the two approaches that take into account also the global an local optimal paths. A more detailed analysis of the results highlighted that, in some cases, local differences between the ref-

Table 1. Identification rates for the different approaches, in terms of Average Precision.

Approach	Mean Average Precision (%)
α	78.7
γ	39.3
δ	51.1

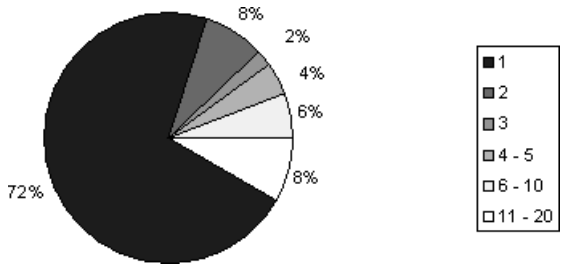


Fig. 6. Rank distributions of correct matches in the α identification test.

erence and the unknown performances gave unreliable results in the alignment, affecting the overall identification rates.

For the particular case of α, more detailed results are reported in Figure 6 that shows the percentages at which the correct audio recording was ranked as the most similar one, and when it was ranked within the first two, three, five, ten and twenty positions. As it can be seen, 43 out of 50 queries (86%) returned correct match among top 3 models, and 36 among them (72%) were correctly identified. Moreover, only 4 queries (8%) returned the correct match after the first 10 positions. These encouraging results allow us to consider the methodology suitable for the development of a supervised system for music identification. A typical scenario could be the one of an user that, after running the identification routines, is provided with a list of potential matches, together with a link to the reference performances that he can listen to, in order to finally identify the unknown recording.

4 Conclusions

A methodology for automatic music identification based on HMMs has been proposed. Three approaches to compute the conditional probability of observing a performance given the model of an indexed audio recording have been tested on a collection of digital acoustic performances. Experimental results showed that, at least for tonal Western music, it is possible to achieve a good identification rate. In particular, the typical approach to recognition based on the used of forward probabilities, which has been defined as the α approach, achieved an identification rate of 72%. Alternative approaches, which take into account the

alignment between the reference and the unknown performances, did not have comparable performances.

These results suggest that the approach can be successfully exploited for a retrieval task, where the user queries the system through an acoustic recording of a music work. The automatic identification of unknown recordings can be exploited as a tool for supervised manual labeling: the user is presented with a ranked list of candidate music works, from which he can choose. In this way, the task can be carried out also by non expert users, because they will be able to directly compare the recordings of the unknown and of the reference performances. Once that the unknown recording has been correctly recognized, it can be indexed and joint to the musical digital library, allowing us to increment the information stored inside it.

A prototype system has been developed, which allows a user, after recording or downloading an excerpt of a performance of classical music, to obtain after few seconds the relevant metadata of the music work.

References

1. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A review of audio fingerprinting. Journal of VLSI Signal Processing **41** (2005) 271–284
2. L. Boney, A.T., Hamdy, K.: Digital watermarks for audio signals. IEEE Proceedings Multimedia (1996) 473–480
3. Orio, N.: Automatic recognition of audio recordings. In: Proceedings of the Italian Research Conference on Digital Library Management Systems. (2006) 15–20
4. Müller, M., Kurth, F., Clausen, M.: Audio matching via chroma-based statistical features. In: Proceedings of the International Conference of Music Information Retrieval. (2005) 288–295
5. Dixon, S., Widmer, G.: MATCH: a music alignment tool chest. In: Proceedings of the International Conference of Music Information Retrieval. (2005) 492–497
6. Hu, N., Dannenberg, R., Tzanetakis, G.: Polyphonic audio matching and alignment for music retrieval. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. (2003) 185–188
7. Seifert, F.: Semantic music recognition – audio identification beyond fingerprinting. In: Proceedings of the International Conference of Web Delivering of Music. (2004) 118–125
8. Aucouturier, J.: Segmentation of Music Signals and Applications to the Analysis of Musical Structure. Master Thesis, King's College, University of London, UK (2001)
9. Choi, F.Y.: Advances in domain independent linear text segmentation. In: Proceedings of the Conference on North American chapter of the Association for Computational Linguistics. (2000) 26–33
10. Reynar, J.: Topic Segmentation: Algorithms and Applications. PhD Thesis, Computer and Information Science, University of Pennsylvania (1998)
11. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ (1993)
12. Raphael, C.: Automatic segmentation of acoustic musical signals using hidden markov models. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(4) (1999) 360–370