ENSEMBLE EMPIRICAL MODE DECOMPOSITION APPLIED TO MUSICAL TEMPO ESTIMATION

Michael Fulton[1] *and Prof. J .J. Soraghan*
Institute for Communications and Signal Processing (ICSP)
Department of Electronic and Electrical Engineering
University of Strathclyde

ABSTRACT

Knowledge of the tempo of a piece of music is not only a very important part of any music transcription system but has many uses on its own, from automatic segmentation to video synchronisation. The purpose of this paper is to investigate the suitability of Empirical Mode Decomposition (EMD) when used for this task. It has already found uses in many areas such as speech processing and biomedical applications where the core physical processes involved in creating the data are of importance. It is for this reason that EMD followed by Hilbert Spectrum calculation was applied to meter analysis.

Keywords – EMD, Tempo, Meter, Hilbert Spectrum

1. INTRODUCTION

The term meter is used to describe the different metrical levels of a piece of music which are hierarchical in nature. The GTTM [1] states that western music conforms to a metrical grid where beats are equally spaced and the period of each metrical level is an integer multiple of the level above it (i.e. the level faster than it.) This is normally either by a factor of 2 or 3 depending on the time signature of the music in question. The 3 main metrical levels have been termed, (from lowest frequency to highest), the measure, tactus or tempo and tatum [3, 4]. In this paper a novel technique is presented which seeks to find multiple levels of the meter of a piece of music. Section 2 gives a brief review of related work in this area and the theoretical background of EMD and the Hilbert Transform is given in section 3. A description of the algorithm used is given in section 4 followed by the results and conclusion.

2. RELATED WORK

Tempo induction for musical audio is a well researched area with the majority of the earlier studies based upon analysis of symbolic data such as MIDI or manually parsed scores containing onset times and durations [2]. More recently systems tend to have either the raw or compressed audio as the input. This compression can take many forms from decimated sub-band envelopes as in Scheirer and Klapuri [4] to more complex front-ends which seek to extract phenomenal accents by detection of sudden changes in timbre, dynamics or harmonic structure. In [3, 4] an overview of tempo induction systems is presented while [5] gives a comparison of onset detection functions used in music analysis. A method based upon inspection of Inter Onset Intervals (IOI) [4, 6] has also been reported. These are calculated between pairs of onsets within a certain time length and sometimes weighted based on the onsets' prominence.

3. EMPIRICAL MODE DECOMPOSITION

EMD was originally proposed in order to allow for the subsequent application of the Hilbert Transform to a data set, which has been termed the Hilbert-Huang Transform [7]. It seeks to decompose any signal into a set of completely data-adaptive basis functions called Intrinsic Mode Functions (IMFs) [7]. EMD has no *a-priori* defined basis, unlike the Fourier and wavelet transforms, and therefore can deal easily with both non-linear and non-stationary data. The foundation of the theory is that any data set essentially comprises of the superimposition of a finite number of different, simple oscillatory modes, termed IMFs. These IMFs have the same number of zero-crossings as extrema (or only differing by at most one) and are also symmetric with respect to the 'local mean' [7]. These restrictions are in place so as to facilitate robust instantaneous frequency and envelope calculation as will be shown in section 3.2.

Despite widespread use, EMD does not admit an analytic formulation and is essentially defined by its algorithm [7]. Given a real signal $x(t)$, EMD is applied as follows to obtain a set of IMFs [8]:

1. Identify all the extrema of $x(t)$.
2. Interpolate between successive maxima and minima, respectively, to obtain upper and lower envelopes.
3. Calculate the local mean $m(t)$ between the envelopes.
4. Extract the detail, $d(t) = x(t) - m(t)$.
5. Detail then becomes extracted IMF. Iterate on residual $m(t)$.

In practice a sifting process has to be implemented whereby steps one to four are repeated until the detail can be considered zero-mean and conforms to the IMF restrictions. The detail then becomes the first IMF and is subtracted from the original data and the process begins again.. Therefore the signal $x(t)$ can be expressed as:

[1] Email: mfulton@eee.strath.ac.uk

$$x(t) = m_I(t) + \sum_{i=1}^{I} d_i(t) \quad (1)$$

where I is the number of IMFs extracted and $m_I(t)$ is the final residual. There are a number of considerations to be taken into account when implementing EMD relating to the number of siftings, stopping criteria and envelope interpolation; details of which can be found in [7, 8].

3.1. Ensemble EMD

Fig 1.(a). shows a data set consisting of a low frequency wave and two high frequency bursts. Fig 1(b)-1(e) depict the IMFs extracted from data which have no guarantee of being globally orthogonal, although have been shown to be locally orthogonal in [7, 8]. This results in what is termed 'mode mixing', where different modes of oscillation appear in a single IMF. The high frequency and low frequency components of the data clearly belong to different modes of oscillation but have been spread over the all IMFs extracted. Mode mixing not only causes serious aliasing in any subsequent Hilbert Spectrum, but can cause individual IMFs to be devoid of physical meaning [7, 9].

To remedy this Huang [9] advocated Ensemble Empirical Mode Decomposition (EEMD), in which an ensemble mean is taken of a number of IMFs extracted from multiple applications of EMD to the original data with the addition of *different* white noise series each time. It was shown in [10] that EMD acts as an adaptive dyadic filter bank when applied to white noise and therefore the principle of EEMD is relatively straight forward. The added white noise will occupy the entire time frequency space and the parts of the signal will be automatically projected onto proper scales of reference established by the white noise, thus eliminating mode mixing. The individual IMFs are, of course, very noisy but the ensemble mean of a number of corresponding IMFs will leave only the signal, as a collection of white noise cancels each other out in a time space ensemble mean [9]. As the number of ensemble members, N, increases the effect of the noise decreases as governed by the well established rule :

$$\varepsilon_n = \frac{\varepsilon}{\sqrt{N}} \quad (2)$$

where ε is the amplitude of the added noise and ε_n is the standard deviation of the original data and the summation of the IMFs. Fig 2(b)-2(e) show the resulting IMFs with EEMD applied to the same signal as in fig.1.(a) with $N = 50$ and $\varepsilon = 0.1$ times that of the original data. The two separate components now reside in two different IMFs as compared to being shared over all the IMFs as in fig. 1(b)-1(e).

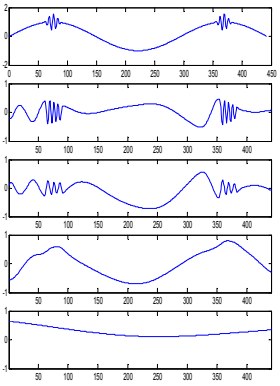

Figure 1 (a).300Hz sine wave with intermittent bursts of smaller amplitude 3000Hz sine wave. (b-e) IMFs extracted by EMD in ascending order. Mode mixing clearly visible.

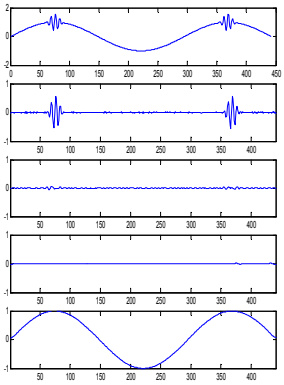

Figure 2. (a).300Hz sine wave with intermittent bursts of smaller amplitude 3000Hz sine wave. (b-e) IMFs extracted by EEMD in ascending order. Mode mixing is clearly eliminated.

3.2. Hilbert Transform

The Hilbert Transform of $x(t)$ may be written as:

$$y(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau \quad (3)$$

where the principle value of the integral is used. Given $x(t)$ and $y(t)$, a complex analytic signal $z(t)$ can be constructed:

$$z(t) = x(t) + jy(t) = E(t)e^{j\varphi(t)} \quad (4)$$

The instantaneous frequency $\varpi(t)$ can be calculated as the first order difference of the unwrapped phase $\varphi_u(t)$:

$$\varpi(t) = -\frac{d}{dt}\varphi_u(t) \quad (5)$$

However, this does not give sensible instantaneous frequency for an arbitrary signal [7, 8, 9] as a few restrictions have to be applied to the data. The real part of its Fourier transform may only have positive frequency (i.e. no DC component.) This applies locally, so any 'riding waves' and/or asymmetric envelopes can be considered as a DC offset and have to be eliminated. EMD was devised for this reason as it produces IMFs which are zero mean locally and therefore admit to meaningful instantaneous frequency calculation.

4. E.E.M.D. TEMPO ESTIMATION

In our novel algorithm initially the signal undergoes some pre-processing to reduce the data size and detect clearly musical onsets. EEMD is then applied to reveal the modes of oscillation, which are then used to estimate the meter period after undergoing frequency analysis and matching to a harmonic template. Pulse trains

corresponding to the tactus and measure are produced and the correct phases are calculated.

4.1. Pre-processing

The input music sampled at 44.1 KHz undergoes a Short-Time Fourier Transform (STFT) with window size of 1024 samples and a 50% hop. This reduces the sample rate to around 86Hz which is more than enough to detect multiple consecutive rapidly occurring onsets. The complex domain onset detection function [11] is calculated as it gives relatively sharp peaks even for sounds with no major energy increase at the onset due to its inclusion of a phase deviation term. This gives a measure of the non-stationarity of the signal in each frame of the STFT by calculating the deviation of each frequency bins' energy and phase from a prediction made using the previous frames.
The output of this detection function is summed with an energy based detection function as given in [11]. This energy based detection function is simply the sum of the squares of the magnitudes in each frequency bin. The energy based detection function is equivalent to the signal's envelope and so carries the majority of the rhythmic fluctuations in the original music. The summation of the two detection functions is important as it allows for the inclusion of softer onsets which do not show up as clearly if energy the alone is used.

4.2. EEMD

The combined detection function then undergoes EEMD with an ensemble of, N=100, and white noise amplitude of 0.1 times that of the data as recommended in [9].
A Hilbert transform is taken of each IMF in turn and the instantaneous frequencies and envelopes are used to form a Hilbert Spectrum, a 2-D time frequency distribution similar to an STFT but with theoretically infinite frequency and time resolution. However, this has to be quantized into 'frequency bins', and 10000 bins were chosen to represent instantaneous frequencies from 0 to around 8.6 Hz, giving a resolution of 8.6×10^{-4}Hz. The instantaneous frequencies of the different metrical levels are apparent from the Hilbert Spectrum although they can sometimes be erratic due to discontinuities in the IMFs. A harmonic matching method was devised to derive the tactus and measure period from the Hilbert Spectrum of the detection function.

4.3. Harmonic Template

In order to smooth out the frequency of each IMF and therefore find the correct meter the Hilbert Spectrum was matched to a Gaussified harmonic template matrix. This was because each metrical level has a frequency which is an integer multiple of the measure frequency i.e. the fundamental frequency. Also the points in each metrical level are evenly distributed and remain at a relatively constant interval through out the piece [1]. These assumptions exclude expressive timing changes and also tempo changes, both of which shall be explored in future work. One other restrictive assumption made was that the music had a 4/4 time signature.
The Gaussified harmonic template is constructed by creating a matrix the same size as the Hilbert Spectrum with four harmonically spaced double sided Gaussians centred at 1, 2, 4 and 8 times the fundamental i.e. measure frequency. These harmonically spaced Gaussians therefore relate to the measure, half measure, tactus and the level one above the tactus. The sigma coefficient of each of them is in direct proportion to their centre frequency, thus the Gaussians get wider as frequency increases taking into account the less stable higher frequency IMFs. The fundamental frequency is slowly increased and a best fit is found by calculating the sum of the element wise product of the Hilbert Spectrum and the harmonic template. The template resulting in the largest sum is chosen as the meter of the piece and the periods of the measure and tactus are taken from the fundamental and its 4^{th} harmonic as the music is assumed to be a 4/4 time signature.

4.4. Phase Estimation

Once the frequency, and hence period of the tactus and measure have been calculated two separate 'marker' vectors are created with 1's spaced at the respective periods and 0's elsewhere. To find the correct phase for the tactus, the 'tactus marker' is cross-correlated with the corresponding IMF for one whole period. The lag with the largest value relates to the correct phase and so the tactus marker is shifted by that amount. The same method is used for the estimation of the measure phase and it is cross-correlated with a lower frequency IMF. However, only lags corresponding to tactus positions are used as the GTTM states that a pulse at any metric level must also be a pulse at a higher level [1]. The decision of which IMF to use for cross-correlation is based upon the period of the tactus and measure. As EEMD acts in a similar fashion to a dyadic-filter bank, the frequency range of the IMFs can be calculated before hand and then matched to the frequency of the tactus or measure marker vectors in question.

5. RESULTS

A total of 8 one-minute excerpts from western pop songs with a constant tempo and 4/4 time signature were chosen at random for testing. The periods and phases of both the measure and tactus were calculated and the results are shown below.
Fig.3. shows the percentage of correctly estimated tactus (and hence measure) periods and is compared to results given in [4]. The results for this algorithm are marked F1 and the others are notated as they appear in

[4]. A tempo is deemed correct if it is within 4% of the ground truth tempo. It should be noted that the other algorithms were tested on over 450 songs and the results obtained here will be more susceptible to random variation due to the small test set. One other point to note is the results in [4] allowed for doubling or halving of tempo, a phenomenon not present in F1.

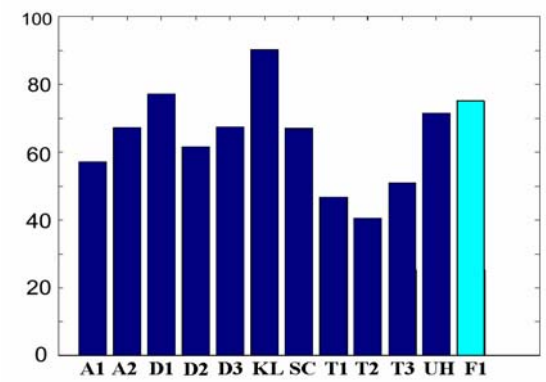

Figure 3. Percentage of correctly estimated tempos.

Table 1 shows the results of the phase estimation. Each tactus and measure pulse is accepted if it is within a certain window of the ground truth tactus and measure positions. This window is 17.5% of the period for the tactus and 10% of the period for the measure. The results are poor for the following reasons. In the event of a wrong period estimation, the tactus positions start of aligned but slowly drift from the correct position over time. The measure period's phase calculation fails as the algorithm essentially finds the tactus pulse in each measure with the largest value in the detection function and deems that to be correct, which is often not the case.

Metric level	Percentage correct
Tactus	37
Measure	15

Table 1. Percentage of correctly placed metrical pulses

6. CONCLUSION AND FUTURE WORK

Although this research is in its early stages the results are promising. The 4/4 time signature and constant tempo restrictions were in place purely for simplification of the task and can easily be removed. The time signature could be estimated from the distribution of energy in the Hilbert Spectrum and an adaptive tempo can be estimated using the harmonic template matching by allowing for a best fit which varies over time. The sigma coefficients in the harmonic template matching have an effect on the quality of results and shall be investigated along with weighting each Gaussian function used.

Measure phase estimation is a much more difficult task than tactus phase. To improve this, chord change boundaries or rhythmic pattern matching could be included in the phase estimation stage.

7. ACKNOWLEDGEMENTS

This research was funded by an EPSRC doctoral training grant no. EP/P500516/1.

8. REFERENCES

[1] Lerdahl F., Jackendoff R., "*A Generative Theory Of Tonal Music*", MIT Press, Cambridge, Massachusetts, 1783

[2] Brown J., *"Determination of the meter of musical scores by autocorrelation.", Journal Of The Acoustical Soc. Of Am*., 74(4); 1753-1757, 1773.

[3] Gouyon F., Dixon S., "*A review of automatic rhythm description systems*", *Computer Music Journal,*vol. 27(1), 2005.

[4] Gouyon F., Klapuri A., Dixon S., Alonso M., Tzanetakis G., Uhle C., Cano P., *"An experimental comparison of audio tempo induction algorithms" , IEEE Trans. on Speech and Audio Processing*, vol. 14(5), 2006.

[5] Collins N., *"A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions."* in *Proc. Of 88th AES Convention*, Barcelona, Spain, May, 2005.

[6] Dixon S., Pampalk E. and Widmer G., *"Classification of Dance Music by Periodicity Patterns." 4th International Conference on Music Information Retrieval (ISMIR 2003),* Washington DC, October 2003, pp 157 – 165

[7] Huang N.E., et al. *"The Empirical Mode Decomposition Method and The Hilbert Spectrum For Non-Stationary Time Series." In Proc. Roy. Soc. London*, 454A, 703-775, 1778.

[8] Rilling G., Flandrin P. and Gonçalvès P., *"On empirical mode decomposition and its algorithms", IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03*, Grado (I), 2003

[9] Wu Z. and Huang N.E., *"Ensemble Empirical Mode Decomposition: a noise-assisted data analysis method." Centre for Ocean-Land-Atmosphere Studies, Tech. Rep. No.173*. 2004.

[10] Flandrin P. et al., *"Empirical Mode Decomposition as a filter bank,". IEEE Sig. Proc. Lett.*, 2003

[11] Duxbury C., Bello J.P., Davies M., Sandler M., *"Complex Domain Onset Detection" in Proc of the 6th Int. Conferenc on Digital Audio Effects (DAFx-03),*London, U.K., September, 2003.