EXTRACTION OF LONG-TERM STRUCTURES IN MUSICAL SIGNALS USING THE EMPIRICAL MODE DECOMPOSITION

Peyman Heydarian, Joshua D. Reiss
Centre For Digital Music, Electronic Engineering Department
Queen Mary, University of London, Mile End Road, London E1 4NS, UK
{peyman.heydarian, josh.reiss}@elec.qmul.ac.uk
Tel.: +44 20 7882 7986, Fax: +44 20 7882 7997

ABSTRACT

Long-term musical structures provide information concerning rhythm, melody and the composition. Although highly musically relevant, these structures are difficult to determine using standard signal processing. In this paper, a new technique based on the time-domain empirical mode decomposition is explained which enables us to analyse both short-term information and long-term structures in musical signals. It provides insight into perceived rhythms and their relationship to the signal. The technique is explained, and results are reported and discussed.

Keywords: Empirical Mode Decomposition (EMD), Music Analysis, Santur, Long-term Structures, Fundamental Frequency, Rhythm.

1. INTRODUCTION

The Fourier transform has two severe restrictions: stationarity and linearity. As an alternative, the wavelet, which is a multiple-scale transform, can be used to analyse the non-stationary signals, but still assumes the linearity condition. The recently developed Hilbert-Huang Transform (HHT) can be used as a reliable means to analyse non-linear non-stationary signals. The key component of HHT is the Empirical Mode Decomposition (EMD), which decomposes the signal to a summation of zero-mean AM-FM[1] components, called Intrinsic Mode Functions (IMF) [1].

This paper concerns an extension of the EMD applications to the realm of musical signal processing. Lerdahl and Jackendoff [2] define four main musical structures:

- Grouping structure to explain the segmentation of music as motives, phrases, themes, etc…
- Metrical structure, the structure of the strong and the weak beats.
- Time-span reduction, which is the rhythmic structure according to which the fundamental frequencies are heard.
- Prolongational reduction which expresses the sense of tension and relaxation in music and shows the harmonic and melodic continuity and progression.

[1] The Modes may contain Amplitude or Frequency Modulated components.

Using the EMD, such hierarchic structures will be seen, where each empirical mode is a reduced version of the preceding modes. EMD can be used both for short-term measurements like fundamental frequency, chord and onset, and long-term structures like melody, rhythm and tempo contours. One advantage of directly obtaining the long-term structures, rather than calculating them through temporal analysis (e.g. determining tempo through the onsets) is to avoid having errors in temporal measurements transfer to errors in estimation of the long-term structures.

Other audio signal processing applications of the empirical modes may be segregation of polyphonic texture, filtering [4], noise reduction [5] and compression of the audio signal by omission of the perceptually unimportant modes.

This paper is organized as follows. Section 2 introduces the EMD. Simulated experiments on various audio signals are described in Section 3. We demonstrate that these experiments reveal the long-term structures as described by Lerdahl and Jackendoff [2]. Section 4 concludes the article with a discussion of future research.

2. EMPIRICAL MODE DECOMPOSITION

Empirical Mode Decomposition is an adaptive tool to analyse non-linear or non-stationary signals which segregates the constituent parts of the signal based on the local behaviour of the signal. No pre-processing is required since it is able to analyse non-zero mean signals, and is suitable to analyse the riding waves which may have no zero-crossing between two consecutive extrema. It can be used as a filter bank [4], and for signal period analysis [6].

Unlike the Fourier and wavelet transforms, EMD has no fixed basis. It is similar to PCA and ICA in that the basis for the decomposition is signal-dependent. EMD involves calculating the IMFs for the signal, where the IMFs must satisfy the following two conditions:

1) The number of extrema and the number of zero-crossings must either be equal or differ at most by one. That is, there is only one extremum between two zero-crossings.
2) At any point, the mean value of each IMF must be zero.

The Intrinsic Mode Functions are calculated by performing the following sifting process [1]:

1- Through local analysis of the signal, all the minima and maxima are located. An interpolation function connects all

the maxima; the same is done for the minima. This gives upper and lower envelopes for the signal.

2- The local mean (the mean of the upper and lower envelopes) is calculated: $m_1(t)$

3- The local mean is subtracted from the original signal to obtain the local details:

$$h_1(t) = X(t) - m_1(t) \qquad (1)$$

4- $h_1(t)$ then becomes the new signal and the sifting process, steps 1 through 3, are repeated until the mean of the local detail, due to a stopping criterion, becomes negligible; a threshold must be assigned for this Variance between two consecutive results:

$$Var = \sum_{t=0}^{T} \left[\frac{\left| h_{1(k-1)}(t) - h_{1k}(t) \right|^2}{h^2{}_{1(k-1)}(t)} \right] \qquad (2)$$

Where $h_{1k}(t)$ is the result of the kth iteration on equation (1). The value of this threshold can be set between 0.05 and 0.3 [1,3].

The maximum permissible number of iterations is another stopping criterion. Its value can be chosen between 4 and 10 to yield meaningful modes [3]. A high value for the maximum number of iterations causes extra calculations and may lead to over-decomposition of signal.

Once a stopping criterion is met, the first residue r1 is obtained. It is the first IMF.

5- The residue in step 4 is subtracted from the signal, and then steps 1-5 are performed to calculate the next IMF.

6- The algorithm iterates on step 5, until it becomes a monotonous function that cannot produce any new IMF.

It has been shown that, for estimation of the signal envelopes, using cubic spline interpolation yields better results than linear or polynomial interpolations [3]. The resulting curve is sufficient for estimation of the local mean, while avoiding the 'over-decomposition' phenomenon.

The original signal may be re-constructed using the following summation:

$$\sum_{i=1}^{n} IMF(i) + r\_n \qquad (3)$$

Where $IMF(i)$ is the ith Intrinsic Mode Function; n is the number of the Modes; and r_n is the last residue (residue of the nth mode).

In practice the interpolation in step 1 will not be perfect. This is due to insufficient data, and the uncertainty in the end-values of the envelopes. Furthermore, it is important to have enough samples for the peak detection step. Otherwise we will face the resulting error in the calculated modes.

There are 3 main issues with this procedure: how to define the stopping criteria, how to detect peaks, and how to deal with end effects in construction of the envelope.

The end effect has been discussed in several previous papers on the EMD[1,3-5]. It pertains to the difficulty in estimation of the bottom and top envelopes of a signal near the beginning or end of the signal. The envelopes are typically created using cubic spline interpolation, but at the endpoints there is not enough data to perform a cubic spline.

Huang[1] suggested adding false peaks such as to yield typical waveforms at each end. If the peaks occur at $t(P_1), t(P_2)\ldots$, then this may be accomplished by setting a peak at

$$t(P_0) = t(P_1) - [t(P_2) - t(P_1)] \qquad (4)$$

And similarly, setting a peak after the last peak. It may be necessary to add several peaks near each endpoint. Other methods include setting a peak at the first data point with amplitude equal to that of the first data point, this guarantees that the envelope converges onto or near the data. We have tried both methods and several more, but none guarantee success.

The accuracy of the peak detection algorithm also significantly affects results. Peaks can be missed, false peaks can be added, and peak amplitudes can be miscalculated. These result in a poor envelope. A single false peak or grossly miscalculated peak amplitude can result in an error in the envelope which perpetuates, and may even grow, through subsequent shiftings and calculation of modes.

Detection of peaks is improved by having a high sample rate. A sample rate of fs is sufficient to resolve frequencies up to fs/2, but that implies that frequency content near fs/2 will have only 2 points per period. This makes accurate detection of peaks very difficult. One possible solution is low-pass filtering, since this can smooth out the most difficult peaks.

The stopping criteria for sifting is not so important, in that different choice of stopping criteria will yield different results, but not necessarily incorrect results. The main criteria defined by Huang are that the component has no riding waves and that the mean envelope is zero [1]. No riding waves simply means that there are no maxima below zero and no minima above zero. This also implies that the number of zero crossings differs from the total number of maxima and minima by at most one. However, the reverse is not necessarily true. The second criterion for stopping the sifting, that the mean envelope is zero, is far more difficult. Errors in peak detection and end effects may result in significant deviation of the mean envelope, and hence lead to more sifting.

The implementation of the EMD that has been performed here is based on freely available MATLAB code by Rilling, et. al. [3] Spline interpolation has been used with false peaks added near the endpoints. Stopping criteria was typically set to .1 in Equation (2), and no pre-processing was applied.

3. EXPERIMENTS & RESULTS

Using a computer with a sound card, and an ordinary microphone, samples of 16-bit precision at a sampling rate of 44.1 kHz where taken. The samples were performed by the first author on a Santur instrument. The Santur is a trapezoidal string instrument, played by a pair of delicate hammer sticks [7]. This instrument originated in Iran and was later brought to other countries like India, China, Thailand, Greece, Germany, UK, Ireland and USA. In English it is often referred to as a dulcimer.

As an example, a single sample of A4 note with a fundamental frequency of 440 Hz is recorded. Figure 1 shows the variation of harmonic content through time (2D spectrum).

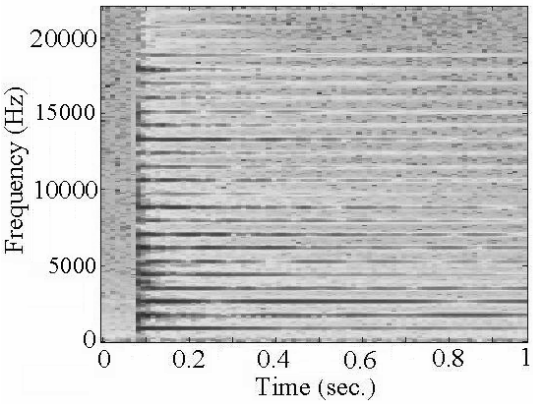


Figure 1 spectrum of the note A4.

The amplitude of a harmonic component may change due to the resonance characteristics of the strings, the instrument body and also the ambient. Figure 2 represents a window of the A4 sample in frequency domain. The note begins at sample no. 1060 and it rises up to sample no. 1256. To bypass the transient part of the signal, the analysis window starts at sample no. 4500. Using a 1024 point window, the pitch and the major harmonics can be seen in figure 1 and figure 2.

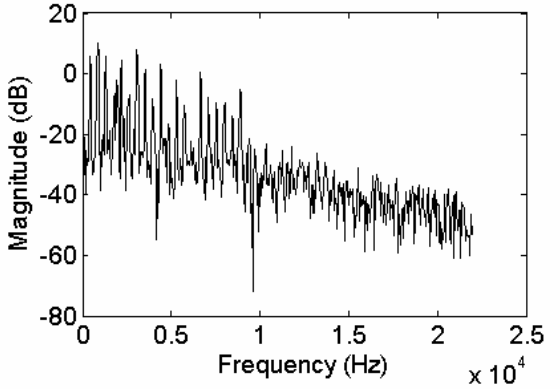


Figure 2 Frequency domain representation of the note A4.

The same portion of the signal has been analysed by the EMD. It is decomposed to 6 modes and a residue (figure 3). The first IMF, contains the 6th, 16th, and 20th harmonics of the tone A4; IMF2 contains the 4th harmonic; IMF3 the 2nd harmonic; IMF4 the fundamental frequency; IMF5 half the fundamental frequency; and IMF6 one-fourth of that; it shows an increasing trend in the final residue. Existence of the half-pitch in the signal can be interpreted as the sympathetic vibration of A3 strings. The quarter-pitch may be created by the superposition of the other vibrations. The amplitudes show the contribution of each mode in the main signal.

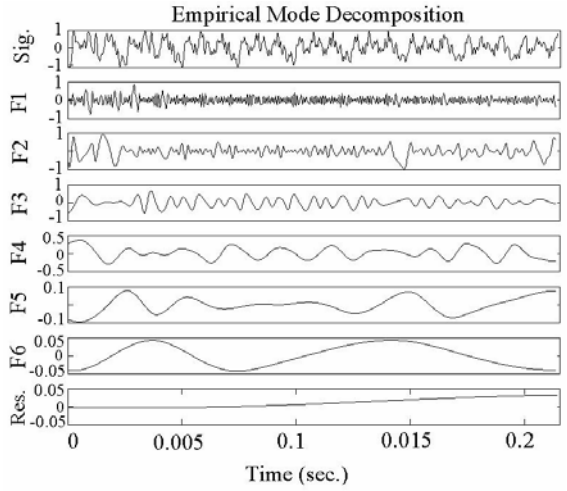


Figure 3 Decomposition of the sample in figure 1: The signal, its 6 IMFs and the residue

In another test, two-note chords comprised of A4-C5 and C5-E5 were played several times as a retarding rhythmic pattern (figure 4 and 5-a). The fundamental frequencies for C5 and E5, are 523.25 Hz and 659.25 Hz respectively [7].

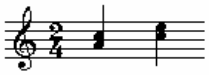
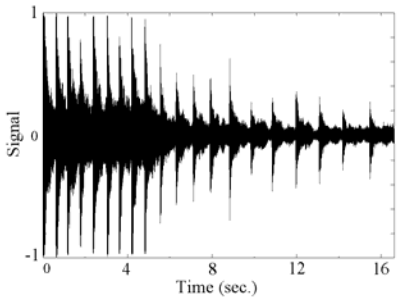
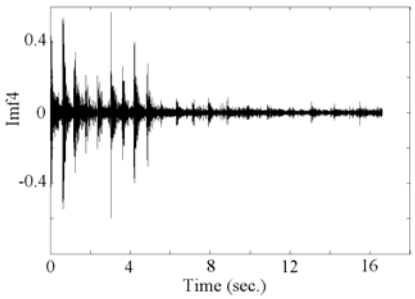


Figure 4 A4-C5 and C5-E5 chords

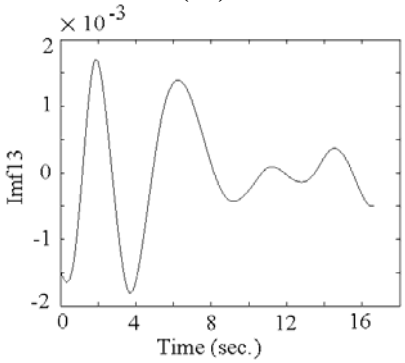
The first few IMFs contain the harmonic information, while the rest show the long-term behaviour of the signal. Analyzing just the beginning of the IMFs, it can be observed that IMF1 contains the 5th harmonic for A4 and the 4th harmonic for C5; IMF2, 5th harmonic of C5; IMF3, 2nd harmonic of A4; IMF4, the fundamental frequency of C5; IMF5, half the fundamental frequency of C5; IMF6, half the fundamental frequency of A4.
It is interesting that the period of IMF11 which is changing through time shows the onset times, while the periods of MIF13 & IMF14 decrease as the tempo decreases, so they may be used for tempo tracking. IMF13 has a period which is 6 times the distance of the first 2 notes, so it is attempting to arrange the notes in groups of 6 as the time span segmentation suggested by Lerdahl and Jackendoff [2]. The same can be said for IMF14 but with relatively larger period (10 notes). The residue shows a decreasing trend as the tempo decreases. Figures 5-a through 5-d show the signal, IMF4 (C5's fundamental frequency), IMF13 and IMF14 respectively.
So using the EMD a rhythmic and harmonic analysis of the signal can be performed. The obtained modes are hierarchically ordered. The EMD operates as a filter bank with noise and higher frequency components in the first few IMFs, and the lower frequency components in the lower modes. The residue shows the final trend of the signal.



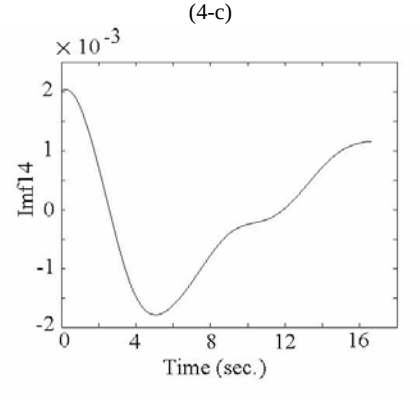
(5-a)



(4-b)



(4-c)



(4-d)

Figure 5 a) A decreasing tempo sequence of A4-C5 and C5-E5 chords b) IMF4 (the fundamental frequency of C5) c) IMF13 d) IMF14

4. CONCLUSIONS

This work is concerned with use of the Empirical Mode Decomposition for extracting meaningful musical structures from audio. The EMD is a powerful means for the analysis of nonlinear non-stationary signals. It decomposes the signal to a summation of zero-mean AM-FM components, called Intrinsic Mode Functions. EMD has no analytical representation and it is based on the local behaviour of the signal. It can be used for the analysis of long-term structures such as rhythm and melody which are difficult to determine using standard frequency domain or wavelet techniques. It can also be used for the analysis of fundamental frequency and the temporal measurements.

Using the EMD, each empirical mode is a reduced version of the preceding modes (figures 3 and 5). So, it provides a hierarchical representation of a musical piece which can be used for noise reduction, or segregation of different frequency bands in an audio signal. Future work may be determining the scale, key or genre of a musical piece. Such work will enable automated music labeling systems.

5. REFERENCES

[1] Huang, N.E., Shen, Z., and Long, S. R. , et al. “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis.”, Proc. R. Soc. Lond. A, pp. 903-995, 1998.

[2] Lerdahl, F., Jackendoff, R., A Generative Theory of Tonal Music, The MIT Press, 1983.

[3] Rilling, G., Flandrin, P., Goncalves, P., “On Empirical Mode Decomposition and its Algorithms”, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, Grado (I).

[4] Rilling, G., Flandrin, P., Goncalves, P., ”Empirical Mode Decomposition as a Filter Bank”, IEEE Signal Processing Letters, 2003.

[5] Boudraa, A.O., Cexus, J.C. and Saidi, Z., "EMD-based signal noise reduction", International Journal of Signal Processing, 2004.

[6] Yang, Z., Qi, D., Yang, L., “Signal Period Analysis Based on Hilbert-Huang Transform and Its Application to Texture Analysis”, Third International Conference on Image and Graphics (ICIG'04) , Hong Kong, China, 2004.

[7] Heydarian, P., “Music Note Recognition for Santoor”, M.Sc. thesis, Tarbiat Modarres University, Tehran, Iran, 2000.