Music meter and tempo tracking from raw polyphonic audio

Conference Paper · January 2004

Source: DBLP

3 authors:

Aggelos Pikrakis
University of Piraeus

Iasonas Antonopoulos
Vodafone Greece

S. Theodoridis
Institute of Electrical and Electronics Engineers

MUSIC METER AND TEMPO TRACKING FROM RAW POLYPHONIC AUDIO

Aggelos Pikrakis, Iasonas Antonopoulos and Sergios Theodoridis
Department of Informatics and Telecommunications
University of Athens, Greece

ABSTRACT

This paper presents a method for the extraction of music meter and tempo from raw polyphonic audio recordings, assuming that music meter remains constant throughout the recoding. Although this assumption can be restrictive for certain musical genres, it is acceptable for a large corpus of folklore eastern music styles, including Greek traditional dance music. Our approach is based on the self-similarity analysis of the audio recording and does not assume the presence of percussive instruments. Its novelty lies in the fact that music meter and tempo are jointly determined. The method has been applied to a variety of musical genres, in the context of Greek traditional music where music meter can be $\frac{2}{4}$, $\frac{3}{4}$, $\frac{4}{4}$, $\frac{5}{4}$, $\frac{7}{8}$, $\frac{9}{8}$, $\frac{12}{8}$ and tempo ranges from 40bpm to 330bpm. Experiments have, so far, demonstrated the efficiency of our method (music meter and tempo were successfully extracted for over 95% of the recordings).

Keywords: music meter tracking, beat tracking, content-based music retrieval

1. INTRODUCTION

Contemporary content-based music retrieval applications have highlighted the need to extract rhythmic features from raw polyphonic audio recordings, in order to increase the efficiency of tools that perform a diversity of tasks, including musical genre classification, query-by-humming and query-by-rhythm, to name but a few, e.g, [1, 22]. Toward this end, several attempts have been made to create an algorithmic perception of rhythm. Most research has focused on tempo tracking, whereas, on the other hand, music meter extraction has attracted significantly less attention.

The first attempts, dating back to the early 90's, involved MIDI signals [2], [3], [4], [5], [6], [7]. However, the need to circumvent the limitations imposed by MIDI signals, led to the development of several tempo-tracking methodologies that were applied on raw polyphonic audio. Goto & Muraoko [10, 11] focused on real-time beat tracking of popular music, assuming a tempo range of 61-120 bpm and music meter 4/4. Shceirer [12] introduced a tempo tracking approach that is independent of musical genre and does not demand a constant beat track. Foote ([13, 14, 15, 16]), investigated the properties of the "self-similarity matrix" and proposed the generation of the "beat spectrum" from audio recordings. A comparative study of tempo trackers was given by Dixon in [8], who also presented a real-time tempo tracker capable of displaying tempo variations in an animated display [9].

This paper [1] presents a method for the extraction of music meter and tempo from raw polyphonic audio recordings, assuming that music meter remains constant throughout the recoding. This assumption is acceptable for a large corpus of Greek traditional dance music, which has been in the center of our study. Our approach is based on the fact that the diagonals of the self-similarity matrix of the audio recording reveal periodicities corresponding to music meter and beat. By examining such periodicities it is possible to jointly estimate the music meter and tempo of the recording, as described in section 2. The method has been applied to musical genres in the context of Greek traditional music whose music meter can be $\frac{2}{4}$, $\frac{3}{4}$, $\frac{4}{4}$, $\frac{5}{4}$, $\frac{7}{8}$, $\frac{9}{8}$, $\frac{12}{8}$ and whose tempo ranges from 40bpm to 330bpm.

Section 2 describes the algorithmic aspects of our method. Section 3 provides implementation details and results of the experiments that have been carried out and finally, section 4 highlights our future research priorities.

2. MUSIC METER AND TEMPO EXTRACTION

At a first step, each raw audio recording is divided into non-overlapping long-term segments, each of which has a duration equal to 10 seconds. The choice for the length of the long-term segments is justified in section 3. Music meter and tempo are then extracted on a segment by segment basis. More specifically, for each long-term segment, a short-term moving window generates a sequence of feature vectors. Approximate values for the length of the short term window and overlap between successive windows are 100 ms and 97 ms respectively, suggesting

[1] Research described by the authors in this paper was funded by the Greek Secretariat of Research and Technology, in the framework of project POLYMNIA - EPAN 4.5

a 3 ms moving window step. Having experimented with a variety of feature candidates and their combinations, we chose to focus on two variations of the mel-frequency cepstrum coefficients (details are given in section 3.1).

Let us denote by $\mathbf{F} = \{f_1, f_2, \ldots f_N\}$, the feature sequence of length N that is extracted from a long-term segment. Sequence $\mathbf{F}$ serves as the basis to calculate the Self Similarity Matrix (SSM) of the segment [13, 14, 15, 16], using the Euclidean function as a distance metric (Figure 1). Since the SSM is symmetric around the main diagonal, in the sequel it suffices to focus on its lower triangle. At

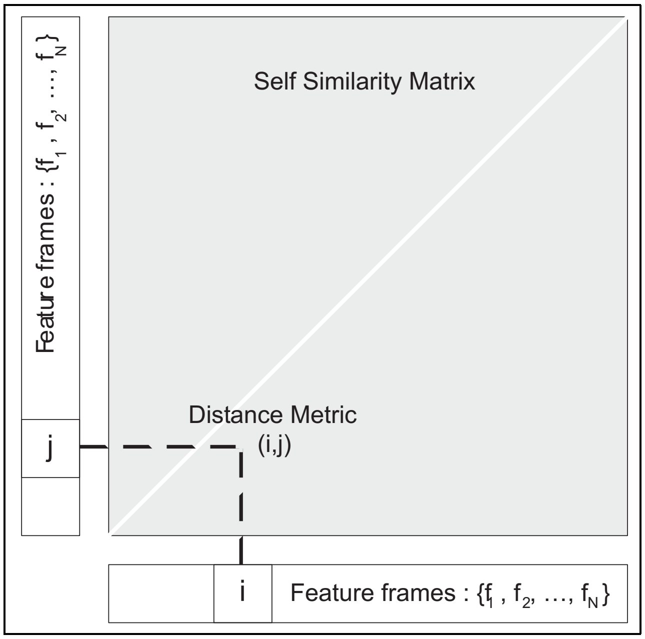

Figure 1. Self Similarity Matrix

a next step, the mean value of each diagonal of the SSM is calculated. If B_k stands for the mean value of the kth diagonal, then:

$$B_k = \frac{1}{N-k} \sum_{l=k}^{N} || \ f_l, f_{l-k} \ || \tag{1}$$

where $N-k$ is the length of the kth diagonal and $||\ .\ ||$ is the Euclidean distance function.

As can be seen in Figure 2, if B is treated as a function of k, then its plot against k exhibits certain local minima (valleys) for a number of k's. Each valley can be interpreted as corresponding to a periodicity, that is inherent in the long-term segment being examined. In Figure 2, the beat of the segment appears as a valley around the 70-th diagonal. This segment has been extracted from a Greek traditional dance of music meter $\frac{7}{8}$. In Figure 3, an overall view of the segment periodicities can be seen, where multiples of the beat, including the music meter itself also, appear as valleys. In the general case, submultiples of the beat are also likely to appear as local minima. It is worth noticing that

a) the global minimum of B does not always coincide with the beat or the music meter

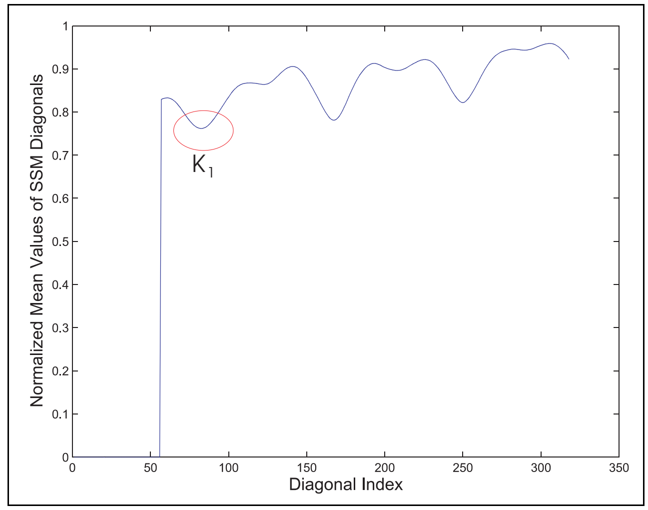

Figure 2. Plot of B_k versus k focusing on the beat range.

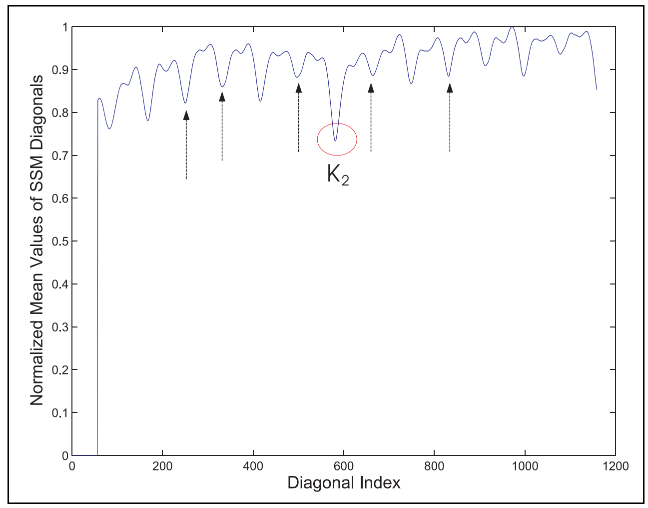

Figure 3. Plot of B_k versus k focusing on the meter range.

b) the indices of the diagonals corresponding to local minima are in most cases approximate multiples or submultiples of the beat index

c) function B decreases (increases) more rapidly around certain local minima and this appears as sharper valleys in the Figures 2 and 3.

Obviously, if diagonal k corresponds to a local minimum, then the time equivalent of the respective periodicity T_k is $T_k = k * step$, where $step$ is the short-term step of the moving window ($100 - 97 = 3$ ms for our study). Therefore, small short term steps (i.e., large overlap) increase the accuracy of the periodicity estimates, while also increasing the computational cost (due to an increase in the length of the feature sequence $\mathbf{F}$).

The aforementioned analysis suggests that, although periodicities corresponding to the music meter and beat are likely to appear as local minima (valleys) of B, further processing of B is required, in order to identify which valleys actually refer to music meter and beat. In order to proceed, we assume that the tempo of the recording, measured in beats per minute (bpm) can vary from 40bpm to

330 bpm. This beat range is applicable to the corpus of Greek traditional music of our study but may require tuning for other musical genres. It also suggests a range of diagonals, i.e., k-values, say $[k_s, k_l]$, in which the beat of the segment is expected to appear as a local minimum of B. Outside this range, i.e., for $k > k_l$, multiples of the beat lag including music meter, are also expected to appear as valley. If k_{max} is the last (downmost) diagonal of interest, music meter is likely to appear as a valley in the range of $(k_l, k_{max}]$. k_{max} must be large enough to account for all music meters and tempo ranges in question. For our music corpus, the time equivalent of k_{max} was set to 3 secs (see section 3.2).

In the sequel, *beat candidates* in the range $[k_s, k_l]$ are examined in pairs with meter candidates in the range $(k_l, k_{max}]$. For each pair of such candidates, let us denote by k_1 the lag in $(k_s, k_l]$ and by k_2 the meter candidate in $(k_l, k_{max}]$. If C_b and C_m are the numbers of beat and meter candidates respectively, then there exist $C_b * C_m$ such pairs. In order to proceed, two different decision criteria are applied on this set of pairs. Each criterion generates independently a meter and beat decision by exploiting a subset of pairs, as is explained below.

2.1. Meter decision criteria

Criterion A

At first, the local minima in the range $(k_l, k_{max}]$, for which the two neighboring local minima possess larger values, are selected. For example, such is the case with meter candidate marked as k_2 in Figure 3, that corresponds to a periodicity indicating music meter $\frac{7}{8}$. In this example, the local minima pointed by the dotted arrows, corresponding to $\frac{4}{8}$, $\frac{6}{8}$, $\frac{8}{8}$, $\frac{10}{8}$ are filtered out. This initial filtering procedure can be useful for audio recordings of music meter $\frac{7}{8}$, $\frac{9}{8}$ and $\frac{12}{8}$ stemming from Greek Traditional music.

At a second step, each beat candidate is examined in pair with the remaining meter candidates. For each such pair, the fraction $\frac{k_2}{k_1}$ should coincide, within an error margin e, with one of the fractions related to the music meters of our study, i.e., $\frac{2}{4}$, $\frac{3}{4}$, $\frac{4}{4}$, $\frac{5}{4}$, $\frac{7}{8}$, $\frac{9}{8}$, $\frac{12}{8}$. All pairs falling outside the error margin are discarded (e is assumed constant for all allowable music meters) and was set equal to 0.3 for our experiments. In other words, each music meter is considered to lie in the center of a bin, whose width is equal to $2e$. If a pair of valleys $\{k_1, k_2\}$ is assigned to a bin, k_1 is considered to be the beat lag. Furthermore, for each pair assigned to a bin, the quantity $C_{\{k_1,k_2\}}$ is calculated:

$$C_{\{k_1,k_2\}} = B_{k_1} + B_{k_2} \tag{2}$$

If more than one pair is assigned to the same bin, the pair generating the lowest $C_{\{k_1,k_2\}}$ is considered to be the winning pair for the bin. After all pairs have been processed, the music meter of the segment is determined according to the bin with the lowest $C_{\{k_1,k_2\}}$ value.

Criterion B

The previous criterion can be modified by taking into account, for the calculation of the $C_{\{k_1,k_2\}}$ value, the slope (sharpness) of the valleys of each pair being examined (and not just their absolute values). This deals with the fact that, a pair of sharp valleys corresponding to the actual music meter and beat does not always coincide with the pair having the lowest sum of absolute values. Therefore, $C_{\{k_1,k_2\}}$ can be calculated as follows:

$$C_{\{k_1,k_2\}} = \frac{slope(B_{k_1})}{B_{k_1}} + \frac{slope(B_{k_2})}{B_{k_2}} \tag{3}$$

where $slope(.)$ is a measure of the sharpness around each valley of function B. Equation 3 suggests that, if both valleys are sharp and deep, then $C_{\{k_1,k_2\}}$ has a large value. Having determined the music meter bin for all pairs, the pair yielding the maximum value of $C_{\{k_1,k_2\}}$ is considered to be the winner and the bin to which it has been assigned stands for the music meter of the segment.

Although in general both criteria result in an acceptable performance, there are cases where one succeeds and the other fails. This is similar to two classifiers, where (from pattern recognition theory [17]), a classifier with a better overall error, can fail in cases where others succeed. The remedy is to combine classifiers. To comply with this philosophy, two music meter decisions are generated from each long term segment. This has turned out to increase the overall performance significantly. If S is the number of long term segments, then $2 * S$ music meter decisions are generated. The most frequently encountered music meter is selected as the meter of the whole audio recording and its frequency of appearance is returned as the certainty of the overall decision.

2.2. Tempo Estimation

There now remains to determine the tempo of the audio recording. For the music corpus of our study, it can be assumed that tempo remains approximately constant throughout the audio recording, and is therefore possible to return an average tempo value. Alternatively, a tempo value per long term segment can be returned, so as to highlight tempo fluctuations, or even, return additionally, the time limits of the segments that produced wrong estimates, since this type of information might be useful to algorithms that extract repeating patterns from audio recordings.

At this point, it has to be noted that certain assumptions have to be adopted concerning the beat range, that may need fine-tuning for musical genres outside the context Greek traditional dance music. As a first assumption, the tempo associated with music meters $\frac{2}{4}$, $\frac{3}{4}$, $\frac{4}{4}$ and $\frac{5}{4}$ cannot be greater than 180 bpm while the tempo of meters $\frac{7}{8}$, $\frac{9}{8}$ and $\frac{12}{8}$ must be over 180bpm and up to 330bpm (as mentioned before). This implies that the range of beat lags $[k_s, k_l]$ is divided into two successive sub-regions, i.e., $[k_s, k_q]$ and $[k_q, k_l]$, which correspond to $\frac{1}{8}$ and $\frac{1}{4}$ periodicities respectively and k_s,k_q, and k_l are the lags of 330bpm (fastest $\frac{1}{8}$), 180bpm (fastest $\frac{1}{4}$) and 40bpm (slowest $\frac{1}{4}$).

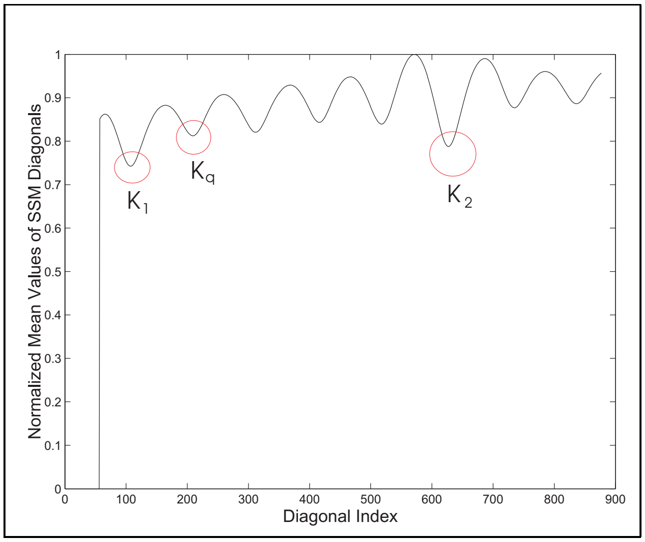

Figure 4. Plot of B_k versus k for an audio meter of $\frac{3}{4}$, where the eighth note is the dominant beat lag.

As it was previously mentioned, each long term segment produces two music meter decisions, each of which is associated with a pair of lags (k_1, k_2). In the ideal case, for a decision that coincides with the overall music meter estimate, k_2 should be the meter lag and k_1 the beat lag. However in practice, for meters $\frac{2}{4}$, $\frac{3}{4}$, $\frac{4}{4}$ and $\frac{5}{4}$, k_1 often lies in the range $[k_s, k_q]$ and refers to the eighth note periodicity instead of the expected quarter note periodicity. For example, in Figure 4, which refers to a segment from an audio recording of music meter $\frac{3}{4}$, for both decision criteria, the dominant pair of lags (marked k_1, k_2 in Figure 4) corresponds to a beat lag of $\frac{1}{8}$ and meter lag of $\frac{6}{8}$, because of a less dominant quarter note periodicity (marked k_q). As a result, meter $\frac{6}{8}$ and $\frac{3}{4}$ can be confused. However, in the context of the Greek traditional dance music that we studied, these can be thought to be equivalent and it therefore suffices to double the value of the beat lag.

3. IMPLEMENTATION DETAILS AND RESULTS OF EXPERIMENTS

The length of the long term segments was set equal to 10 secs with zero overlap between successive segments. This segment length is large enough to capture periodicities of slow tempo values in the range of 40bpm. For the short term analysis, the moving window size was approximately 93 ms (4096 samples for sampling frequency 44.1kHz) and the moving window step was set equal to $\cong$ 3 ms (128 samples for sampling frequency 44.1kHz). It has to be noted that the moving window step reflects the beat accuracy. Smaller values produce more accurate beat estimates but increase computational complexity significantly. For slow tempo recordings, we also experimented with longer short term windows up to 186 ms (8192 samples for sampling frequency 44.1kHz). However, for fast tempo values (as is the case with music meter $\frac{7}{8}$, etc, in the context of Greek traditional music), large short term windows result in poor valleys in the beat range. Although in this case smaller short term windows would be desirable, it would not be possible to achieve tone resolution in the low frequency range as imposed by the chroma-based MFCCS (see 3.1).

3.1. Feature selection details

For the short term analysis of each long term audio segment, we considered both energy and mel frequency cepstral coefficients (MFCCs) [18, 19]. In addition to the standard MFCCs, which assume equally spaced critical band filters in the mel scale, we also experimented with a critical band filter bank consisting of overlapping triangular filters, whose center frequencies follow the equation:

$$C_k = 110 * 2^{\frac{k}{12}} \tag{4}$$

That is, the filter bank is chosen to align with the chromatic scale of semitones (starting from 110 Hz and reaching up to approximately 5KHz). If whole-tone spacing is adopted, equation (4) becomes:

$$C_k = 110 * 2^{\frac{k}{6}} \tag{5}$$

Our variation of the mel frequency cepstrum bears certain similarities with the “chroma vector” [21].

Compared with energy, the two variants of the MFCCs, although computationally expensive, yield significantly better results, and this is mainly because periodicities corresponding to beat and meter are emphasized (Figure 5). It has to be noted though, that energy gave good results for a significant number of recordings of music meter $\frac{2}{4}$ and $\frac{3}{4}$, but failed for most of the recordings with music meter $\frac{5}{4}$, $\frac{7}{8}$, $\frac{9}{8}$ and $\frac{12}{8}$. Depending on the frequency content distribution of the recording, especially in the case of dominant singing voices, our variant of the mel frequency cepstrum led to an improved performance compared to the standard approach. This was mainly observed in the cases of $\frac{5}{4}$, $\frac{9}{8}$ and $\frac{12}{8}$. The standard MFCC's were computed using Slaney's auditory toolbox [20].

3.2. Self Similarity Analysis details

For the distance metric, we adopted the standard Euclidean distance function (also used in [21]). The use of the cosine distance ([13, 14, 15, 16]) in our experiments tended to lead to inferior performance.

Due to the assumptions adopted in section 2, we only need to focus on a subset of the diagonals of the SSM. For sampling frequency 44.1KHz and moving window step 3 ms, the range $[k_s, k_l]$ is mapped to diagonals $[63, 517]$, k_q (fastest quarter note) corresponds to the 115-th diagonal and k_{max} is mapped to the 1159-th diagonal. For our music corpus, k_{max} was chosen large enough to account for music meter periodicities of $\frac{2}{4}$ and tempo values in the range of 40bpm. In general, k_{max} (as well as k_s, k_q and k_l) needs fine tuning depending on the music genre. For example, if $\frac{4}{4}$ audio recordings of slow tempo also need to be taken into consideration, the k_{max} value should at least be doubled.

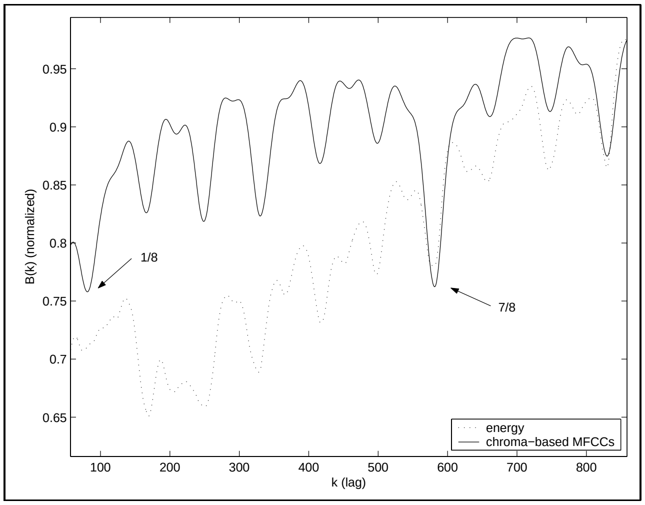

Figure 5. Plot of B_k versus k for energy and chroma-based MFCCs, for ten second audio extract of music meter $\frac{7}{8}$.

3.3. Results of experiments

The music corpus of our study consists of 300 raw audio recordings of Greek dance folklore music and neighboring Eastern music traditions. Throughout each recording, music meter remains constant. This corpus was assembled under musicological advice and focuses on most frequently encountered folk dances, exhibiting significant rhythmic variety over beat and music meter (see Table 1).

music meter	tempo range (bpm)	# of recordings
$\frac{7}{8}$	200-280	45
$\frac{9}{8}$	250-330	45
$\frac{12}{8}$	260-330	10
$\frac{2}{4}$	40-160	90
$\frac{3}{4}$	80-170	60
$\frac{4}{4}$	70-140	90
$\frac{5}{4}$	90-120	10

Table 1. Description of music corpus.

Approximately one third of the audio corpus consists of live performances and digitally remastered recordings. For live performances, certain beat fluctuation was observed and for these recordings it makes more sense to return a beat value per long term segment, instead of an average beat value.

For the majority of the recordings (over 95%), the rhythmic features in question were successfully extracted. Most mistaken results were produced by confusion of music meter $\frac{2}{4}$ with $\frac{4}{4}$, $\frac{5}{4}$ with $\frac{9}{8}$ or $\frac{4}{4}$ and $\frac{7}{8}$ with $\frac{3}{4}$ or $\frac{4}{4}$. The main reason for the above cases of confusion, is that the dominant periodicities in the beat range, often deviate significately from the desired values and as a result the pair (beat lag, meter lag) is assigned to an incorrect (neighboring) meter bin. Especially in the case of meter $\frac{2}{4}$, confusion with $\frac{4}{4}$ may also occur because a very strong periodicity at four quarter-notes is observed.

In addition, for certain long term segments, due to the nature of the signal, the features that have been employed fail to capture any periodicities at all. As a last remark, in certain cases, especially for meter cases of $\frac{7}{8}$, $\frac{9}{8}$ and $\frac{12}{8}$, a dominant quarter note, appears in the beat range instead of the expected eighth note, thus leading to an incorrect meter and beat estimate. As a remedy to this situation, it is possible to divide by two all valleys in the range $[k_q, k_l]$ and treat these new values as candidate beat lags.
All experiments were carried out using the Matlab workbench.

4. CONCLUSIONS AND FUTURE WORK

We have presented a method for the extraction of music meter and tempo from raw audio recordings, assuming that music meter remains constant throughout the recording. The method was applied on a music corpus consisting of recordings stemming from Greek Traditional Music and neighboring music traditions. In the future, feature selection will be expanded to cover more feature candidates and their combinations. In addition, we will investigate ways to pre-process the SSM prior to calculating the mean of diagonals, in order to detect subsections of the diagonals that emphasize the inherent periodicities. Toward this end, Dynamic Time Warping techniques are expected to be employed [22]. Finally, it is our intention to investigate the effectiveness of the methodology in the context of other music genres.

5. REFERENCES

[1] George Tzanetakis and Perry Cook, "Musical Genre Classification of Audio Signals", *IEEE Transactions on Speech and Audio Processing, vol. 10, No. 5, July 2002*

[2] Allen, P. & Dannenberg, R. "Tracking musical beats in real time", *Proceedings of the 1990 International Computer Music Conference, pages 140143. International Computer Music Association*, San Francisco, USA, 1990.

[3] Large, E. & Kolen, J. "Resonance and the perception of musical meter", *Connection Science, 6:177208*, 1994

[4] Large, E. "Beat tracking with a nonlinear oscillator" *Proceedings of the IJCAI95 Workshop on Artificial Intelligence and Music, pages 2431* International Joint Conference on Artificial Intelligence, 1995

[5] Large, E. "Modeling beat perception with a nonlinear oscillator" *In Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 1996

[6] Dixon, S. "A lightweight multi-agent musical beat tracking system" *PRICAI 2000: Proceedings of the*

Pacific Rim International Conference on Artificial Intelligence, pages 778788, Springer, 2000

[7] Dixon, S. & Cambouropoulos, E. "Beat tracking with musical knowledge" *ECAI 2000: Proceedings of the 14th European Conference on Artificial Intelligence, pages 626630,* IOS Press, 2000

[8] Dixon, S. "An Empirical Comparison of Tempo Trackers" *Proceedings of 8th Brazilian Symposium on Computer Music, pp 832-840* Fortaleza, Brazil, 31 July - 3 August 2001

[9] Dixon, S. et al, "Real Time Tracking and Visualization of Musical Expression" *ICMAI 2002: Proceedings of the 2nd International Conference on Music and Artificial Intelligence, pages 58-69*, LNAI 2445, Springer-Verlag, 2002

[10] Goto, M. & Muraoka, Y. "A real-time beat tracking system for audio signals" *Proceedings of the International Computer Music Conference, pages 171174* Computer Music Association, San Francisco, USA, 1995

[11] Goto, M. & Muraoka, Y. "Real-time beat tracking for drumless audio signals" *Speech Communication, 27(34):331335, 1999*

[12] Scheirer, E. "Tempo and beat analysis of acoustic musical signals" *Journal of the Acoustical Society of America, 103(1):588601, 1998*

[13] J. Foote "Visualizing Music and Audio using Self-Similarity" *Proceedings of ACM Multimedia 99, pp. 77-80* Orlando, FL, USA, ACM Press, 1999

[14] J. Foote, and S. Uchihashi "The Beat Spectrum: A New Approach to Rhythm Analysis" *Proceedings of International Conference on Multimedia and Expo (ICME) 2001*

[15] Jonathan Foote, Matt Cooper, and Unjung Nam "Audio Retrieval by Rhythmic Similarity" *Proceedings of Third International Symposium on Musical Information Retrieval (ISMIR), pp. 265-266* Paris, France, September 2002

[16] J. Foote "Automatic Audio Segmentation using a Measure of Audio Novelty" *Proceedings of IEEE International Conference on Multimedia and Expo, vol. I, pp. 452-455* 2000

[17] Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern Recognition (2nd Edition)*, Academic Press, 2003.

[18] Lawrence Rabiner and Bing-Hwang Juang, *Fundamentals Of Speech Recognition* Prentice Hall, New Jersey, USA, 1993

[19] John R. Deller Jr, John G.Proakis and John H.L.Hansen *Discrete-Time Processing Of Speech Signals* Prentice Hall, New Jersey, USA, 1987

[20] Auditory Toolbox Malcom Slaney, Technical Report #1998-010 , Interval Research Corporation malcolm@interval.com

[21] Roger B. Dannenberg & Ning Hu "Discovering Musical Structure in Audio Recordings" *Proceedings of 2nd International Conference on Music and Artificial Intelligence, pp 43-57*, Edinburg, Scotland, September 2002

[22] A. Pikrakis, S. Theodoridis, D. Kamarotos, "Recognition of isolated musical patterns using context dependent dynamic time warping", *IEEE Transactions on Speech and Audio Processing, vol. 11(3), May 2003*