AN APPLICATION OF EMPIRICAL MODE DECOMPOSITION ON TEMPO INDUCTION FROM MUSIC RECORDINGS

Aggelos Pikrakis and Sergios Theodoridis
Dept. of Informatics and Telecommunications
University of Athens, Greece
{pikrakis, stheodor}@di.uoa.gr, http://www.di.uoa.gr/dsp

ABSTRACT

This paper presents an application of Empirical Mode Decomposition (EMD) on the induction of notated tempo from music recordings. At a first stage, EMD is employed as a means to segment music recordings into segments that exhibit similar rhythmic characteristics. At a second stage, EMD is used in order to analyze the diagonals of the Self-Similarity Matrix of each segment, so as to estimate the tempo of the recording. The proposed method has been employed on various music genres with music meters of $\frac{2}{4}$, $\frac{3}{4}$ and $\frac{4}{4}$. Tempo has been assumed to remain approximately constant throughout each recording, ranging from 60bpm up to 220bpm.

1 INTRODUCTION

Tempo extraction is generally acknowledged to be useful in a variety of Music Information Retrieval applications and has been often studied in close relationship with beat tracking. A recently published study and comparison of well known algorithms is presented in [1].

Most proposed approaches assume that tempo remains approximately constant throughout the recording which is also the case with our method. In addition, our method focuses on music recordings where music meter can be one of $\frac{2}{4}$, $\frac{3}{4}$ and $\frac{4}{4}$, with tempo ranging from 60bpm up to 220bpm. This paper is an attempt to apply a relatively new transform, called Empirical Mode Decomposition (EMD), in the context of tempo extraction. EMD was originally introduced in the context of non-stationary time-series analysis [2] and its relationship with dyadic filter banks was later investigated [3]. The origin of EMD is algorithmic in nature and lacks a solid theoretical framework. To bridge this theoretical gap, certain attempts have been made recently [4]. This fact has not, however, discouraged researchers from applying EMD especially in the context of geophysical and biosignal processing. To our knowledge, the EMD technique has so far met very limited recognition in the context of Music Information Retrieval [5].

In this paper EMD is used in conjunction with Self Similarity Analysis of music recordings [6] in order to achieve tempo induction. Previous work in the field has shown that the diagonals of the Self Similarity Matrix can reveal signal periodicities and that tempo also manifests itself as a signal periodicity, e.g., [7]. In particular, if the mean value of each diagonal is computed, then the resulting sequence of mean values exhibits certain minima that correspond to inherent signal periodicities. We propose that if EMD is used to decompose this sequence, signal periodicities manifest themselves more clearly in the resulting components. By processing these components, it is possible a) to achieve a rough segmentation of a music recording into clusters of segments that exhibit similar rhythmic characteristics and b) to extract reliable tempo estimates from the generated clusters of segments.

The reason we chose EMD is because, by its algorithmic nature, it considers signals at the level of their local oscillations and examines the evolution of a signal between consecutive local extrema, e.g., consecutive local minima. This fits nicely with the nature of the sequence of mean values of diagonals that is extracted from the SSM of music recordings, if this sequence is treated as a signal.

The paper is organized as follows: the next section describes the feature extraction stage, Section 3 proposes how EMD can be used to provide a rough segmentation of the recording and Section 4 describes how EMD can be applied on the resulting segments in order to achieve tempo induction. Results are presented in Ssection 5; conclusions and ideas for future work are given in Section 6.

2 FEATURE EXTRACTION

At a first step, the music recording is split into overlapping long-term segments. Each long-term segment is 5 seconds long with 4 seconds overlap between successive windows. The energy envelop of each long-term window is then extracted by means of a short-term processing technique. Suggested values for the length, w_s and hop size h_s of the short-term window are 95ms and 5ms respectively. The extracted energy envelop is then used to generate the Self-Similarity Matrix (SSM) of each segment [6]. To this end, the Euclidean Distance function is used as the similarity metric.

Once the SSM is generated, the mean value of each diagonal is computed. Let $\mathbf{B}(k)$ denote the mean value of the k-th diagonal, $k = 1 \ldots D$, where D is the total number of diagonals. If $\mathbf{B}(k)$ is treated as a function of k, it can be observed that signal periodicities appear as local minima (valleys) of $\mathbf{B}$. This can be seen in Figure 1. The deeper a valley, the stronger the periodicity. In the sequel, we will also refer to index k as the "lag" index. The relative positions of lags that correspond to valleys reveal useful information about the rhythmic characteristics of a segment. Obviously, if $\mathbf{B}(k_1)$ is a local minimum, the corresponding periodicity, measured in seconds, is $(k_1 - 1) * h_s$.

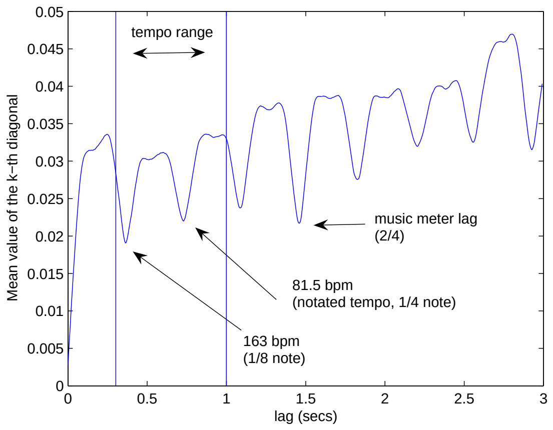

Figure 1. Plot of $\mathbf{B}$ for a segment stemming from a $\frac{2}{4}$ recording with notated tempo ≈81.5 bpm. For convenience, only periodicities up to 3 seconds are shown. It can be seen that notated tempo does not correspond to the deepest valley.

Previous work in the field ([7]) has suggested that notated tempo also manifests itself as a periodicity, i.e., a valley of B, although not always the deepest one, as is the case in Figure 1. In the sequel, we treat the function $\mathbf{B}$ as the *"rhythmic signature"* of the long-term segment from which it is extracted.

3 SIGNATURE CLUSTERING USING EMD

Our next goal is to group *"rhythmic signatures"* into clusters and compute the mean signature of each cluster. This is because reliable tempo estimates cannot be extracted from all signatures, due to the fact that certain regions of a music recording contain introductory or transitive parts that distort periodicities. If a number of signatures form a cluster, this is indicative of an underlying rhythmic similarity and it is expected that it will yield more reliable tempo estimates.

In order to perform clustering, EMD is applied separately on each signature. The basic steps of EMD, given a signal $x(t)$ can be summarized as follows [2], [3]:

1. Identify all extrema of x(t)

2. Interpolate between minima (resp. maxima), ending up with some "envelope" $e_{min}(t)$ (respectively $e_{max}(t)$).

3. Compute the average $a(t) = (e_{min}(t) + e_{max}(t))/2$

4. Extract the detail $d(t) = x(t) - a(t)$, also known as the IMF. Iterate on the residual $a(t)$ until a stopping criterion is satisfied, i.e., $a(t)$ is reasonably zero everywhere [3].

Let $\mathbf{B}_m$ denote the signature of the m-th long-term segment, $m = 1 \ldots M$, where M is the total number of long-term segments. If c_m is the number of components (IMFs) generated by the EMD for $\mathbf{B}_m$, then, at a first step, all $\mathbf{B}_m$'s that have generated the same number of components are grouped to form a single cluster. For example, the signature in Figure 1 is decomposed into 5 components and will be part of a cluster where all signatures have 5 components. If no other signature yields five components, this signature will form a cluster of its own.

At a next step, all signatures in a cluster are further examined in order to form sub-clusters. To this end, let IMF_i be the i-th component of a signature $\mathbf{B}_m$ that has been assigned to some cluster, where $i = 1 \ldots m_K$ and m_K is the number of components of $\mathbf{B}_m$. By the nature of EMD, IMF_{m_K} (i.e., the last component, also known as the residual) only captures the slowly varying nature of the signature and does not provide any useful information about inherent periodicities. This is why we choose to ignore this component while refining the clusters that have been already formed. To continue, the energy of all remaining IMF_i's is computed and components are sorted in descending order, according to their energy values. The resulting order is then used to form sub-clusters within every cluster, i.e. segments that yield the same *order* of components are considered to be similar. For the example of Figure 1, the components (excluding the last one) are ordered in terms of energy values as follows: $\{2nd, 3rd, 4th, 1st\}$.

Figures 2 and 3 show the two components with higher energy values for the example in Figure 1. When this refined clustering is complete, the mean signature for each cluster is computed by simply averaging the B_m's of each

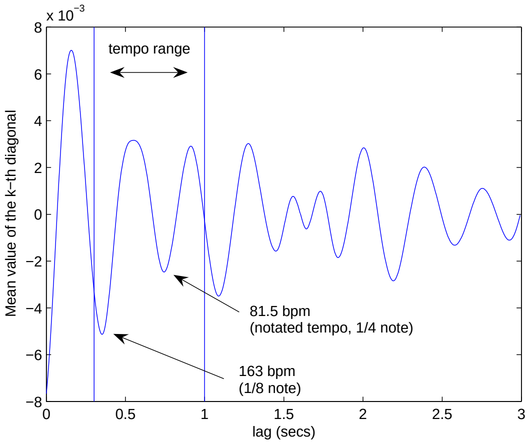

Figure 2. The second component of the signature in Figure 1.

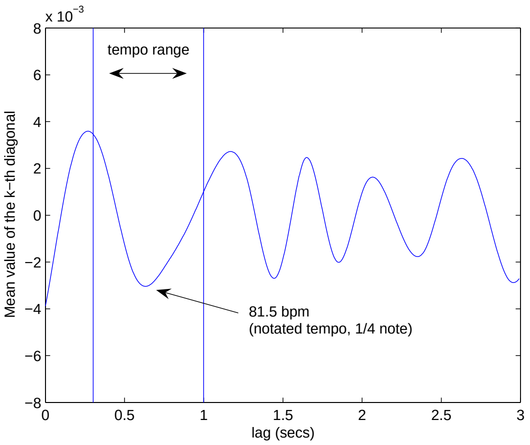

Figure 3. The third component of the signature in Figure 1.

cluster. This is possible because all signatures have the same length due to the fixed size of the long-term window. At the end of this stage, the initial music recording has been split into regions that are determined by the respective clusters. Certain regions may, of course, contain non-adjacent segments.

4 TEMPO INDUCTION

We then focus on clusters that consist of more than two signatures. The mean signature R_l of the l-th cluster is then used to provide two separate tempo estimates for the segments belonging to the respective cluster. To this end, R_l is decomposed using EMD. The energy of each component of R_l is computed and the resulting components are again sorted in descending order, according to their energy values. Then, we focus on the two components that possess the higher energy values and from each component a tempo estimate is extracted, thus yielding two tempo estimates per cluster. For convenience, let us assume that the components in Figures 2 and 3 are the two high-energy components of a mean signature R. It can be observed, that the component in Figure 2 exhibits two dominant periodicities inside the tempo limits, whereas in the component of Figure 3 only one periodicity appears inside the tempo region and this periodicity refers to the notated tempo value. By the nature of EMD, if IMF_{i_1} and IMF_{i_2} are two components with $i_1 < i_2$, IMF_{i_2} is expected to capture longer periodicities. Therefore, it is not a surprise that the presence of the periodicity of the $\frac{1}{8}$ note in Figure 2 is weakened in Figure 3, whereas the opposite holds for the longer periodicity of the $\frac{1}{4}$ note.

In order to extract a tempo estimate from each one of the two components, the following procedure is applied on each component:

1. All valleys of the component are detected, including valleys to the right of the tempo region.
2. Each valley in the tempo region, is then examined against all valleys with larger lags. Let k_m be the lag of a valley in the tempo region and k_i be the lag of the valley against which it is examined. The ratio $\frac{k_i}{k_m}$ is then computed. If the roundoff error of this ratio is smaller than 0.1, then k_2 is considered to be a multiple of k_m, i.e., k_m is treated as a fundamental periodicity and k_i as its multiple. This procedure is repeated for all possible pairs, yielding a set L_{k_m} of multiples (for lag k_m in the tempo region). In the end, the following sum is computed for k_m, i.e., $P_{k_m} = c_{k_m} + \sum_{\forall k_i \in L_{k_m}} c_{k_i}$, where c_{k_i} is the value of the EMD component for lag k_i.
3. The above step is repeated for all valleys that fall within the allowable tempo limits. The valley with the highest sum is selected as the winner and the corresponding lag as the periodicity of the tempo estimate.

After tempo extraction has been completed for all clusters, all tempo estimates are placed in a histogram and the tempo corresponding to the highest peak is selected as the tempo of the music recording. It is often the case, that the histogram exhibits lobes around peaks because EMD tends to slightly displace periodicities. This is why an averaging of the lobes (histogram smoothing) is needed prior to selecting the highest peak.

5 RESULTS

The proposed method has been applied on a variety of music genres, including western pop/rock music and Greek Traditional music. A total of 400 recordings were studied. The tempo of these recordings was notated from musicologists. The complete list of the titles of recordings, along with the respective notated tempi is available at `http://www.di.uoa.gr/pikrakis/tempo.html`. Recordings were chosen on the basis of the following criteria:

- Notated tempo remains approximately constant throughout each recording. The tempo ranges from 60bpm up to 220bpm.
- Music meter remains constant throughout each recording and can be one of the following: $\frac{2}{4}$, $\frac{3}{4}$ and $\frac{4}{4}$.
- Instrumentation varies, including non-percussive recordings and absence of vocals.

Table 1 provides a rough distribution of recordings among genres and music meters.

5.1 Performance of the clustering scheme

As it was described in Section 3, we choose to estimate tempo from clusters consisting of at least 3 signatures. At an average, 25% of the signatures in each audio recording is grouped into such clusters. This suggests that, on average, 25% of the length of an audio recording takes

	Music Meter		
Broad Music Genre	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{4}{4}$
Contemporary Pop/Rock	20%	5%	40%
Traditional Greek Folk music	10%	15%	10%

Table 1. Distribution of music tracks among genres and music meters.

part in the tempo extraction process. This is because the clustering criteria are quite strict, in the sense that certain components returned by the EMD contain a very low percentage of the energy of the signature and could actually be omitted, thus loosening the clustering criteria. For the recordings that we studied, the number of signatures in these clusters can vary significantly depending on the recording, with large clusters containing approximately 20 signatures.

5.2 Performance of the tempo extraction algorithm

When the proposed method yields correct tempo estimates, the average accuracy of the extracted tempo value lies within 3.5% of the respective notated tempo value. This is due to histogram smoothing (see Section 4) and the displacement of valleys that is ususally more apparent in EMD components that capture longer periodicities. Table 2 summarizes the cases where the algorithm fails to return the notated tempo (percentages refer to the total number of recordings of the music corpus). It can be seen that in the majority of cases, the returned tempo value is twice or half the notated value, with the exception of fast contemporary music recordings of music meter $\frac{3}{4}$, where the returned value is three times the notated value, i.e., coincides with the periodicity that corresponds to a whole music meter. Another interesting case of confusion stems from Greek Folk music of meter $\frac{2}{4}$ where the dotted quarter-note is often perceived by humans as a dominant periodicity. For certain recordings of this particular type of music, our method also returns as notated tempo the periodicity of the dotted quarter-note. Table 2 suggests that the notated tempo was successfully inducted (within a 3.5% accuracy) from 79% of the recordings. It has to be noticed that, in the vast majority of failures, notated tempo is present in the histogram of tempo values (prior to selecting the winner peak) as the second highest peak and quite often its height is very close to the highest peak. The study in [1] suggests that there exists an upper bound of approximately 80% for such algorithms. However, it has to be noticed that the algorithms in [1] were compared on a different dataset and in addition our paper treats a subset of music meters encountered in [1]. Since this is the first time our work is reported, further improvements are expected by refinements in the clustering stage.

	Music Meter		
Broad Music Genre	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{4}{4}$
Contemp. Pop/Rock (2xbpm)	3%	0%	2%
Contemp. Pop/Rock ($\frac{1}{2}$bpm)	1.5%	0%	3%
Contemp. Pop/Rock (3xbpm)	0	1.5%	0%
Greek Folk Dances (2xbpm)	1%	0%	3%
Greek Folk Dances ($\frac{1}{2}$bpm)	2%	0%	2%
Greek Folk Dances (3xbpm)	0%	0%	0%
Greek Folk Dances (1.5xbpm)	2%	0%	0%

Table 2. Distribution of tempo induction failures.

6 CONCLUSIONS

This paper presented the application of EMD on tempo extraction of music recordings. EMD was used in a twofold manner: a) as a means to generate similarities among rhythmic signatures, thus yielding a rough audio segmentation and b) as a decomposition whose components emphasize periodicities that are related with the rhythmic characteristics of the music recording. The results are very encouraging, indicating that EMD is a very promising tool for Music Information Retrieval tasks. In the future, we will revisit the clustering criteria so that the segmentation stage covers a larger percent of the music recording and will also investigate extending our approach to non-binary meters such as $\frac{7}{8}$, $\frac{9}{8}$ and $\frac{5}{4}$. We will also study the performance of the method in connection with perceived tempo values from music recordings.

7 REFERENCES

[1] F. Gouyon et al., “An experimental comparison of audio tempo induction algorithms”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14(5), pp. 1832-1844, Sept. 2006.

[2] N.E. Huang et al., “The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis, *Proceedings of R. Soc. London*, vol. 454, pp. 903-995, 1998.

[3] P. Flandrin, G. Rilling and P. Concalves, “Empirical mode decomposition as a filter bank”, *IEEE Signal Processing Letters*, vol 11(2), pp. 112-114, Feb. 2004.

[4] E. Delechelle, J. Lemoine and O. Niang, “Empirical mode decomposition: an analytical approach for sifting process”, *IEEE Signal Processing Letters*, vol. 12(11), pp. 764-767, Nov. 2005.

[5] P. Heydarian and J. D. Reiss, “Extraction Of Long-Term Structures In Musical Signals Using The Empirical Mode Decomposition”, *Proceedings of DAFX-05*, Sep. 2005, Madrid, Spain.

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

[7] A. Pikrakis, I. Antonopoulos and S. Theodoridis, “Music Meter and Tempo Tracking from Raw Polyphonic Audio”, *Proceedings of ISMIR 2004*, pp. 192-197, Sep. 2004, Barcelona, Spain.