COMBINING MULTILEVEL AND MULTIFEATURE REPRESENTATION TO COMPUTE MELODIC SIMILARITY

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

ABSTRACT

In the proposed approach, melodic similarity is computed as a content-based information retrieval task. To this end, the initial incipit is considered as the query in a query-by-example paradigm and the ranked list of potentially similar documents is given by the list of retrieved documents. The approach to retrieval is based on document indexing, where each document is described by alternative melodic features of note substring with different lengths. The document collection is separetely indexed for each feature. The similarity between the query and the other documents is computed only on indexes, obtaining a number of rank lists of potentially similar documents. The final result is given by the application of a data fusion technique on the single rank lists..

Keywords: Melodic similarity, Melodic segmentation, Multifeature representation

1 INTRODUCTION

The approaches for computing melodic similarity can be coarsely divided in two groups: the ones that consider the melody as a whole and the ones that segment it in smaller units that are used as content descriptors. The former group can exploit a direct modeling of possible differences between perceptually similar melodies, usually at a higher computational cost because the number of comparisons is linear with the number of documents. The latter group can take advantage of a direct comparison between content descriptors, which is efficiently implemented using ad-hoc data structures to match the small units, but it has the drawback that its effectiveness heavily depends on the kind of segmentation that is applied to the melody.

The algorithms presented in this paper belongs to the second group. Melodic segmentation is exploited to extract a number of content descriptors, which are used to index the melodies. Different segmentations and different sets of features are exploited to describe each document with a number of sequences of symbols, taken from a finite alphabet. The distances between the query document and the other documents in the collection are then computed in a text retrieval fashion and by applying data fusion techniques to merge individual indexes.

2 THE APPROACH

A typical task of a Music Information Retrieval (MIR) system, based on the query-by-humming paradigm, is to retrieve all the documents that are instances of a particular music work, starting from an approximate excerpt of the that work. The excerpt is normally provided by the user as a short monophonic melody, sung or played. All the documents that are not instances of the searched work are usually considered irrelevant for the retrieval task, even if they are perceptually similar to it. For this reason, most of the approaches to content-based MIR consider the query as a substring of the main melody of the document, with possible local mismatches due to user's mistakes in performing the query, to imprecisions in the query transcription, and to errors in the documents.

The task of retrieving documents being similar to a query document can be considered a variant of the more general query-by-example approach. In this case, the query is exactly in the same form of the documents, and the differences between the query and the documents are not due to random errors, but to the fact that documents are expected to be instances of different music works. Notwithstanding these differences, the approach proposed in this paper is to extend the results of previous research work by Neve and Orio (2004) on query-by-humming MIR to a melodic similarity task.

2.1 Document Segmentation

The first step consists in segmenting the original melodies to obtain a set of short note sequences, to be used as document descriptors. Document segmentation is based on a simple N-grams approach. That is, the melody is considered as a sequence of events, and it is segmented by extracting all its subsequences of exactly N events. The idea of describing music documents through N-grams is quite popular in the research area of MIR – see for instance Downie and Nelson (2000) or Uitdenbogerd and Zobel (1998), yet other approaches have been applied to the segmentation task in order to take into account also some a priori information about the music language, as for Melucci and Orio (2004).

In the N-gram approach the choice of the size of the N-grams, that is the value of N, is crucial. From an information retrieval point of view, low values of N are expected to give high recall, because short note sequences are more

probably shared by similar documents than longer sequences. As usual, the increase in recall can have the drawback of a decrease in precision, because also very different documents may share the same short excerpts. On the other hand, high values of N are expected to increase precision at the cost of a lowering in recall, especially for local mismatches between the query document and the other documents in the collection.

A set of experiments, carried out on a collection of monophonic melodies, highlighted that the two drawbacks can be partially compensated by using a *multilevel* segmentation. That is, the documents are segmented by using N-grams for different values of N. Each value gives a different degree of overlapping between documents, that is a different number of excerpts shared by documents: the shorter the segment the higher the overlap. Different combinations have been tested with the training set provided by MIREX, in order to compute experimentally the optimal segmentation for the melodic similarity task. Given that the collection is made of music incipits, the maximum value of N has been bounded by the minimum length of documents in the experimental collection.

2.2 Melodic Features

The choice of which features describe a melody is as important as the way documents are segmented. As usual, there is a tradeoff between precision and recall, which can be tuned by different implementation choices. In particular, a complex combination of features, which corresponds to a close representation of documents content, is expected to improve precision, while simple features, which give a more vague and general description of documents content, are expected to improve recall.

It has been chosen to increase the number of alternative representations rather than choosing a single one, that is to have a *multifeature* representation of music documents. Preliminary experiments showed that good results can be obtained by using respectively: only the interonset interval (IOI), only the pitch contour or delta pitch (DP), and the combination of both features. In this way, documents are alternatively described by their rhythmic information only, by the pure melodic profile, and by the complete melodic information. The approach can be extended using additional features, such as the local tonality center or the relative frequence of pitches.

Both duration and pitch information can be quantized in a fixed number of classes. For instance, DP can be assigned to a number of interval classes, from a coarse representation such as unison, ascending and descending – or Same, Up, and Down as proposed by Ghias et al. (1995) – to a fine grained representation such as unison, ascending second, ascending third, and so on. It is also possible to not perform any quantization at all, and to take into account the complete information about the number of semitones in the interval. Similar considerations apply to the information about IOI and on IOI ratio. Also in this case, the optimal configuration has been computed experimentally, using the training set available by MIREX. Tests showed that the best results in terms of retrieval effectiveness are obtained without applying any quantization, apart from clipping the maximum DP to an octave, and the maximum IOI to a whole measure.

Because the number of different symbols is known in advance, each N-gram of a given feature can be described in a compact way using a text-like representation. It can be noted that quantization allows for reducing the number of symbols, with a decrease in computational complexity that can be relevant for a fully functional system.

2.3 Computing Similarity

After that each document has been represented by its multilevel N-grams of multiple features, the melodic similarity is computed using a classical text information retrieval approach. In particular, all the documents are indexed using their segments as content descriptors. Document indexing is usually carried out to speed up retrieval, because indexing is performed off-line and then the similarity between the query and the documents is carried out only on indexes. It can be noted that, for the particular task of melodic similarity in MIREX, document indexing may not give advantages in terms of computing time, because the increase due to the time taken for creating the index is not compensated by the speed up at retrieval since there were only 11 query documents for the complete task.

The N-gram approach to segmentation may give a high number of segments that have little or no musical meaning. Moreover, some segments including scales, repeated notes, or similar musical gestures, are likely to appear in many documents and hence to be poor discriminants among documents. In general, the degree by which a segment is a good index may vary depending on the segment and on the document. This is a typical situation of textual information retrieval, where different words may describe a document to a different extent. For this reason it is proposed to apply the classical $tf \cdot idf$ measure, which can be found in Baeza-Yates and Ribeiro-Neto (1999). A document is described by a sparse array, where each element is associated to a different pattern in the collection. The value of each element is given by the $tf \cdot idf$ value, that is the number of times a segment appears in the document (the tf term) is multiplied by the inverse of the fraction of documents that contain that segment, computed in log scale (the idf term).

The index is built as an inverted file, where each term of the vocabulary is a different segment. Each entry in the inverted file corresponds to a different segment, and can efficiently be computed in an expected time $O(1)$ with an hashing function. Given the multilevel and multifeature representation, a number of inverted files are built. Inverted files can be efficiently stored in memory, eventually using compression, and accessed at retrieval time – see Baeza-Yates and Ribeiro-Neto (1999) for details in the implementation. The size of the inverted file and the implementation of the hashing function depend on the number of different segments of the complete collection.

The melodic similarity is carried out using the Vector Space Model, which widely used in text retrieval systems, that allows for computing the distance between the vector of N-grams representing the query document and the vector of N-grams representing each document. Given

a query document, for each document in the collection a Retrieval Status Value (RSV) is calculated, the higher the RSV, the closer the document with the query. A rank list of potentially relevant documents is computed from each RSV, obtaining a number of lists equal to the number of levels multiplied by the number of features.

2.4 Merging the Individual Results

The similarity between couple of documents can be considered as a multidimensional vector, which dimensions depend on the number of the N-grams lengths and the number of used features. In principle, each dimension can give a different rank list, where the relative order of the documents may differ. For instance two documents may have many 3-grams with exactly the same IOI but totally different 4-grams of DP.

In the proposed approach, the individual results are merged using a data fusion technique, as described in Lee (1997), which is usually exploited by meta search engines. In particular the different similarity scores are averaged, obtaining a fused ranked list of similar documents. The fusion is carried out using directly the RSVs of each element in the ranked list, with equal weights. The choice of assigning equal weights to all the levels and all the features is due to previous results on the performances of a MIR system – see Neve and Orio (2004) – even if the tests have been carried out on a query-by-humming paradigm, which is quite different from a similarity task.

Experimental evaluation, with a query by query analysis, showed that data fusion gave a consistent improvement of the retrieval effectiveness in comparison with individual rank lists. In particular, for each of the query documents, the fused rank list always had higher precision than the one of the best single list.

3 RESULTS

Before submitting the algorithms to MIREX, a number of experiments have been carried out on the training set, which consisted of 11 query documents and 582 candidate documents in the collection. This preliminary evaluation allowed us to set the optimal configuration of the parameters, in terms of lengths of N-grams and quantization of the features. A preliminary test has also been carried out comparing alternative weighting schemes for the data fusion step, showing that the sum of the individual RSVs with equal weights gave the best performances.

The retrieval effectiveness has been tested using the *Average Dynamic Recall* (ADR), which is the measure used for the final comparison of the algorithms, together with another well-known measure in the information retrieval community, the *Average Precision* (Av.Prec). It turned out that the two measures gave very similar results in terms of ranking of the retrieval effectiveness when the parameters were varied.

The effect of the lengths of the segments has been tested by running 20 experiments where the maximum length MAX_l of the N-grams were varied from 4 to 8 notes. The minimum length min_l was varied accordingly from 2 to $M_l - 1$ notes. Three different levels of DP quantization have been used – with classes of $q_l \in \{10, 15, 25\}$ symbols respectively – while no quantization has been applied to IOI. The highest ADR and Av.Prec with the training set were

$$ADR = 68.25 \quad Av.Prec = 54.03$$

and they have been obtained with the following configuration of the parameters (which is the one that has been submitted to MIREX):

$$min_l = 3 \quad MAX_l = 6 \quad q_l = 25$$

It is interesting to note that, with $q_l = 25$, the highest ADR has always been obtained with $min_l = 3$ for any MAX_l, while when a coarse quantization has been applied, the highest ADR has been obtained with $min_l = 4$ for any MAX_l. This may mean that short segments become relevant when the DP representation is close to the original melody, while longer segments are needed when the DP quantization becomes coarser.

The results with the training set were slightly better than the ones of the official runs that, with the same configuration of the parameters, have been:

$$ADR = 64.96 \quad Av.Prec = 42.96$$

It would be interesting to carry out a query by query analysis of the official runs for highlighting the motivations of the different performances, in particular if the drop in performances has been due to a general decrease in the retrieval effectiveness or to same special case.

Comparing the results with the other algorithms submitted to the task, the performances of this approach are encouraging. The algorithm placed itself at the second highest position for almost all of the retrieval measures, with 1% difference with the first one for ADR measure.

References

R. Baeza-Yates and B. Ribeiro-Neto, editors. *Modern Information Retrieval*. ACM Press, New York, NY, 1999.

S. Downie and M. Nelson. Evaluation of a simple and effective music information retrieval method. In *Proc. of the ACM-SIGIR*, pages 73–80, Athens, GR, 2000.

A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith. Query by humming: Musical information retrieval in an audio database. In *Proc of ACM Digital Libraries (DL)*, pages 231–236, New York, NY, 1995.

J. H. Lee. Analysis of multiple evidence combination. In *Proc. of ACM-SIGIR Conference*, pages 267–275, Philadelphia, USA, 1997.

M. Melucci and N. Orio. Combining melody processing and information retrieval techniques: Methodology, evaluation, and system implementatio. *JASIST, Wiley*, 55(12):1058–1066, 2004.

G. Neve and N. Orio. Indexing and retrieval of music documents through pattern analysis and data fusion techniques. In *Proc. of the ISMIR*, pages 216–223, Barcelona, ES, 2004.

S. Uitdenbogerd and J. Zobel. Manipulation of music for melody matching. In *Proc. of ACM Multimedia Conference*, page 235240, Bristol, UK, 1998.