Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors

Article · January 2003

CITATIONS
49

READS
113

2 authors:

Fabien Gouyon
Institute for Systems and Computer Engineering, Technology and Science (INESC …
97 PUBLICATIONS **2,428** CITATIONS
SEE PROFILE

Perfecto Herrera
Escola Superior de Música de Catalunya (ESMUC)
175 PUBLICATIONS **3,714** CITATIONS
SEE PROFILE

Audio Engineering Society

Convention Paper

Presented at the 114th Convention
2003 March 22–25 Amsterdam, The Netherlands

Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors

Fabien Gouyon, Perfecto Herrera

Music Technology Group, IUA-Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT

We address the problem of classifying polyphonic musical audio signals by their meter: the number of beats between regularly recurring accents (or downbeats). The problem is simplified to a 'duple'/'triple' decision. Experiments have been conducted on a 70 instances database (20s excerpts from pieces of music without particular genre nor timbre restriction). Our approach aims to test the hypothesis that acoustic evidences for downbeats can be measured on signal low-level features; focusing especially on their temporal recurrences. We experimented several approaches to the problem of feature selection and report some interesting results: measurements of a very small set of beat descriptors (*i.e.* 4) and subsequent processing (based on autocorrelation functions) permit to reach around 95% of correct classification. Using only the temporal centroid, almost 90% of correct classification can be achieved.

1. INTRODUCTION

1.1. Meter

The metrical structure of a musical piece is based on periodic recurrences of musical events. These periodicities involve two (or more) temporal scales, or levels, that show integer ratios. Along this line, Cooper *et al.* (1960) define the meter as "the number of pulses between the more or less regularly recurring accents." Yeston (1976) defines it as "an outgrowth of the interaction of two distinct levels (two differently-rated strata), the faster of which provides the elements and the slower of which groups them." For Gasser *et al.* (1999), it is "an abstract structure in time based on the periodic recurrences of pulses."

However, humans accept important violations of this definition without loosing the sense of meter. First, actual musical signals never show *integer* ratios between metrical levels. Many theoretical attempts include the notion of temporal deviations to this too rigid framework, *e.g.* (Large *et al.* 1994), (McAuley 1995), (Gasser *et al.* 1999), (Honing 2001), (Large *et al.*

2002). Furthermore, in addition to the exigency to deal with *approximate* periodicities, there is a need to provide a suitable definition for the musical events ("pulses", "accents" or "downbeats") that occur periodically, as they are encountered in musical signals. Indeed, these musical events are sometimes perceived as sounded events and sometimes as unsounded time points (Large *et al.* 2002). And they *need not be precisely equivalent* (in a physical meaning) to be perceived as instances of the same musical concept.

1.2. Downbeats

Music theories define the notion of *strong* beats (or downbeats) as time points at which several metrical levels coincide (Lerdahl *et al.* 1983). Most of the attempts to determine which musical events define metrical accents entail the processing of lists of onsets, as could be derived from a score (Brown 1993), or parsed from MIDI data (or other symbolic formats). Onset features such as inter-onset intervals (IOIs), onset-offset durations, pitches, dynamics or harmonic roles are usually hypothesized as relevant cues to meter determination.

For instance, Brown (1993) proposes the hypothesis that the frequency of occurrence of notes is greater on strong metrical time points (*i.e.* where several metrical levels coincide) than elsewhere. Under this hypothesis, time differences between onsets should reveal patterns. She uses the autocorrelation method to detect periodicities of IOIs. She also acknowledges that using an additional musical feature, note durations, slightly improves results.

Another example is the approach followed by Meudic (2002). One of its aims is to extend Brown's (1993) algorithm to other musical features. Here, the input is a list of onset features (MIDI features) and time indexes of onsets that correspond to beat indexes. A method resembling Brown's (1993) is applied, not on IOIs, but rather on a sequence of "scores", one score for each beat segment. The determination of the salience of a beat segment (its score) depends upon (is the linear combination of) 5 features that are parsed from the input data: dynamics of the first note in the segment, interval between the segment pitch extrema, possible presence of a rest after the first note in the segment, duration of the first note and number of notes in the segment.

1.3. Dealing with audio

The previously commented algorithms (and many others) process symbolic inputs (lists of onset features). As argued elsewhere (Scheirer 1998), processing lists of onset features is not directly transposable to acoustic signals. In a *continuous* musical flow, how can we define musical events? And what is it that makes events (actually differing in their acoustic properties) sound like occurrences of the same, particularly salient, pulse?

A possible rationale could be to envisage a transcription of audio material into symbolic data (as *e.g.* MIDI) as a preprocessing step to meter determination. For instance, the first processing achieved in Goto's (2001) algorithm is the transcription of the percussion instruments (onset detection and bass-drum/snare-drum recognition) when dealing with percussive music signals, and a harmonic transcription (onset detection and chord recognition) when dealing with non-percussive signals.

We choose a different approach. Instead of processing onset times and conventional musical features (as would be notes, chords or specific timbres, etc.), we propose to deal with data of a lower level of abstraction. Our objective is to ground the determination of the meter of musical audio signals onto the hypothesized recurrences of low-level descriptors.

This framework raises the following issues:

1. Which are the relevant low-level features?
2. Which are the relevant temporal boundaries for the computation of these features?

1.4. Chosen approach

We propose to consider the *beat as the relevant temporal resolution* to compute descriptors. (Beat indexes being extracted in a semiautomatic manner to provide reliable input to the problem of interest here.) This is the only use of data that might be considered to entail a high level of abstraction. (Although it is still not clear whether the perception of beats entails a higher level of abstraction than that of notes –or chords–. The necessity of modeling cognitive processes for inducing beats is still object of controversy (Scheirer 1998).)

Our approach resembles Seppänen's (2001) estimation of phenomenal accents (see Section 4.3). Differences are the following: First, we focus on beat descriptors where he focuses on Tatum descriptors. Second, we do not account for onset features (as *e.g.* "onset spectrum brightness") for most onset detection methods are unreliable when dealing with unrestricted polyphonic audio. Third, we do not propose a *downbeat model, i.e.* a specification of *what values* a specific set of features should take to indicate the presence of a downbeat. Our approach is rather

qualitative than quantitative. We intend to determine *which* are the relevant features to focus on. That is, we seek the set of low-level features whose periodicities most likely correspond to specific meters. A labeled database is used to seek patterns of feature recurrences, rather than patterns of feature values.

Situating the debate in a pattern recognition framework, we simplify the problem of meter determination in restricting it to a two-class decision: *duple* (groupings of two beats) or *triple* (groupings of three beats). (See Figure 1.) Indeed, we need a way to automatically and objectively assess our results. When dealing with written music, or MIDI, a reference can be taken as the score time signature. But there is no ground truth regarding the concept of meter of audio signals. As an illustration, it is our belief that there could be endless discussions on whether the beats of a given excerpt would be better grouped by 2, 4 or 8. But there would certainly be no doubt that for this particular excerpt, 2 would be a better grouping factor than 3.

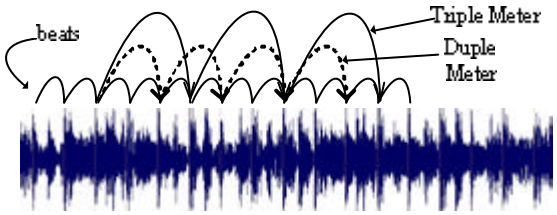

Figure 1: Illustration of the two possibilities for beat grouping: duple or triple?

In the remainder of this paper, we detail our algorithm proposal for meter determination. Then, we propose justifications for some aspects of this algorithm, giving details of experiments relative to feature selection and classification techniques. This paper ends with a discussion and considerations regarding future work.

2. METER DETERMINATION ALGORITHM

Beat indexes are extracted in a semiautomatic manner to provide reliable input to the problem of interest here (algorithms dealing with polyphonic audio signals are reported in the literature, *e.g.* (Scheirer 1998), (Dixon 2001)).

The basic steps are the following:

1. Frame descriptor computation

We set a frame size of 20 ms, and a hop size of 10 ms. On each signal frame, a few low-level features of interest are computed. Our investigations led us to use:

- f_1: Energy (see Figure 2)
- f_2: Spectral flatness, *i.e.* ratio geometric mean/arithmetic mean (for this feature, frames are multiplied by a Hamming window before DFT computation)
- f_3: Energy in upper-half of first Bark band (approximately 50-100 Hz)

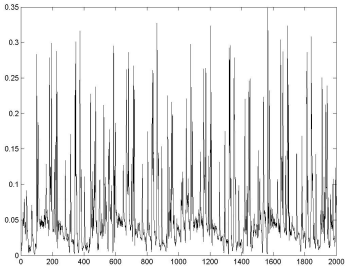

Figure 2: Evolution of f_1 over the frames of 20 seconds of "A lo Cubano" (*Orishas*, Cuban Hip-Hop)

2. Beat segment descriptor computation

Beat boundaries are matched with frame indexes. For each beat, three regions of interest are defined:

- R0: The whole beat segment, *recentered* around the beat index
- R1: The 120 ms region surrounding the beat index
- R2: The rest of the beat segment, *i.e.* $\left(R0 \cap \overline{R1}\right)$

Four beat descriptors are defined as the standard deviation of f_1 over R0, the average of f_2 and f_3 over R1, and the temporal centroid over R0 (this descriptor does not entail frame feature computation). Values of the descriptors are normalized (mean is subtracted and they are divided by the standard deviation).

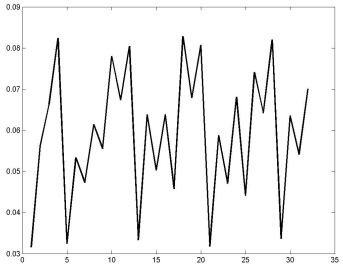

Figure 3: Evolution of the standard deviation of f_1 over R0s (same song, same temporal scale on the X-axis as Figure 2, but measured in beat indexes).

Each musical excerpt is then represented by 4 temporal sequences, whose lengths correspond to the

number of beats of this excerpt. Each sequence is the evolution of a specific descriptor over the different beat segments. (See Figure 3.)

3. Periodicity detection

The (normalized) autocorrelation is computed for each sequence as follows.

Let x be the subsequence corresponding to beat indexes 0 to I, and y the subsequence corresponding to beat indexes l to $(l+I)$.

$$acf(l) = \frac{x \cdot y}{\sqrt{\sum_{i=0}^{I}(x_i)^2}\sqrt{\sum_{i=0}^{I}(y_i)^2}}, \forall l \in \{0 \cdots U\}$$

U is the upper limit for the lag l (set to 8 beats), and I the integration time (set to 10 beats).

High peaks in a descriptor autocorrelation function indicate lags for whose this descriptor reveals recurrences along the sequence. (See Figure 4.)

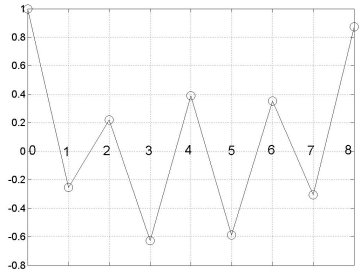

Figure 4: Autocorrelation coefficients of the sequence represented in Figure 3

4. Computation of decisional features

We specify the criterion M for making decisions regarding the 'duple' or 'triple' nature of excerpts:

$$M = \left(acf(2) + acf(4) + acf(8)/3\right) - \left(acf(3) + acf(6)/2\right)$$

M is a real number; the farther from zero in the positive values, the more it is representative of duple meters; the farther from zero in the negative values, the more it represents triple meters. There is one value of M for each descriptor. Henceforth, the relevant features for the 'duple'/'triple' decision are the values of M corresponding to each descriptor.

In the example illustrated in the figures above, relative to a single feature (the evolution of the energy standard deviation over R0s), $M = 0.6313$, the meter is effectively duple.

5. Classification

Excerpts are represented by four features: the criteria M relative to the four beat descriptors (see step 2 above). The decision regarding the meter of a test excerpt shall be taken according to the set of values for these descriptors. Deriving a class membership from a set of descriptor values can be achieved by several pattern recognition techniques. For instance, Discriminant Analysis (DA) derives *regions* of class memberships in the feature space from a statistical modeling of labeled data –in our case, by the very definition of M, the region boundaries are around zero–. This technique gave us fairly good results. Even a simple rule, relative to a single feature, seems to give error rates relatively acceptable. Namely: "For a given excerpt, if the feature M computed from the beat temporal centroids (values over R0s) is greater than -0.108046, then this excerpt has a duple meter, otherwise its meter is triple."

We report hereafter on several experiments and discuss shortly criteria to consider when designing a classifier.

3. EXPERIMENTS

The algorithm detailed above involves a series of decisions that we will intend to justify in the remainder of this paper.

Basically, once decided to use the autocorrelation as the method for periodicity detection, and once defined potential descriptors of interest (they can be numerous, we first envisaged a total of 277 signal descriptors), the issue lies in the *reduction of the dimensionality*.

Indeed, the actual problem would be to determine the importance of *each lag*, of *each autocorrelation function* (one function being relative to one descriptor), as for the classification 'duple' or 'triple'. Let N be the number of descriptors, and L the number of autocorrelation lags. Then, the dimensionality is $N \times L$.

We seek periodicities of lag 2 or 3 as reflected in the autocorrelation functions. Therefore, we propose to address a relaxed problem, a first reduction of dimensionality is achieved by specifying the criterion M previously introduced. (This criterion was motivated by the fact that the autocorrelation function of a periodic signal is periodic, of the same period.) One value of M corresponds to one descriptor; therefore, N dimensions remain.

The next reduction of dimensionality concerns the number of input features to the classification algorithm. It must be reduced, first, because of

measurement computational and memory costs, second, to ease the understanding of the data; and third, it should also be reduced to avoid the "curse of dimensionality": if the number of training samples that are used to design a classifier is small relative to the number of features, adding features may actually degrade the performance of the classifier. The ratio of sample size to dimensionality should be at least 10 (at least 10 times more training samples than dimensions). (Jain *et al.* 2000)

In the following, we detail the initial set of features used, then experiments relative to selections among these features, and also experiments relative to classification techniques.[1]

3.1. Sounds and features

A database of 70 sounds (format 44100 Hz, 16 bit, mono) was used for experiments. Each excerpt lasts 20 seconds. Boundaries for beginnings and ends were set randomly. We intended to keep generality w.r.t. genres and timbres: excerpts are all polyphonic, the majority multitimbral (there are few monotimbral excerpts: guitar or piano), and the genres are diverse (Hip-hop, Pop, Opera, Classical, Jazz, Flamenco, Latin, Hard-rock, etc.). There are 34 triple meters and 36 duple meters. 33 excerpts have drums or percussion (10 triple, 23 duple) and 37 do not (24 triple, 13 duple). The list of songs can be found on the first author's web page.

The four features detailed in Section 2 are a subset of an initial group of 277 features.

46 features have been computed on a frame-by-frame basis: energy, zero-crossing rate, spectral centroid, spectral kurtosis, spectral skewness, two measures of the spectral flatness (one is the ratio geometric mean/arithmetic mean and the other is the ratio harmonic mean/arithmetic mean), 13 Mel-Frequency Cepstrum Coefficients (MFCCs) and the energy in 26 non-overlapping spectral bands, as defined by a Bark decomposition (the first two traditional Bark bands are divided in two subbands).

Beat descriptors have been designed as statistics of frame features over the 3 regions mentioned above (R0, R1 and R2). Means over R0, means over R1, means over R2, standard deviations over R0, standard deviations over R2 and ratios of means over R2 and R1 define 276 descriptors. The 277th beat descriptor is the temporal centroid computed over R0. Descriptor values are normalized. At that point, a musical excerpt is represented by 277 temporal sequences, which length corresponds to its number of beats. Each sequence is the evolution of a specific descriptor over the different beat segments.

As detailed in Section 2, a feature M is computed for each sequence. Hence, each excerpt is represented by 277 scalar features.

3.2. Feature selection

3.2.1. Background

Feature *selection* techniques differ from feature *extraction* techniques (as *e.g.* Principal Component Analysis). The latter *create* new features by transforming the original set of features (by linear or non-linear combinations), where the former *select* a subset of features among the original set. (Jain *et al.* 2000)

When a set of labeled instances is available, we can use *supervised* feature selection, otherwise *unsupervised* feature selection may be appropriate. In both cases, the key issue resides in defining a criterion for evaluating the relevance of a given feature subset.

Liu *et al.* (1998) discuss different relevance criteria to be applied when selecting features: inter-class distances, association or dependency, entropy or information, consistency, divergence and precision. The many combinations of features define a space to be searched. Exhaustive search being impractical for reasons of computational cost, different feature selection techniques can be defined by combining one or more criteria with a specific search strategy (evaluating growing feature sets, shrinking feature sets, one feature at a time, or more complex schemes – "plus n, minus m"–, using heuristics, etc.).

We can also consider different approaches to feature selection, according to the relationships between the feature selection and the induction (*i.e.* class learning) algorithm that is used:

- *Embedding* is used when the induction algorithm incorporates in its own machinery some feature selection operation. For example, Discriminant Analysis (DA) uses an F-statistic that is computed in connection with other necessary indexes and matrices data. Selection of features is here a process that is "naturally" embedded in the own DA algorithm.
- *Filtering* considers the feature selection as a pre-processing operation that can be performed independently from the induction

[1] Experiments have been done using the commercial software Systat (http://www.systat.com/), the open-source software Weka (http://www.cs.waikato.ac.nz/~ml/) and some implementations by the authors.

algorithm. This way all the irrelevant features are discarded before the application of the induction algorithm. Some filtering approaches yield (and evaluate as a whole) subsets of features whereas others yield ranked lists, according to the selection criteria.

- *Wrapping* connects feature selection with the induction algorithm in such a way that the accuracy of the feature subsets is evaluated according to their results with the induction algorithm. It can be similar to embedding by the fact that selection and induction are explicitly connected, and it can be also similar to filtering by the fact that the feature selection technique is not intimately related with the induction technique. The key point is in the way of doing the evaluation of features, as here it depends not only in the specific selection technique but also it is mediated by the induction technique selected by the experimenter. The drawback of this approach is the extreme slowness (which increases as the number of features increases), because the induction algorithm is called for every evaluation operation of the feature candidates.

3.2.2. Experiments

We first tried an unsupervised technique. The goal in this framework is to filter out "similar", or "redundant" descriptors, without considering class memberships. We implemented two measures of feature "similarity": correlation between features and the "maximal information compression index" proposed in (Mitra *et al.* 2002).

Given a measure of similarity between features, features can be grouped into *clusters* (by *e.g.* hierarchical clustering, k-means clustering, etc.). (See Figure 5.) Selecting relevant features requires a subsequent step: choosing a *representative feature* for each cluster. We implemented the algorithm proposed by Mitra *et al.* (2002) for this task (based on the k-NN principle). For instance, when reducing the dimensionality from 277 to 20 features with this algorithm, one of the many steps discards the means of MFCC12 over R2 and R1, their representative feature being the mean of MFCC12 over R0.

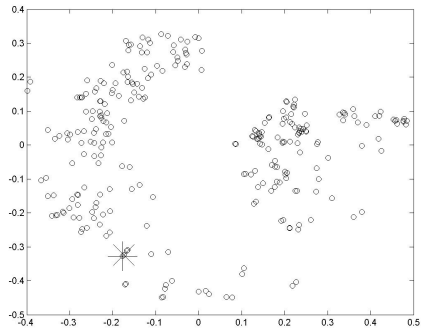

Figure 5: 2D representation of similarities between the initial 277 features (obtained by Multidimensional Scaling, similarity measure is the correlation). Each circle represents a feature. The plus sign represents the ZCR mean over R0, the cross represents the ZCR mean over R2, they are very close as one would expect.

We also tried diverse supervised techniques in the framework of the software Weka: filters that implement different relevance criteria (Correlation-based Feature Selection (CFS), Consistency, InfoGain, InfoGainRatio, SymmetricalUncertainty, OneR, ChiSquare and ReliefF), and wrappers around different induction techniques.

We will not give details here of these experiments, but their analyses showed that some features co-occurred as relevant ones with many techniques. Among them, most importantly, the temporal centroid over R0, but also the three others listed in Section 2. Figure 6 illustrates the projection of 70 songs over two of these dimensions.

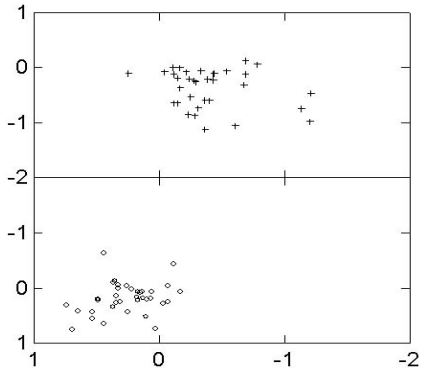

Figure 6: Plots of the 70 songs in a 2D space. The X-axis is the feature M computed from the temporal centroid over R0, the Y-axis is that computed from the standard deviation of the energy over R0. Triple meters are in the upper box, duple meters in the lower box.

Regarding the technique of Mitra *et al.* (2002), we ran experiments on both sequences of beat descriptors (before autocorrelation computations) and on features

M (after autocorrelation computations). Our conclusion is that, even though it is a much faster process than any supervised technique, it should not be exploited too far in the discarding of features. It may rather be used as a preprocessing (some sort of rough filtering) to a subsequent *informed* selection of features. Indeed, reducing the number of features by just a few may be very useful, as supervised techniques are typically very sensitive, w.r.t. time consumption, to the number of features. (Recall that they entail search procedures.)

3.3. Classification

We have tested different approaches ranging from non-parametric models (Kernel Density estimation) to parametric ones (Discriminant Analysis), and including rule induction, neural networks, 1-Nearest Neighbor (1-NN), or Support Vector Machines (SVMs). Let us report on some of these experiments.

We considered three feature subsets: (a) those kept after a CFS selection –filter method, 27 features[2]–, (b) those selected by CFS with a wrapping around the induction algorithm subsequently used for classification (usually far less than 27 features), and finally (c) the sole feature M computed from the temporal centroid values over R0s. (With all the following techniques used as wrappers for the CFS algorithm, the number of features varying from 2 to 8, it is worth noticing that the feature subset (b) always accounts for the temporal centroid.) Table 1 presents results obtained by ten-fold cross-validations (average of the error rates over 10 trials with 10% of the samples for testing and 90% for training, randomly chosen)

	(a)	(b)	(c)
Naïve Bayes	94.2	92.8	91.4
Kernel density	92.8	97.1	91.4
1-NN	92.8	88.5	88.5
Support Vector Machine	92.8	81.4	88.5
C4.5 (tree)	90	88.5	88.5
PART (rules)	82.8	88.5	88.5

Table 1: Percentage of correct classification (10-fold cross-validations)

[2] Means over R0 of MFCCs 1 and 13, spectral skewness, spectral flatness (harmonic/arithmetic) and energies in Bark bands 1, 2, 3, 19 and 20. Means over R1 of spectral flatness (geometric/arithmetic) and energies in Bark bands 2, 9, 13 and 24. Standard deviations over R0 of energies in Bark bands 6 and 13. Standard deviations over R2 of energy (whole band), spectral flatness (harmonic/arithmetic) and energies in Bark bands 1, 9 and 19. Ratio of means over R2 and R1 of MFCC 1 and energies in Bark bands 12, 18, 20 and 22. Temporal centroid over R0.

Error rates for all the cases were found to lay below 17.2% when 27 descriptors were used (the best technique –Naïve Bayes– yielded 5.8% only, where a rule induction technique yielded 17.2%). When drastically decreasing the number of features, varying between 8 to 2, and then 1, performances degrade slightly but error rates can be kept around 10%. (With the exception of the SVM which performance drops down when using 3 features, selected by CFS wrapped around SVM, and stays around 10% error with the sole temporal centroid.)

In another experiment, we performed a Discriminant Analysis with four features that co-occurred in the results of many feature selection techniques (see Section 3.2): the four features M advocated in Section 2. Ten-fold cross-validations yielded a 5.2% error rate.

4. DISCUSSION AND FUTURE WORK

In this paper, we proposed an algorithm for the determination of the meter (duple vs. triple) of musical audio signals. The main hypothesis tested is that acoustic evidences for downbeats can be measured on temporal recurrences of signal low-level features. Our investigations indicate that a small set of beat descriptors might be relevant for the task. We herein detailed experiments that led us to design a specific algorithm, based on a set of four features. However, this hypothesis, tested empirically here, should be further discussed on a theoretical ground. Indeed, we are aware that this approach might not support the widespread concept of meter as a construct without reality in the stimulus itself, an abstraction from the stimulus properties (see *e.g.* (Clarke 1999)).

We presented diverse means to select relevant features. This task is still part of our ongoing investigations regarding the understanding of the perception of the meter. One might consider the selection of relevant features as a fundamental problem, unrelated to practical applications. On the other hand, the choice of a final classification algorithm that assigns the value for the meter (duple or triple) depends on the priorities of the system one intends to build. One might seek a high precision in classification, whatever the computational or memory cost. Or, one might prefer a compact decision function, even if it looses a small amount of predictive power. This trade-off can be observed in our data: the rule induction algorithm (PART) seems to be the less successful, w.r.t. the three feature sets, however, it is the most compact and simple to implement.

Our experiments point out the relevance of the temporal centroid for the task. Let us indulge in

explanation insights. Let us propose the assumption that note occurrences have a direct correlation with increases in the waveform amplitude, and thus with the value of the temporal centroid. Then, one might hypothesize that downbeats would be characterized by patterns of note timings. That is, the main difference between downbeats and offbeats would reside in the regularity of note timing patterns: along the musical sequence, patterns of note onset times would show greater similarity on downbeats than on offbeats. This hypothesis is an extension of the widespread hypothesis that the frequency of note occurrences would be greater on strong metrical time points; it would not be really that there are more notes on downbeats than on offbeats, but rather that they would show more regular patterns. However, it is important to notice that the temporal centroid is probably the most sensitive feature to the extraction of the beat indexes.

Future work is relative to the extension of the database. It is our belief that the size of the database used for the aforementioned experiments is reasonable for making a first step into the problem of meter determination of audio signals; it permits to set up a framework for research and to state our hypotheses. However, we would need to mine a much larger database to seriously claim that our algorithm is general and scales up well.

Another important research to be pursued concerns the determination of the "phase" of the downbeat. So far, we propose an algorithm for choosing among different groupings of beats, but we did not address the issue of determining *which* of the beats in the grouping is the first (*i.e.* the 'one' in *e.g.* 'one-two-three-one-two-three...'). (See Figure 1.)

We will also extend research regarding the definition of the criterion M. Instead of defining it *a priori*, we might want to address the underlying issue of determining the relevance of *each lag*, of *each autocorrelation function* (one function being relative to one descriptor), as for the classification 'duple' or 'triple'.

5. ACKNOWLEDGMENT

Part of this research has been supported by the European IST project CUIDADO. We wish to thank Benoit Meudic (IRCAM) and Pedro Cano (UPF) for interesting discussions in the first stage of the project and Amaury Dehamel (UPF) for the Bark band decomposition code.

6. REFERENCES

Brown J. (1993), Determination of the meter of musical scores by autocorrelation. *Journal of the Acoustical Society of America* 94(4).

Clarke E. (1999), Rhythm and Timing in Music. *The Psychology of Music, 2nd edition*, Deutsch D. (Ed.), Academic Press.

Cooper G., Meyer L. (1960), *The rhythmic structure of music*. University of Chicago Press, Chicago.

Dixon S. (2001), Automatic Extraction of Tempo and Beat from Expressive Performances. *Journal of New Music Research* 30(1).

Gasser M., Eck D., Port R. (1999), Meter as mechanism: a neural network that learns metrical patterns. *Connection Science* (1).

Goto M. (2001), An Audio-based Real-Time Beat Tracking System for Music with or without Drums. *Journal of New Music Research* 30(2).

Honing H. (2001), From time to time: The representation of timing and tempo. *Computer Music Journal* 35(3).

Jain A., Duin R., Mao J. (2000), Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Machine Intell.* 22(1).

Large E., Kolen E. (1994), Resonance and the Perception of Musical Meter. *Connection Science* (6).

Large E., Palmer C. (2002), Perceiving temporal regularity in music. *Cognitive Science* 26.

Lerdahl F., Jackendoff R. (1983), *A generative theory of tonal music*. MIT Press, Cambridge.

Liu H., Motoda H. (1998), *Feature selection for knowledge discovery data mining, 2nd edition.* Kluwer Academic, Boston.

McAuley J. (1995), *Perception of time as phase: Towards an adaptive-oscillator model of rhythmic pattern processing.* Ph.D. Thesis, Indiana University, Bloomington.

Meudic B. (2002), Automatic Meter extraction from MIDI files. *Proceedings of Journées d'Informatique Musicale*, Marseille.

Mitra P., Murthy C., Pal S. (2002), Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Machine Intell.* 24(3).

Scheirer E. (1998), Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America* 103(1).

Seppänen J. (2001), *Computational models of musical meter recognition*. M.Sc. Thesis, Tampere University of Technology.

Yeston M. (1976), *The stratification of musical rhythm.* Yale University Press, New Haven.