Experiments on Segmentation Techniques for Music Documents Indexing

Nicola Orio
Department of Information Engineering
Via Gradenigo, 6/A
35131 Padova – Italy
orio@dei.unipd.it

Giovanna Neve
Department of Information Engineering
Via Gradenigo, 6/A
35131 Padova – Italy
giovanna.neve@virgilio.it

ABSTRACT

This paper presents an overview of different approaches to melody segmentation aimed at extracting music lexical units, which can be used as content descriptors of music documents. Four approaches have been implemented and compared on a test collection of real documents and queries, showing their impact on index term size and on retrieval effectiveness. From the results, simple but extensive approaches seem to give better performances than more sophisticated segmentation algorithms.

Keywords: indexing, melodic segmentation.

1 INTRODUCTION

The access to music digital libraries is usually content-based. The user provides an example that describes his information need, the system analysis the query document, extracts a number of features and compares them with the ones extracted from documents in the collection: documents that are more similar, under a given metrics, to the query document are presented to the final user in order of similarity. The main idea underlying content-based approaches is that a document can be described by a set of its features which, for most of the approaches in the literature, are based on the melody.

The research work on content-based music accessing and retrieval can be roughly divided in two categories: *on-line searching techniques*, which compute a match between the sequence representing the query and the ones representing the documents each time a new query is submitted to the system; *indexing techniques*, which extract off-line all the relevant information that is needed at retrieval time and perform the match directly between query and documents indexes. On-line techniques can be computationally expensive, because their complexity is proportional to the collection size; the work presented in [1] reports a comparison of different variants of an on-line technique, discussing their computational complexity and scalability.

Research work on off-line document indexing is usually based on text information retrieval techniques, which give high scalability but do not model possible mismatches between the query and the documents. Text retrieval techniques can be applied providing that a music document can be described by a set of lexical units, which play the same role of words in a text. The approaches proposed in the literature differ on the way document indexes are computed and how they are eventually normalized to overcome sources of mismatch between documents and queries. In [2] work, melodies have been indexed through the use of N-grams, each N-gram being a sequence of N pitch intervals, while note duration was not used as a content descriptor. Another approach to document indexing has been presented in [3], where indexing has been carried out by automatically highlighting lexical units using an automatic segmentation algorithm based on music theory. Units could undergo a number of different normalization, from the complete information of pitch intervals and duration to the simple melodic profile. Melodic and rhythmic patterns have been used in [4] as lexical units. The computation has been carried out without using knowledge on music structure or cognition. Separate indexes has been computed for melodic and rhythmic patterns, using a data fusion approach for merging results.

This paper reports a study on the effects of different segmentation techniques aimed at off-line document indexing. Four algorithms based on different approaches have been tested tested using a set of real queries and a collection of documents of popular music from which the melodies have been automatically extracted.

2 APPROACHES TO MELODIC SEGMENTATION

Music, both in acoustic and in notated form, is a continuous flow of events without explicit separators. It is therefore necessary to automatically detect the lexical units of a music document to be used as content descriptors to build the index of the collection. Different strategies to melodic segmentation can be applied, each one focusing on particular aspects of melodic information.

2.1 Segmentations Based on Document Content

A simple segmentation approach consists in the extraction from a melody of all the subsequences of exactly N notes, called N-grams (NG approach). N-grams may overlap, because no assumption is made on the possible starting point of a theme, neither on the possible repetitions of relevant music passages. The idea underlying this approach is that the effect of musically irrelevant N-grams will be compensated by the presence of all the relevant ones. It may be advisable to choose small values for N, in order to increase recall, which is considered more significant than the subsequent lowering in terms of precision.

Segmentation can be performed considering that typical passages of a given melody tend to be repeated many times, because of the presence of different choruses in the score or of the use of similar melodic material. Each sequence that is repeated at least K times can be used for the description of a music document. This approach based on the analysis of repetitions (AR) is an extension NG, because AR units can be of any length, with the limitation that they have to be repeated inside the melody. Segments can be truncated by applying a given threshold, because it is unlikely that a user will remember long sequences of notes.

2.2 Segmentations Based on A Priori Knowledge

Melodies can be segmented by exploiting a priori knowledge on the music domain. For instance, accordingly to theories on human perception, listeners have the ability to segment the unstructured auditory stream into smaller units, which may correspond to melodic phrases or motifs. It is likely that perceptually based (PB) units are good descriptors of a document content, because they capture melodic information that appears to be relevant for users. Even if the ability of segmenting music may vary depending on the level of musical training, a number of strategies can be generalized for all listeners. Computational approaches have been proposed in the literature for the automatic emulation of listeners behavior [5]. PB units do not overlap and are based on information on note pitch and duration of monophonic melodies.

Another approach to segmentation is based on knowledge on music theory, in particular for classical music. According with music theorists, music is based on the combination of musical structures [6], which can be inferred by applying a number of rules. It is likely that a musicologically oriented (MO) approach can be extended also to less structured music, like popular music. MO units should be computed using the complete score, but comparable results have been obtained by an algorithm that exploits local information only [7]. Structures may overlap in principle, but the current implementations do not take into account this possibility.

3 QUERY AND DOCUMENT PROCESSING

Content-based retrieval requires that both documents and queries are processed correspondingly to compute a measure of similarity between them. As a first step, the melody has been automatically extracted from documents, using a classification technique based on the Nearest Neighbor approach. A set of 50 documents have been used as training set, while 50 more documents have been used to validate the results. Automatic extraction of melodies gave a rate of correct classifications of 80% when the first nearest neighbor was used, which increased to 100% for the training set and 92% for the validation set with three nearest neighbors. The use of three neighbors increases the number of false alarms, but this has the only side effect that more than one melodic line per document is indexed.

The extracted melodies have been segmented using four different algorithms, which implement the four approaches to segmentations presented in the previous section. It is important to note that the tested algorithms are only particular implementations of general approaches to segmentation, and the experimental results may be different if other implementations had been used. Nevertheless, we believe that results may show a general trend of the relationship between the segmentation technique and the retrieval effectiveness. Several tests have been carried for each algorithm to evaluate experimentally the best configuration of its parameters. Results have been obtained with an optimal configuration for each algorithm, which is related to the particular collection used for the experiments.

The information on pitch and timing has been processed for all the segmented melodies, by considering the pitch intervals and the ratio of interonset intervals. Both dimensions have been quantized. In particular, the pitch intervals have been divided in 7 classes: unison and ascending or descending intervals within a major third, interval within a major sixth, and interval over a minor seventh. The ratios of note durations have been uniformly quantized in 9 classes, from a ratio lower than $1/3$ to a ratio higher than 3. These choices of quantizations are the ones that gave the best results for all the algorithms.

Query processing poses additional problems. Queries normally are audio recordings of users singing melodic excerpts, and have to be translated in a suitable form to be matched with documents. Any automatic transcription introduces a number of errors in pitch and onset detection that may affect the retrieval performances. Moreover, it is not guaranteed that the user will sing in the same key of the original melody, or that the tempo will be comparable to the original one. Finally, queries are error prone, because the user cannot remember exactly the excerpt – since the user is using a search engine, this is usually the case – and because untrained singers may make mistakes in pitch contour or may introduce tempo fluctuations. In our experiments, we used a set of 36 queries provided by the MAMI - Musical Audio Mining project [8]. The transcriptions of the audio queries, in the format of pitch in Hertz plus onset and offset times, are already available together with the original queries and a description of how the recordings have been carried out [9]. The use of the MAMI material allows for experimenting the effect of different segmentation techniques independently from the particular pitch tracker used to transcribe the queries.

The same segmentation processing cannot be directly applied to queries. In fact, it has to be considered that users normally provide short examples, which do not allow for the computation of patterns, and query boundaries may not correspond to patterns, perceptual units, or musicological structures. With the aim of partially overcome this problem, the extraction of lexical units from queries has been carried out taking all possible sequences of notes. Most of these sequences will be musically irrelevant, giving no contribution to the similarity computation. On the other hand, this exhaustive approach guarantees that all possible lexical units are used to query the system. We tested also the application of PB and MO algorithms also to queries, in order to segment the query consistently with documents, but the experimental results did not show any improvement on the retrieval effectiveness. After segmentation, the queries have been quantized using the same approach applied to documents.

4 EXPERIMENTAL COMPARISON OF SEGMENTATION TECHNIQUES

A music test collection of popular music has been created using 1004 documents in MIDI format; the melodies automatically extracted by documents had an average length of 372.6 notes, ranging from 124 up to 1407 notes. Queries have been added to the collection by downloading 36 annotated queries from the MAMI Web site [9]; average query length was 38.2 notes, from 5 to a very long query of 79 notes. All the segmentation algorithms have been tested with the same retrieval engine, which was based on the Vector Space Model, using the classical $tf \cdot idf$ measure to compute the similarity between the query and the documents. The retrieval engine has been extensively tested and evaluated in previous work, using different collections and sets of automatically created queries: The evaluation showed that the performances of the retrieval engine are comparable to the ones of other systems described in the literature.

Experimental results presented for NG have been obtained with N-grams of three notes, while AR has been computed applying a threshold of five notes. PB and MO performances have been computed using the algorithms presented in [10], where the *clangs* of PB [5] have been used as lexical units and all the local maxima have been used as evidence of a boundary in MO [7]. These choices are the ones that gave the best performances for the four algorithms.

Table 1 shows the main characteristics of lexical units extracted by the four algorithms. As it can be seen, AR has an index size four/five times bigger than the other algorithms, because of the high overlapping among segments of different lengths, requiring more storage and longer access times. On the other hand, the four algorithms gave comparable results in terms of average length of segments. According to a study on manual segmentation [11], all the approaches had the tendency to oversegment melodies.

The average number of unique segments per document, and the average number of documents that have a particular segment are also shown in Table 1. Consistently with results on index size, AR had the highest number of unique segments per document, while the other three algorithms had similar values. It is interesting to note that the number of segments which are present in different documents is quite low for all the algorithms. This result may suggest that a weighting scheme based on the $tf \cdot idf$ measure may not be completely suitable for a music information retrieval task, at least because the inverse document frequency (idf component) will give a small contribution to the final ranking of relevant documents. We carried out a number of tests using only the term frequency (tf component) in the weighing scheme and we found that all the algorithms had only a small decrease in their retrieval performances.

Table 1: Index size and average values for the different segmentations algorithms.

	NG	AR	PB	MO
Index size	$21,620$	$105,093$	$25,385$	$23,047$
Seg length	3.0	4.1	5.7	4.4
Seg/Doc	53.0	159.6	40.1	43.8
Doc/Seg	2.5	1.5	1.6	1.9

The results in terms of retrieval effectiveness are presented in Table 2, which reports the percentage queries that gave the correct document within the first k positions (with $k \in \{1, 5, 10, 20\}$), and the ones that did not retrieve the relevant document ("not found"), as representative measures.

The success rate of presenting the correct document at top rank was quite low for all the algorithms. For AR and NG, which gave the best results, respectively only 15.1% and 12.6 of the queries gave the corresponding document at top rank, while for PB and MO the percentages drop to 5.6 and 2.8. Analyzing the results at different levels of retrieved documents, it can be seen that for AR the correct document was presented within the first five documents in 50% of the queries, and that more than $3/4$ of the queries presented it within the first 20 documents. NG had almost comparable performances, which slightly lower values. PB and MO algorithms gave poorer results, with PB better than MO for each measured level.

The number of documents that are not retrieved at all is much lower with NG and AR approach than with PB and MO, as it can be seen from the last row of Table 2. It is interesting to note that PB had an higher number of unretrieved documents than MO, showing that the better results in terms of precision are paired to poorer results in terms of recall.

From these results, it seems that the presence of overlapping segments – as for NG and AR algorithms – can improve retrieval effectiveness. Moreover, the increase of the index size, due to a small overlap between documents and a high number of unique segments per document – as for AR algorithm – can give a further improvement. More complex approaches to segmentation did not seem to compete with simpler ones, even if it has to be noted that these results may be biased by the particular implementations of the corresponding algorithms, which had not been developed for a music information retrieval task but for musicological studies [10].

Another measure that can show the performances of

Table 2: Retrieval effectiveness.

	NG	AR	PB	MO
$= 1$	12.6	15.1	5.6	2.8
≤ 5	38.5	50.0	16.7	6.9
≤ 10	62.3	72.2	16.7	13.9
≤ 20	69.0	77.8	25.0	16.7
not found	2.8	2.8	33.3	23.6

the different techniques is the average precision, which is directly computed from the rank of the correct document. Average precision is related to the values presented in Table 2, and in fact from the first row of Table 3 it can be seen that AR gave better results than NG. The values of PB and MO are lower, yet they are somehow biased by the number of unretrieved documents. For this reason, average precision has been computed also on a reduced set of queries, which is different for each of the approaches because it contains only the ones that retrieved the correct document in any position. From the second row of Table 3 it can be seen that MO has an higher average precision, showing that, when the correct document was retrieved, its rank was usually higher than for the other approaches.

Table 3: Average precision for all the queries and for a reduced set.

	NG	AR	PB	MO
Whole query set	0.29	0.31	0.21	0.12
Reduced query set	0.30	0.32	0.30	0.39

5 CONCLUSIONS

This paper presents an overview of different approaches to segmentation for extracting music lexical units. At the state of the art, it is not clear which approach is more successful for an effective description of music documents being also robust to limitations in the query content. To this end, a comparison of four different algorithms in terms of index terms size and in terms of retrieval effectiveness has been carried out using a test collection.

From these results, algorithms based on simple approaches, which also allows the presence of overlapping lexical units, seem to give better performances when compared to more complex approaches, yet this can be due to the particular implementations of the algorithms and to the repertoire of the test collection. Probably a pattern analysis approach is more suitable for popular music than approaches to segmentation created in the context of classical music. Moreover, local errors in the queries may be compensated by the overlap between segments, affecting the retrieval results only partially.

An interesting result is that, for two of the tested algorithms, the percentage of queries that did not retrieve at all the correct document is very low. Hence, the overall performances in terms of retrieval can be improved by reranking the documents using an on-line approach. This way, off-line indexing can be used to filter out great part of the document collection, and a pattern matching approach can be used on a reduced set of the collection with the aim of giving a high similarity score to the relevant documents.

ACKNOWLEDGMENTS

The work is partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

References

[1] Hu, N., Dannenberg, R.B.: A Comparison of Melodic Database Retrieval Techniques Using Sung Queries In *Proc. of the ACM/IEEE JCDL*, Portland, OR (2002) 301–307

[2] Downie, S., Nelson, M.: Evaluation of a Simple and Effective Music Information Retrieval Method In *Proc. of the ACM-SIGIR*, Athens, GR (2000) 73–80

[3] Melucci, M., Orio, N.: Combining Melody Processing and Information Retrieval Techniques: Methodology, Evaluation, and System Implementation JASIST, Wiley, Vol. 55, Issue 12 (2004) 1058–1066

[4] Neve, G., Orio, N.: Indexing and Retrieval of Music Documents through Pattern Analysis and Data Fusion Techniques In *Proc. of the ISMIR*, Barcelona, ES (2004) 216–223

[5] Tenney, J., Polansky, L.: Temporal Gestalt Perception in Music Journal of Music Theory, Vol. 24, Issue 2 (1980) 205–241

[6] Lerdhal, F., Jackendoff, R. *A Generative Theory of Tonal Music* MIT Press, Cambridge, MA (1983)

[7] Cambouropoulos, E.: Musical Rhythm: a Formal Model for Determining Local Boundaries In Leman, M. (Ed.) *Music, Gestalt, and Computing*, Springer Verlag, Berlin (1997) 277–293

[8] Lesaffre, M. et al.: The MAMI Query-By-Voice Experiment: Collecting and annotating vocal queries for music information retrieval In *Proc. of the ISMIR*, Baltimore, MA (2003) 65–71

[9] MAMI Musical Audio Mining `http://www.ipem.ugent.be/MAMI/`, visited on April 2005.

[10] Eerola, T., Toiviainen, P.: MIR in Matlab: The Midi Toolbox In *Proc. of the ISMIR*, Barcelona, ES (2004) 22–27

[11] Melucci, M., Orio, N.: Evaluating Automatic Melody Segmentation Aimed at Music Information Retrieval In *Proc. of the ACM/IEEE JCDL*, Portland, OR (2002) 310-311