A Comparison of Manual and Automatic Melody Segmentation

Massimo Melucci
University of Padua
Department of Information Engineering
Via Gradenigo, 6/A
35131 Padova, Italy
melo@dei.unipd.it

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

ABSTRACT

This paper reports an investigation on the effects of exploiting melodic features for automatic melody segmentation aimed at content-based music retrieval. We argue that segmentation based on melodic features is more effective than random or N-grams-based segmentation, which ignore any context. We have carried out an experiment employing experienced subjects. The manual segmentation result has been processed to detect the most probable boundaries in the melodic surface, using a probabilistic decision function. The detected boundaries have then been compared with the boundaries detected by an automatic procedure implementing an algorithm for melody segmentation, as well as by a random segmenter and by a N-gram-based segmenter. Results showed that automatic segmentation based on melodic features is closer to manual segmentation than algorithms that do not use such information.

1. INTRODUCTION

The main contribution of this paper is an investigation on the effects of exploiting melodic features for automatic melody segmentation aimed at content-based music retrieval. Melody segmentation helps detecting *boundaries* between elements of melody that highlight musical phrases, or melodic surfaces, which can be used as descriptors of the music document content. If available, segments, built using either melodic features or not, can be organized in indexes to speed up searches, without implementing any string matching-based algorithm. We argue that segmentation based on melodic features is more effective than random or N-grams-based segmentation, which ignore any musical context. We have carried out an experiment employing experienced subjects: composers, musicians, and music students. Subjects were asked to segment manually a set of 20 music scores, each subject segmenting all the music scores. The analysis on subjects' judgments shows that the subjects have segmented in a consistent way one to each other, that is with a relatively high degree of inter-segmenter consistency, by thus providing us with a quite homogeneous testbed consisting of scores and associated segments, which can be exploited for the subsequent steps of the experiments.

Manual segmentation results have been processed to detect the most probable boundaries in the melodic surface, using a probabilistic decision function. The detected boundaries have been compared with the boundaries detected by an automatic procedure implementing an algorithm for melody segmentation, as well as by a random segmenter and by a N-gram-based segmenter. The effectiveness of the algorithm, as well as of the other automatic segmenters has been evaluated computing the probability of *miss* and the probability of *false alarm*, the former being the probability that the algorithm does not insert a boundary detected by the subjects, and the latter being the probability that the algorithm inserts a boundary that has not been detected by the subjects. Since subjects typically exploit melodic features to segment melody, their own segmentation results are employed as a baseline to compare automatic segmentation algorithms. The best algorithm is the one being closest to the manual segmentation. Results showed that automatic segmentation based on melodic features is closer to manual segmentation than algorithms that do not use such information, like the ones based on random or fixed-size segments (i.e., N-grams). This means that automatic melodic features-based segmentation can be designed and implemented to provide efficient and effective semantic content-based access to music document collections.

2. BACKGROUND

2.1 Music Information Retrieval

The detection of content descriptors for text, which is at the basis of automatization of document indexing, has been less difficult than for other media, like music, because textual words are lexical units separated by non-alphanumeric characters, as regards Western languages at least. Textual token recognition is affected by little ambiguity, though some symbols, such as periods, may be interpreted either as a separator or as a token element. In contrast, music language, like other non-textual languages, lacks of those explicit separators because there is no counterpart of textual blanks or commas in music notation. However, listeners usually perceive in the music flow the presence of distinct lexical units that form musical structures, notwithstanding the absence of predefined separators. Thus, some theory on musical structures can be defined: Musicologists proposed different theories, the most relevant being those reported in [13] and [20], which imply the possibility of segmenting music to form lexical units that can be used as descriptors of music document content.

2.2 Digital Libraries

The access to digital libraries is widely spread to users of any type, who may not have a deep knowledge of music language. Among the different features that characterize music, melody seems to be the most suitable for inexperienced users. In fact, almost everybody can recognize simple melodies and perform them at least by singing or humming. Other than being the easiest perceptive feature, melody plays a central role in a wide range of works and it is the most important feature in some specific genres, e.g. folk music. Thus, the investigation of the role of melody is a necessary, yet not sufficient, step to describe the content of music works. If melody is employed to describe music content, lexical units are *melodic segments*, that is, short excerpts of the melody

which are perceived as a single musical gesture. The role being played by melodic segments may be similar to that being played by keywords in text retrieval: As textual keywords are meaningful descriptors of the semantic content of documents, melodic segments may describe the *content* of music documents. The exploitation of melodic segments for music IR implies some sort of melody segmentation processes.

2.3 Melody Segmentation

Melody segmentation has a perceptual and subjective nature. Different persons may produce different results at the same time, and a person may differently segment the same excerpt at different times. Subjectivity can also be caused by the author, who may have inserted ambiguous music structures that are then interpreted differently by listeners. The perceptual nature of melody segmentation would suggest that the "best" segmentation result would be reached through manual work. The superiority of manual segmentation can be due to the different dimensions of music, like rhythm and harmony, that interact with melody to form the content representation of a music document and that can be detected by humans. In order to reach levels of effectiveness being comparable with manual segmentation, the automatization of segmentation would imply the detection of the different dimensions of music, which is a very difficult task. Though limited to melody only, manual segmentation becomes an infeasible task for current large collections of music documents, which are currently managed by digital library systems. The latter evidence suggests the use of *automatic melody segmentation* algorithms, which allow for the automatic extraction of melodic segments from large music document collections, eventually simplifying the process by sacrificing the detection of some complex interactions between melody and other music dimensions. One of these algorithms is based on the "Local Boundaries Detection Model", reported in [3], which is at the basis of the algorithm tested in this paper.

2.4 Topic Detection

The need and usefulness of segmenting documents to extract meaningful parts to be processed individually dates back to the research in document retrieval, structuring, and topic detection. In the case of expository texts, the automatic process of detecting the topics that are addressed in a large textual document is a feasible task [2, 11, 22]. The effectiveness of these results are sometimes due to the fact that logical structure reflects semantic structure, which allows systems for exploiting the former to infer the latter, and to detect topics. Less impressive results are likely to be observed if textual documents address topics in a logical structure being different from the semantic structure. In content based-music retrieval, the main problem is due to the absence of a clear and unquestionable logical structure that would help segmentation. Indeed, the lack of an explicit structure of the melodic profile may be a cause of a possible partial agreement between human segmenters.

2.5 Related Work

Many are the research works being proposed on music information retrieval based on melody. The reader is suggested to refer to [1, 4, 7, 8, 14, 16, 25] that address music indexing and melody segmentation, to [6, 23] that address music retrieval evaluation, or to [5, 9, 10, 15, 17, 19, 21] that describe working systems.

3. EXPERIMENTS

We investigate on the effects of exploiting melodic features for automatic melody segmentation, that is whether evidence being provided by melody helps segmenting music document more effectively than procedures that are not based on melodic features. Note that our aim was not to test the effectiveness of a specific melody segmentation algorithm; the algorithm we used is an instance of the class of algorithms that can be used to segment music. Our aim was rather to investigate if there exists an algorithm that performs quite well using melodic features, and then that that algorithm can be improved to perform even better than the version we used. The experiments carried out to test that hypothesis aimed to:

1. *Preparation of a baseline*: Highlight a reference segmentation based on human judgments, to which candidate automatic procedures can be compared. The preparation of the baseline required the following steps:
   (a) *Test excerpts*: select a representative sample test excerpts to be segmented by experienced music scholars;
   (b) *Subjects*: select a set of subjects, with a background in music, who were asked to segment manually the test excerpts; comments were recorded by thus permitting us to report their behavior;
   (c) *Cluster Analysis* and *Multidimensional Scaling*: measure the degree to which the subjects performed segmentation consistently so that the baseline can be used as the reference for comparison; this measurement has been carried out using Cluster Analysis and Multidimensional Scaling;
   (d) *Boundary detection*: detect most plausible boundaries for each test excerpt; the baseline then consisted in the excerpt together with markers separating segments that most likely correspond to boundaries.
2. *Analysis of results*: Compare candidate automatic procedures and decide what procedure perform best with respect to the baseline.

In the following, we describe the experiments in detail.

3.1 Test Excerpts

One of the basic elements of the experiments is creation of a testbed. We asked an expert musicologist to select a number of music excerpts of tonal Western music, which represent a good sampling of various typologies of melodic structure. The musicologist proposed 20 excerpts of different lengths, ranging from 7 to 26 bars and from 36 to 192 notes, depending on the melodic structure and on the length of the main theme. The complete list of the music works, from which excerpts were taken, is reported in Table 3.1. (all of the excerpts are the incipit of the music works).

3.2 Subjects

Our experiments are based on the intellectual expertise of music scholars. We believe that subjects are the candidate "producers" of good music segmentation because segmentation has a perceptual and subjective nature. We asked a group of 17 subjects to segment manually the 20 excerpts. All subjects were expert musicians. Note that we asked experienced scholars to perform the segmentation task, instead of inexperienced end users. In fact, the task of detecting significant melodic segments from *scores* requires a noticeable intellectual work that needs the exploitation of knowledge on music theory, and it can be effectively performed only by them. Tests could be carried out also with inexperienced

Table 1: List of music works used for the segmentation test, with the length in bars of each excerpt.

No.	Title	Bars
	J. S. Bach	
1	Sinfonia cantata no. 186, Adagio	7
2	Orchestral Suite no. 3, Aria	6
3	Orchestral Suite no. 2, Bourreé	13
4	Chorale	26
5	Preludium n. 9, BWV 854	8
	L. Van Beethoven	
6	Symphony n. 5, 4th movement	22
7	Symphony n. 7, 1st movement	21
8	Sonata n. 14, 3rd movement	12
9	Sonata n. 7, Minuetto	17
10	Sonata n. 8, Rondò	18
	F. Chopin	
11	Ballade no. 1, op. 23	11
12	Impromptu op. 66, 2nd movement	16
13	Nouvelle Etude no. 3	21
14	Waltz no. 7	16
15	Waltz no. 9	17
	W. A. Mozart	
16	Concerto no. 1, K313	10
17	"Don Giovanni", Aria	18
18	"Le Nozze di Figaro", Aria	10
19	Sonata no. 11, K331	18
20	Sonata no. 9, K310	22

listeners if recordings of performances would have been used instead of scores, because in this case listeners would have been able to perceive boundaries in terms of local segmentation points in the music flow. The choice of using directly scores is due to the fact that each performance is an interpretation of the score, and the performer may suggest the presence of some musical phrases depending on his personal choices. On the other hand, consistency between segmenters is likely to be higher if segmentation is carried out by experienced subjects, as it could be observed from studies on inter-indexer consistency [12].

Each subject was given a package containing the 20 melodic excerpts transcribed on music sheet. There was no maximum time for returning the compiled tests. Subjects could help themselves by playing the excerpts on their instrument and correct previous choices. By directly playing the excerpt, subjects were not biased by any external interpretation that may suggests a particular segmentation. Each page had some empty lines where subjects were encouraged to add comments and explanations of their choices. Because of the absence of time constraints and the need of detailed segmentation results, collecting all the segmented excerpts from all the subjects required a couple of months.

The packages were added with instructions and motivations of the test. The major indication was about an operative definition of the musical segments they had to highlight, which were expressed as "the lexical units of melody, which we may define also as *musical gestures*, that play a similar role of words in the spoken language." Instructions suggested to use two different graphic signs to be drawn at the end of a musical phrases, a simple and a double bar respectively indicating the presence of a normal or of a strong separator between musical phrases.

3.3 Subject Behavior

The first, quite surprising, result was that more than a half of the subjects followed the given instructions only partially, though they provided us indirectly with a useful feedback. The instructions had the implicit assumption that melodic lexical units do not overlap. Some subjects, i.e. 8 out of 17 subjects, disregarded this assumption and invented a new sign – different among subjects, but with the same meaning, as could be understood by their explanations on the tests – that clearly indicated that some notes were both the last of a musical phrase and the first of the next one. For all the subjects, the eventual overlap was of only one note length. This result implies that, for these subjects, the concept of melodic contour cannot be applied, unless we take into account the fact that contours may overlap of, at least, one note. Since this result could not be ignored, we decided to deal with this new kind of marker, as described in Sections 3.5 and 3.7.

Another result is that subjects very seldom highlighted the presence of a strong boundary by drawing a double marker. The number of double markers represent the 4.5% of the overall number of markers – including also the ones used for overlapping phrases – thus preventing for a quantitative analysis of strong separators between musical phrases. This result can be partially explained considering that, in most cases, musical excerpts were too short to allow the presence of strong separators. It is likely that the musicologist who suggested which music works should be used for the test, decided to truncate the excerpt in coincidence with the first strong separator. We decided to not differentiate between double or single markers.

3.4 Algorithm

The main aim of our research work was to test the hypothesis that a melodic feature-based algorithm performs melody segmentation more effectively than algorithms that do not use melodic features. To do that, we need one algorithm as representative of the class of melodic feature-based algorithms. The algorithm we tested is based on a model due to Cambouropoulos [3], who proposed the Local Boundaries Detection Model (LBDM). The basic idea of LBDM is that a listener perceives the presence of a boundary in a melody whenever there are changes regarding the musical intervals and the note durations. Melodic boundaries are uncertain events because some listeners perceive them, while some others do not – this is in contrast with textual data, which include separators between tokens that can be detected deterministically (we are aware that sometimes additional intelligence is needed to detect words including symbols, such as dots). Because of this uncertainty, the LBDM detects boundaries by giving a weight to all the possible places where a boundary may occur. A weight represents the degree of uncertainty of the presence of the corresponding boundary. The boundaries can be detected by analyzing the weights trend: Cambouropoulos proposed that they are associated to the presence of local maxima in the weight function. We have then developed an algorithm by implementing the LBDM in terms of melodic features and rules to compute the weights. We have extended the algorithm adding four normalization levels to tune retrieval precision and recall. Specifically, the algorithm can combinatorially transpose pitches, normalize durations, normalize pitches, and remove durations. Previous experiments show that the normalization method is consistent with what one would expect since high levels of normalization produce high recall, as observed in many other information retrieval experiments. Details of the algorithm are reported in [16].

3.5 Modeling Boundaries

It is worth distinguishing between the notions of marker, boundary, and position. We define as *marker* the personal choice of a given subject of highlighting a boundary. A *boundary* is the "parameter" signaling the fact that a musi-

cal phrase is ending and the subsequent phrase is beginning; note that there might be more than one note at which the boundary might occur. A boundary need to be estimated since it is unknown a priori, and a marker is the observed symbol between two segments, which can be used to estimate boundaries. A *position* around a note can occur just before, just after, or exactly over the note; thus, a position is associated with two notes, since the position being just after a note coincides with the one just before the subsequent note.

A preliminary qualitative analysis of the tests we collected from the subjects showed that:

1. Some melodic segments *overlap*, that is, subjects clearly attributed some notes both to a musical phrase and to the subsequent one, as explained in Section 3.3;

2. There was a good agreement among them in placing markers *around* notes, that is there were notes around which markers placed by different subjects concentrated, rather than at exact positions.

The presence of overlapping phrases confirms the hypothesis that melody segmentation presents a characteristic that is normally not taken into account when segmenting other media, like video or transcribed speech. It is important to note that the possibility of overlapping phrases is considered by music theorists [13] and the motivation of overlapping phrases has been reported by more than half of the subjects in their comments. The presence of overlapping segments may be related to the fact that there was an higher agreement among subjects in placing markers around notes. If a note can belong to two phrases, it is likely that some subjects will assign it to both phrases, while some others will assign it only to the first or to the second phrase.

Obviously, in general there is a lower agreement in placing markers at exact positions than around notes, since a marker placed around a note can be instantiated as two different positions. Despite the observation of this disagreement at exact positions, we believe that the notes around which a major agreement was observed are likely to be related to a boundary, independently of the exact position at which markers were placed. Indeed, many markers were inserted just *before*, just *after*, or exactly *over* (or subjects invented a different sign expressing that the note should be split in two) the note around which a boundary was very likely to occur. This means that, if there exists a note at which the subjects agree about the event "a boundary exists", they indifferently placed markers before, after, or over the note to mean the same event. Preliminary analyses showed that highly concentrated markers were well distinct from the rest of the scores where few, or no markers occurred.

With the aim of taking into account also the case when subjects perceived the presence of a boundary, but not necessarily agreed in assigning a note between two subsequent segments, we introduce a representation of subjects' choices in placing markers on the excerpts. For each excerpt e, the choices of subject s are represented by an array of weights w_{se} of size being equal to the number of possible markers positions i, that is $2N_e - 1$ where N_e is the number of notes of excerpt e and 2 refers to the fact that, apart from the last note, there are two positions at which a marker can be inserted for each note – i.e., over and after. The following weighting rules, which take into account the local effect of a marker, are applied:

$$w_{se}(i) \quad = \quad \begin{cases} 1 & \text{if a marker at } i-1,\ i, \text{ or } i+1 \\ 0 & \text{otherwise} \end{cases}$$

The weighing scheme is quite simple and basically states that a marker implies the maximum weight, whereas the absence implies the null weight. The choice of a binary weight is basically caused by the experimental evidence we collected that suggested that there was no preference among observed position, that is all positions are equally important.

3.6 Statistical Analyses

To have a baseline to which compare algorithms, we have assessed the consistency among subjects, under the reasonable assumption that the more the subjects are consistent one to each other in placing markers, the more the eventual baseline is a good reference for comparison. We have also compared the consistency between the LBDM algorithm and the subjects to have a preliminary idea on the performance of the algorithm itself. To have a detailed representation of the consistency between subjects and algorithm, we have employed different graphical and numerical tools that can provide a measure of the degree to which two subjects (or a subject and the algorithm) are "close" one to each other in segmenting the set of test excerpts. The numerical measure can then be tested to assess its statistical significance, while the corresponding graphical tool provides visual representation. Starting from the vectorial representation of excerpts and from the scheme used to weigh markers, a symmetric matrix of distances $\mathbf{D}_e$ between pairs of subjects s and t for each excerpt e, was calculated according to the formula:

$$D_e(s,t) = 1 - \frac{w_{se}^T \cdot w_{te}}{\|w_{se}\| \ \|w_{te}\|} \quad \text{with} \quad \|w_{xe}\| = \sqrt{\sum_{i=1}^{N_e} w_{xe}^2(i)}$$

Hence, as in usual distances based on the cosines of the vectors, $D_e(s,t) = 0$ means that judgments of subjects s and t are perfectly equal and $D_e(s,t) = 1$ means that the two subjects did not draw any common marker. We apply the same weighting scheme to the choices made by the algorithm, which is considered as subject $s = 18$. One of the reasons why we employed the cosine was caused by design: cosine can be applied indifferently on diverse weighing schemes, so it can be reused once experimental evidence will suggest that positions should be weighted differently, and then w_{se} will have non-binary values.

Two statistical techniques were employed to analyze the relationships between the subjects and algorithm. Cluster Analysis (CA) was carried out on any single excerpt using $\mathbf{D}_e$ as input. CA allows to have a pictorial representation, i.e. dendrograms that depict how subjects and algorithm are grouped within the vector space given by excerpts. Multidimensional Scaling (MDS) has been performed using $\mathbf{D}_e$ as input to produce a graphical description of closeness between subjects and the algorithm. Results are reported in Section 4.

3.7 Boundary Detection

As stressed above, the preparation of the baseline is one of the fundamental steps of the experiments, since the goodness of the comparison of the algorithms with the baseline depends on the goodness of the baseline itself. We assumed that the more the subjects placed markers around a note, the higher the probability that a boundary exists at that note. This means that, after subjects have segmented scores, it is necessary to "normalize" them and to detect the most

plausible boundaries. To this end, we extend the previously described weighting scheme to a stochastic model. This extension is due to the intrinsic uncertainty of the occurrence of boundaries. As a boundary is unknown and can be only observed through markers, the decision as to whether it exists, given a series of observation, is affected by uncertainty.

Let us start to model what is observed, i.e. markers, and $Y_i = (Y_{i,0}, Y_{i,1}, Y_{i,2})$ be a random variable describing the possible 2^3 outcomes after inserting one, two, or three marker around note i. Specifically:

$$Y_{i,i+\delta} = \begin{cases} 1 & \text{if a marker is at position } i+\delta-1 \\ 0 & \text{otherwise} \end{cases}$$

where $\delta \in \{0, 1, 2\}$. The choice of using a three-variate random variable, rather that a n-variate one, provided n is the number of notes, is caused by the empirical observations yielded by the tests. Indeed, almost all the markers signaling a highly probable boundary differed in 0, 1 or 2 positions, as pointed out in Section 3.5, while all the others were negligible.

Since we have extended the weighting scheme to a stochastic model, Y is associated to a probability distribution. The distribution of probability of Y is defined as:

$$Pr(Y_{i0} = y_{i0}, Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) = \prod_{j=0}^{2} p_{ij}^{y_{ij}} (1 - p_{ij})^{y_{ij}}$$

where $y = (y_{i0}, y_{i1}, y_{i2})$ is an outcome of three binomial independently distributed variables; specifically, $y_{ij} = 1$ if and only if a marker was inserted at position $j-1$ around note i and p_{ij} is the probability that $y_{ij} = 1$. Independence of the three random variables is a simplification due to computational and statistical reasons: if assumed, dependence would implied the estimation of a much larger number of parameters, by thus deteriorating the goodness of the actual estimation and then of the experiments.

From the manual segmentation results, it can be observed that the great majority of outcomes relate to the presence of just one marker occurring at either after, before, or over a note, whereas two or three markers rarely were put by the same subject around the same note. Therefore, the event that a boundary exists at a note is likely to correspond to the event that one marker exists around the note. Thus, let R_i be a random variable being defined as follows:

$$R_i = \begin{cases} 1 & \text{if there is a boundary at note } i \text{ if and only if there is at least one marker around } i \\ 0 & \text{if there is are no boundaries at note } i \text{ if and only if there are no markers around } i \end{cases}$$

Given a set X_i of outcomes resulted from the insertion of markers around note i by N_s subjects, we accept the hypothesis that a boundary exists at note i if and only if $Pr(R_i = 1 \mid X_i) > Pr(R_i = 0 \mid X_i)$, that can expressed by the inequality $Pr(R_i = 1 \mid X_i) > \frac{1}{2}$, where $X_i = (X_{i1}, ..., X_{iN_s})$ is the set of outcomes and $X_{ij} = (X_{ij0}, X_{ij1}, X_{ij2})$ is the outcome resulted from the j-th subject. We define the probability that a boundary exists as

$$Pr(R_i = 1 \mid X_i) = \sum_{x \in A} Pr(X_i = x_i)$$

where

$$A = \begin{cases} (0,0,1) & (0,1,0) & (1,0,0) & \\ (0,1,1) & (1,0,1) & (1,1,0) & (1,1,1) \end{cases}$$

is the set of outcomes corresponding to $R_i = 1$. The computation of $Pr(R_i \mid X_i)$ requires the estimation of p_{i0}, p_{i1}, p_{i2} that can be obtained from the likelihood function, that is:

$$Pr(X_i = x_i) = \prod_{j=1}^{N_s} \prod_{k=0}^{2} p_{ik}^{x_{ijk}} (1 - p_{ik})^{x_{ijk}}$$

where $x_{ijh} = 1$ if and only if subject j inserted a marker at position $i+h-1$, under the assumption that subjects marked the score independently one of each other. The maximum likelihood parameter estimators are given by:

$$\hat{p}_{ih} = \frac{\sum_{j=1}^{N_s} x_{ijh}}{N_s}$$

We decide that a boundary exists at note i if and only if $\widehat{Pr}(R_i = 1 \mid X_i) > \frac{1}{2}$ where

$$\widehat{Pr}(R_i = 1 \mid X_i) = \sum_{x \in A} \widehat{Pr}(X_i = x_i)$$

and

$$\widehat{Pr}(X_i = x_i) = \prod_{j=1}^{N_s} \prod_{k=0}^{2} \hat{p}_{ik}^{x_{ijk}} (1 - \hat{p}_{ik})^{x_{ijk}}$$

Therefore, the baseline of an excerpt is given by a sequence of notes such that the probability $\widehat{Pr}(R_i = 1 \mid X_i) > \frac{1}{2}$, i.e. $\widehat{Pr}(R_i = 1 \mid X_i)$ is computed for each note i and a decision is made.

It is important to note that the stochastic model used to detect the most plausible boundaries is based on the notion of likelihood, and then on the assumption that the observed markers and their frequencies is the most trusted source of evidence. Since the most plausible boundaries are those related with the highest observed frequencies, other segmentations being less frequently observed are discarded, yet they might be plausible too. However, we needed to have a model to interpret the results and to detect the boundaries in order to compare and evaluate the automatic segmenter, so we had to choose. Alternative stochastic models might have been used to describe the degree of subjectivity underlying the choice of a segmentation. For instance, a Bayesian model could integrate a prior probability distribution that could mitigate the influence of the likelihood by for example assessing all the segmentation as possible. Unfortunately, the choice of the prior probability distribution would have implied a subjective and perhaps arbitrary decision which could not be taken at this stage of the work.

4. DISCUSSION

4.1 CA and MDS

CA was carried out both on any single excerpt and on an matrix of average distances, but did not highlight the presence of clusters of subjects. This results may be due to the fact that subjects were all expert musicians, who attended the same type of music school (Italian Conservatory) even if in different cities of northern Italy. The analysis always showed only a single cluster, which regularly increased at each step, centered around subjects 3, 10, and 12. The analysis of the profile of these users did not show any particular similarity in their background; they play different instruments and they studied with different teachers. The judgments of the

segmentation algorithm were not distant from subjects' judgments, even if the algorithm never entered the cluster at the first steps of the analysis. Moreover, CA did not highlight any subject that should be ignored because of a too high distance from the others – due, for instance, to a misunderstanding of the task required by the test.

Results of CA were confirmed by the bi-dimensional plot obtained through MDS, which is depicted in Figure 1 for the distances averaged over all the 20 excerpts. As it can be seen, the judgments are spread along the plane, with the algorithm judgments (number 18) close enough to subjects' judgments. The plot shows that subjects 1 and 2 are the ones with the highest distance from the other ones. Anyway, we decided to include also them in the calculation of potential boundaries from marker positions.

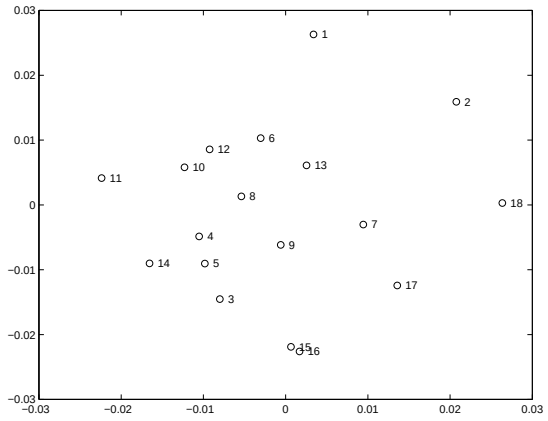

Figure 1: Multidimensional Scaling plot computed on judgments by subjects (1-17) and by the algorithm (18).

4.2 Detection of Boundaries from Markers

Results from CA and MDS shows that there is consistency among subjects choices. This means that it can be possible to extract a *baseline* from their choices and to evaluate the automatic segmentation by direct comparison with the baseline. We already introduced the statistical technique we used to compute "real boundaries" from subjective marker positions.

The number of boundaries depends on the excerpts, as well as their regularity on the score. In Figure 2.A and in Figure 2.B the frequencies of markers and the detected boundaries are reported for excerpt 1 and excerpt 9. These excerpts can be considered as representative of the subjects behavior: Even if there are positions were there is only little agreement (low values of the markers frequency), there exist good agreement at given score positions that may be considered as the baseline of boundaries.

4.3 Performances of Automatic Segmenters

The performance of automatic segmenters can be measured using a technique introduced in [2] for text segmentation in coherent segments and successively adopted in the Topic Detection and Tracking (TDT) framework [24]. To this end, it is defined the probability of agreement as:

$$P_{Agree} = \sum_{ij} D(i,j)\delta_R(i,j)\delta_A(i,j)$$

where the $\delta_Z(i,j) = 1$ if two notes i, j belong to the same segment produced by Z, and 0 otherwise, Z being A (algorithm) or R (subjects). Moreover, $D(i,j) = 1$ if the notes

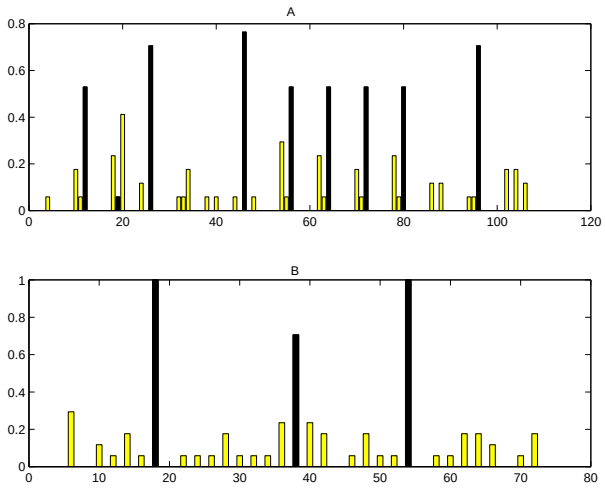

Figure 2: Histograms of markers positions, in gray, for excerpt 1 (A) and 9 (B); detected boundaries are in black.

are k notes apart and 0 otherwise; hence P_{Agree} is a measure of how often a segmentation is correct with respect to two positions at distance k. It can be shown that the complement of P_{Agree} can be computed in terms of the probability of missing a boundary P_{Miss} or placing a marker where there is no boundary P_{False}, and can be expressed as:

$$P_{Disag} = P_{Miss}P_{Seg} + P_{False}(1 - P_{Seg})$$

where P_{Seg} is the a priori probability of a segment and it can be computed from the average distance between boundaries and the choice of k. In our case, given that the average segments length is 15.04, following the guidelines of TDT framework we set k to approximatively its half, hence $k = 8$, and then $P_{Seg} = \frac{8}{15.04} = 0.532$.

Table 4.3 reports the comparison between the tested algorithm and other algorithms which are not based on melodic information. The two algorithms used for the comparison calculate respectively segments of random length (from 10 to 20) and N-grams of length 8 and 15; the former randomly selected a number r between 10 and 20 and inserted a segment after r notes, the latter inserted a marker every N notes. It can be seen that the LBDM has better performances than other techniques that do not exploit melodic information, even if the performances need improvement, in particular because of the high probability of false alarms. The algorithm has the tendency of *over-segment* the excerpts. This behavior is confirmed also by the average segment lengths, which is 15.0 positions for the subjects and 8.6 positions for the algorithm. This is the reason why we chose N-grams of length, respectively, of 15 and 8. There is no significant variation in P_{Disag} between the two N-grams, meaning that gram length does not affect the effectiveness of segmentation but it is the method to detect boundaries that actually

Table 2: Probability of misses, false alarms, and disagreement of four different segmentation algorithms.

Algorithm	P_{Miss}	P_{False}	P_{Disag}
LBDM	0.054	0.342	0.189
Random	0.653	0.304	0.476
Fixed ($N = 8$)	0.421	0.558	0.485
Fixed ($N = 15$)	0.720	0.286	0.517

improves the performances.

The tendency of over-segmenting, together with the relatively low value of P_{Miss}, may imply that many of the segments obtained by analysis of subjects are split in two different segments by the algorithm. This result may be important if the same algorithm is applied to queries, because the latter are much shorter than documents and hence query segments can be consistent with document segments.

The superiority of LBDM with respect to the N-gram and Random algorithms might seem not too surprising – indeed, it was a quite easy fact that something more "intelligent" than splitting every N or a randomly computed number of, notes were more effective. While this almost certainly true for Random, it is worth noting, however, that several research works in music IR advocated that N-grams are an effective means to index music, perhaps if retrieved through approximated string matching algorithms. The use of Random was mainly motivated by the need of having a baseline representing the worst case.

4.4 Scope of the Work

The study concentrated on the evaluation of melody segmentation because our previous work and the work by other researchers were conducted primarily on melody. As stressed above, the addressed issue is on the use of melody for segmentation, rather that on the evaluation of a specific melody segmentation algorithm, such as LBDM. We have left aside the issue of music query, which is a debated problem in music retrieval community [23]. Nevertheless, the results in automatic segmentation may apply to the query-side since queries may need to be segmented before retrieving music documents, and "good" segmentation algorithms would help represent music query effectively.

The discussion on the use of melody for segmentation, is still open and is addressed at different levels of music retrieval system design. This paper provides with useful insights about the effectiveness of melody. Looking at other features, such as timbre or rhythm may be of great value, but the studied sample of works and the number of subjects would have been much larger. It should be noted that some rhythm information can extracted from melody; however, if other features were been considered and analyzed, the interpretation of results would be confusing, especially if it would be done in relation with the interpretation of results regarding melody alone. Thus, we preferred to concentrate on one feature to isolate external factors and to provide some useful insights on melody.

Like many user studies, the sample data used to evaluate the segmentation algorithm requires a non-large dataset, while much larger ones are exploited to conduct laboratory and computer based experiments in textual document retrieval evaluation. The choice of a non-large sample is also due to the experimental design choice of asking the subjects to segment *all* the excerpts and to provide qualitative, e.g. full-text comments on their own segmentation decisions, which has been exploited during the process of result analysis. Moreover, subjects were asked to play several times the melodies so that the decision regarding where to place boundaries was as more founded and definitive as possible. As consequence, the value of the dataset lies in the quite high number of different segmentation being placed at each melody by every subject, and in the availability of comments and suggestions that have explicitly been provided by the subjects.

The algorithms over-segments melodies. While this is interesting in its own right, there would not be implications for retrieval effectiveness. If the same algorithm is used to segment both the documents and the user-query, then both documents and queries will be over-segmented, but since the indexing and retrieval sub-systems are consistent with each other, this may not be a major problem as far as retrieval effectiveness is concerned.

5. FUTURE WORK

Further study can be conducted in the future to evaluate the role of timbre, rhythm or, harmony on segmentation, in similar way to that used in this paper. However, robust methods for their segmentation aimed at music retrieval have to be addressed. We will investigate on the effect of considering melodic features in query segmentation, since music queries are more likely to be affected by errors and are typically much shorter than documents. In this work, we have observed that there is a good agreement among subjects if marker positions around notes are considered instead of exact positions. This peculiarity of melody segmentation has to be considered in the design of indexing techniques of segmented melodies, which is an aspect beyond the aims of this study and that will be addressed in the future.

6. REFERENCES

[1] D. Bainbridge, C.G. Nevill-Manning, I.H. Witten, L.A. Smith, and R.J. McNab. Towards a digital library of popular music. In *Proceedings of ACM Digital Libraries (DL) Conference*, pages 161–169, Berkeley, CA, August 1999.

[2] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, special issue on Natural Language Learning, C. Cardie and R. Mooney, editors, 34(1-3), pages 177–210, 1999.

[3] E. Cambouropoulos. Musical rhythm: a formal model for determining local boundaries. In E. Leman, editor, *Music, Gestalt and Computing*, pages 277–293. Springer-Verlag, Berlin, 1997.

[4] E. Cambouropoulos. The Local Boundary Detection Model (LBDM) and its Application in the Study of Expressive Timing. *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001.

[5] CANTATE. Computer Access to Notation and Text in Music Libraries, Jan. 2002. `http://www.svb.nl/project/cantate/cantate.htm`.

[6] J.S. Downie and M. Nelson. Evaluating a simple and effective music information retrieval method. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 73–80, Athens, Greece, 2000.

[7] A. Friberg, R. Bresin, L. Frydén, and J. Sunberg. Musical Punctuation on the Microlevel: Automatic Indentification and Performance of Small Melodic Units. *Journal of New Music research*, 27(3):271-292, 1998

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

[9] HARMONICA. Accompanying Action on Music Information in Libraries, Jan. 2002. `http://www.svb.nl/project/harmonica/harmonica.htm`.

[10] J. Harvell and C. Clark. Analysis of the quantitative data of system performance. Deliverable 7c, LIB-JUKEBOX/4-1049: Music Across Borders, 1996. See also `http://www.sb.aau.dk/Jukebox/edit-report-1.html`, Jan. 2002.

[11] M.A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 59–68, 1993.

[12] F.W. Lancaster and A.J. Warner. *Information Retrieval Today*. Information Resources Press, Arlington, VA, 1993.

[13] F. Lerdhal and R. Jackendoff. *A generative theory of tonal music*. MIT Press, Cambridge, MA, 1983.

[14] R.J. McNab, L.A. Smith, I.H. Witten, C.L. Henderson, and S.J. Cunningham. Towards the digital music library: Tune retrieval from acoustic input. In *DL'96: Proceedings of the 1st ACM International Conference on Digital Libraries*, Multimedia Digital Libraries, pages 11–18, 1996.

[15] Rodger J. McNab, Lloyd A. Smith, David Bainbridge, and Ian H. Witten. The New Zealand Digital Library: MELody inDEX. Technical Report may97-witten, D-Lib Magazine, May 15, 1997.

[16] M. Melucci and N. Orio. Musical information retrieval using melodic surface. In *Proceedings of ACM Digital Libraries (DL) Conference*, pages 152–160, Berkeley, CA, August 1999.

[17] M. Melucci and N. Orio. SMILE: a system for content-based musical information retrieval environments. In *Proceedings of Intelligent Multimedia Information Retrieval Systems and Management (RIAO) Conference*, pages 1246–1260, Paris, France, April 2000.

[18] M. Melucci and N. Orio. An Evaluation Study on Music Perception for Music Content-based Information Retrieval. In *Proceedings of the International Computer Music Conference*, pages 162– 165, Berlin, Germany, August 2000.

[19] Musica. The International Database of Choral Repertoire, Jan. 2002. `http://www.MusicaNet.org/`.

[20] E. Narmour. *The analysis and cognition of basic melodic structures*. University of Chicago Press, Chicago, MI, 1990.

[21] RISM. Répertoire International des Sources Musicales, Jan. 2002. `http://www.rism.harvard.edu/rism/Welcome.html`.

[22] G. Salton, A. Singhal, C. A. Buckley, and M. Mitra. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264(5164):1421–1426, 1996.

[23] E. Selfridge-Field. What motivates a musical query? In *International Symposium on Music Information Retrieval*, Plymouth, MA, 2000. `http://orange.cs.umass.edu/music2000/papers/invites/selfridge_invite.pdf`, Jan. 2002.

[24] Topic Detection and Tracking, Phase 2. Jan. 2002. `http://morph.ldc.upenn.edu/Projects/TDT2/`

[25] A. Uitdenbogerd and J. Zobel. Manipulation of music for melody matching. In *Proceedings of ACM Multimedia Conference*, pages 235–240, Bristol, UK, 1998.