Audio Source Separation from the Mixture using Empirical Mode Decomposition with Independent Subspace Analysis

[1]Md. Khademul Islam Molla, [1]Keikichi Hirose, [2]Nobuaki Minematsu

[1]Graduate School of Frontier Sciences, [2]Graduate School of Information Science and Technology
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
{molla, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

In this paper we decompose the Hilbert Spectrum of an audio mixture into a number of subspaces to segregate the sources. Empirical mode decomposition (EMD) together with Hilbert transform produces Hilbert spectrum (HS), which is a fine-resolution time-frequency representation of a non-stationary signal. EMD decomposes the mixture signal into some intrinsic oscillatory modes called intrinsic mode function (IMF). HS is constructed from the instantaneous frequency responses of IMFs. Some frequency independent basis vectors are derived using independent component analysis (ICA). Kulback-Laibler divergence based k-means clustering algorithm is proposed to group the basis vectors into number of desired sources. Then projecting HS on to the grouped basis vectors derives the independent source subspaces. The time domain source signals are assembled by applying some post processing on the subspaces. We have also produced some experimental results using our proposed separation algorithm.

1. Introduction

The usual approach of single mixture audio source separation is to project the mixture on to the time-frequency plane and to analyze auditory scenes. Most of the CASA (computational auditory scene analysis) based approaches employed Fourier transform method for time-frequency representation of mixture signal assuming that the audio signal is piecewise stationary [1, 2]. In [3], a number of sources from two mixtures have been separated considering the phase and amplitude variations between two sensors. They have also assumed that not more than one source is active at any time-frequency point in the spectrogram. N. Roman et al [4] used binaural mixtures of the convolution of speech and various noises from different azimuth and separated the speech signals using binary masking method. Sam T. Roweis [5] proposed learning based statistical pattern recognition process to separate the sources from single mixture.

Our proposed separation algorithm can separate the audio sources from their single mixture without any prior knowledge about the sources. This system has also taken into account that the audio signals are mostly non-linear and non-stationary. Empirical Mode decomposition (EMD) is a new technique for nonlinear and non-stationary time series analysis (Huang *et al* [6]) has recently been used in many signal processing application including water and wind wave analysis, tone separation etc. In this paper, we have employed the EMD together with Hilbert transform for time-frequency representation of the audio signals. EMD decomposes the mixture signal as collection of some intrinsic mode functions (IMFs) and this action is very much similar to the filter bank analysis [7]. Instantaneous frequency of each IMF is calculated using analytic signal method. The Hilbert Spectrum (HS) of the mixture signal is constructed by properly arranging the frequency responses of the IMFs along time and frequency axes.

The data space corresponding to the instantaneous frequency response of the IMFs and the mixture signal are used to derive some frequency independent basis vectors using ICA. Kulback-Laibler divergence (KLd) based clustering algorithm is employed to group the basis vectors into the number of sources. The projection of the HS to each group of frequency basis corresponds to a subspace of the individual source signal. We have simulated our proposed system and presented some experimental results to separate the sources from single mixture of two audio signals. The results show that this method can produce applicable results in source separation arena.

Regarding the organization of this paper, we have described the basics of EMD in section two, the proposed separation algorithm is presented in detail in section three, some experimental results are demonstrated in section four, and section five contains some concluding remarks and also future planes with this work.

2. EMD Basics

Empirical mode decomposition is an adaptive process to decompose a signal into oscillating components obeying some basic properties. EMD has recently been pioneered by N.E. Huang *et al.* for representing non-stationary signals as sums of zero-mean AM-FM components [6].

The principle of the EMD technique is to decompose a signal $s(t)$ into a sum of the functions $imf_i(t)$ called intrinsic mode function (IMF). Each IMF satisfies two conditions: (i) in the whole data set the number of extrema and the number of zero crossing must be same or differ at most by one, (ii) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. The first condition is similar to the narrow-band requirement for a stationary Gaussian process and the second condition modifies a global requirement to a local one, and is necessary to ensure that the instantaneous frequency will not have unwanted fluctuations as induced by asymmetric waveforms [8]. There exist many algorithmic approaches of EMD [7], [8]. At the end of EMD the signal $s(t)$ is represented as:

$$s(t) = \sum_{i=1}^{n} imf_i(t) + r_n(t) \quad (1)$$

where n is the number of IMFs and $r_n(t)$ is the final residue.

Another way to explain how EMD works is that it extracts out the highest frequency oscillation that remains in the signal. Thus, locally, each IMF contains lower frequency oscillations than the one extracted just before. An audio mixture signal

(mixture of speech and flute sound) and the decomposed IMFs are shown in Figure 1.

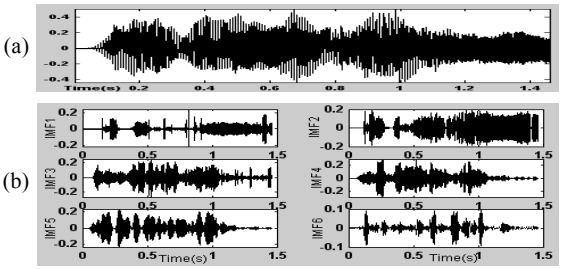


Figure 1. EMD of an audio mixture, (a) mixture (speech and flute sounds) audio (b) first six IMFs out of total 14 IMFs.

2.1. Instantaneous Frequency

Every IMF is a real valued signal. Analytic signal method is used to calculate the instantaneous frequency of the IMFs. The analytic signal corresponding to a real signal $x(t)$ is defined as

$$z(t) = x(t) + jH[x(t)] = z(t) = a(t)e^{j\theta(t)} \quad (2.1)$$

where $H[]$ is the Hilbert transform operator, $a(t)$ and $\theta(t)$ are instantaneous amplitude and phase respectively. So the instantaneous frequency ω(t) can easily be derived as:

$$\omega(t) = \frac{d\theta(t)}{dt} \quad (2.2)$$

Using Equations (2.1) and (2.2), the analytic signal associated with each of the IMFs and thus the instantaneous frequency of each of them is calculated.

2.2. Hilbert Spectrum

Hilbert Spectrum describes the joint distribution of the amplitude and frequency content of the signal as a function of time. This distribution is designated as Hilbert amplitude spectrum $H(\omega,t)$ or simply Hilbert spectrum. To build $H(\omega,t)$, the instantaneous frequency of each IMF is first scaled according to the given frequency bins. Then for every $imf_i(t)$, if $\omega_i(t)$is the corresponding instantaneous frequency, we represent the time-frequency plane the triplet $\{t, \omega_i(t), a_i(t)\}$ where $a_i(t)$ is the amplitude of the analytic signal associated to $imf_i(t)$. It is noted that the time resolution of H is equal to the sampling rate. Figure 2 represents the Hilbert spectrum of the audio signal of Figure 1 using 256 frequency bins.

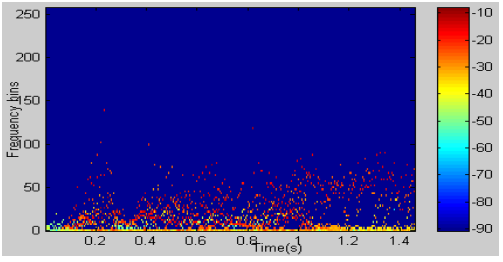


Figure 2. Hilbert amplitude spectrum (or simply Hilbert spectrum) using 256 frequency bins. The amplitude is in dB.

During the construction of the Hilbert spectrum, the phase matrix $\phi(\omega,t)$ representing the phase information corresponding to each time-frequency point of $H(\omega,t)$ is saved to be used in re-synthesis of the extracted sources.

3. Source Separation Algorithm

We have already derived the Hilbert (amplitude) spectrum (HS) from the time domain signal. Our approach of source separation is to decompose the Hilbert spectrum of mixture signal into a number of HSs corresponding to each independent source.
The overall Hilbert spectrum H can be represented as the superposition of N independent source Hilbert spectrums h_i as:

$$H = \sum_{i=1}^{N} h_i \quad (3.1)$$

h_i is also uniquely represented as the outer product of an invariant frequency basis vector F_i, and corresponding amplitude envelope A_i (time basis vector) which describes the magnitude variation of the frequency basis vectors over time [2].

$$h_i = F_i A_i^T \quad (3.2)$$

For the source containing b basis vectors-

$$F_i = [f_1^{(i)} f_2^{(i)} \cdots f_b^{(i)}] \quad (3.3)$$
$$A_i = [a_1^{(i)} a_2^{(i)} \cdots a_b^{(i)}]$$

Then corresponding h_i is represented as

$$h_i = \sum_{j=1}^{b} f_j^{(i)} a_j^{(i)} \quad (3.4)$$

So by summing all the Hilbert spectrums:

$$H = \sum_{i=1}^{N} F_i A_i^T \quad (3.5)$$

The amplitude weighting vector A_i corresponding to frequency (independent) basis vector F_i is obtained by projecting H against F_i as:

$$A_i = F_i^T H \quad (3.6)$$

To decompose the spectrum H into some independent h_i, it is urged to determine some frequency independent basis over the whole H. It is also assumed that the derived frequency independent basis vectors are stationary over the whole time sequences of H. To meet this condition the mixture signal is segmented into some blocks with almost stationary energy distribution [1,2].

3.1. Selection of Data Space

The Hilbert spectrum is actually the resultant effect of the Hilbert spectrums of individual IMF [6]. Instead of decomposing the high dimensional data space H we have derived the new vectors each of which is the spectral projection of the mixture on to an IMF. These projection vectors are then used to derive some frequency independent basis vectors to produce the source subspaces. The spectral projection of the mixture signal x on to the ^{n}th IMF is defined as

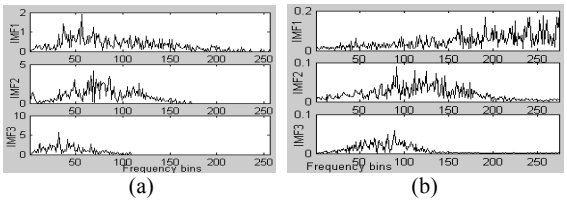
$$P_{xn}(\omega) = \frac{|C_{xn}(\omega)|^2}{S_x(\omega)s_n(\omega)} \quad (3.7)$$

where $C_{xn}(\omega)$ is the cross marginal spectrum of mixture and nth IMF, $S_x(\omega)$ and $s_n(\omega)$ are the marginal power spectra of mixture and nth IMF respectively at frequency index ω.

The projection term $P_{xn}(\omega)$ is a quantitative measure of how much the mixture is correlated with ^{n}th IMF at $^{\omega}$th frequency band. There some benefits to use the spectral projection vectors instead of using whole H for subspace decomposition. It reduces the computational complexity and boosts the convergence during ICA. If $imfH_n(\omega,t)$ is the Hilbert spectrum of ^{n}th IMF, its corresponding marginal spectra can be defined as:

$$s_n(\omega) = \int_0^T imfH_n(\omega,t)dt \quad (3.8)$$

where T is the total data length. The marginal spectra of the mixture is calculated by replacing $imfH_n(\omega,t)$ by $H(\omega,t)$.The marginal spectra of first three IMFs using 256 frequency bins are shown in Figure 3(a). The Fourier transform (512 point FFT) of that IMFs are also presented in Figure 3(b). It is observed that, $s_n(\omega)$ has a totally different meaning from the Fourier spectra (*for detail* [6]). The data matrix X containing the spectral projections of mixture on the IMFs is used to derive the frequency independent basis vectors by using ICA as described in the following sections.



(a) (b)

Figure 3. Frequency response of the first four IMFs. (a) Marginal spectra using 256 frequency bins (b) Fourier spectra using 512 point FFT (displaying only 256 point)

3.2. Constructing Independent Subspaces

Usually the number vectors in the data matrix X is greater than the number of frequency basis vectors required for subspace decomposition. The dimension of the spectral projection matrix X is first reduced by principal component analysis (PCA) [1, 2].

The basis vectors obtained by PCA are only uncorrelated but not statistically independent. To derive the independent basis vectors a further procedure called ICA must be carried out. The ICA model [9] expresses the observation signals x (comes from PCA, i.e. reduced X) as the product of mixing matrix A and vectors of statistically independent signals s,

$$x = As \quad (3.9)$$

where A is $l \times m$ (pseudo-) invertible mixing matrix with orthogonal columns, s is random vector of m signals, and X is an l dimensional vector of observation with $l \geq m$. JadeICA algorithm [10] is used here to estimate the demixing matrix W such that:

$$F = Wx = WAs \quad (3.10)$$

where F is the collection of independent basis vectors.

Once the frequency independent basis vectors F have been obtained the corresponding amplitude envelopes A can be obtained by Equation (3.6). The basis vectors are then grouped into the number of sources. Vectors F and A are grouped into F_i and A_i subsets respectively. For a two source mixture problem $i=1,2$ i.e. two subsets of F and A. Then the Hilbert spectrum of individual source is constructed using each group of F and A as in Equation (3.2).

We have introduced a Kullback-Laibler divergence (KLd) based k-means clustering algorithm for the grouping process. Symmetric KLd measures the relative entropy between two probability mass functions $p(x)$ and $q(x)$ over a random variable X as:

$$KLd(p,q) = \frac{1}{2}\{\sum_{x \in X} p(x)\log\frac{p(x)}{q(x)} + \sum_{x \in X} q(x)\log\frac{q(x)}{p(x)}\} \quad (3.11)$$

Each basis vector is normalized and transformed to its corresponding probability mass function. Then KLd is used for distance measure between two basis vectors during k-means clustering whereas traditional k-means measures the Euclidean distance. KLd being information theoretic measure performs better. The value of k is selected manually.

3.3. Source Re-synthesis

We have implemented a reversible process from Hilbert spectrum to get back the time domain signal to re-synthesis the extracted source signals. As shown in equation (2.1) Hilbert transform only imposes the imaginary part (keeping same the real component) to produce the analytic signal. It implies that time domain signal is assembled by filtering out the imaginary part from the HS. When the Hilbert spectrum (h_i) of every source is derived, the spectrum of the real component of each time-frequency point is calculated by element wise multiplication of h_i and the cosine of the phase matrix ϕ as:

$$h_i^R(\omega,t) = h_i(\omega,t) \cdot \cos[\phi(\omega,t)] \quad (3.12)$$

Then the time domain source signal $s_i(t)$ is obtained by summing the real components over the frequency bins for every time instant as:

$$s_i(t) = \sum_{\omega} h_i^R(\omega,t) \quad (3.13)$$

The same process is repeated for each stationary block of he mixture. Concatenating the individual source of every block using windowed overlap-add method produces the whole separation.

4. Experimental Results

We have used some mixtures of two audio streams to test the efficiency of our proposed separation algorithm. The individual test stream is a mixture of male speech and other sound like female speech, violin, telephone ring, jazz music, white noise etc. The audio streams m1, m2, m3 and m4 are the mixture of male speech plus telephone ring, jazz music, female speech and flute sound respectively. All mixtures are with 16kHz sampling rate and 8-bit amplitude resolution. We have applied the separation algorithm on the audio segments with no large variance of spectral distribution and then concatenate the

extracted source signals (of every segment) to produce the separation over the entire audio stream.

It is not easy to propose a strong feature for quantitative measure of the separation performance. The average value of the running short-term relative energy between original and separated signal is used here for quantifying the separation efficiency. It is termed here as original to separated signal ratio (OSSR) and mathematically defined as:

$$OSSR\ (t) = \log 10 \left(\frac{\sum_{i=1}^{w} s_{original}{}^{2}(t+i)}{\sum_{i=1}^{w} s_{separated}{}^{2}(t+i)} \right) \quad (4)$$

where $s_{original}$ and $s_{separated}$ are the original and separated signal respectively, w is window length and it is 10 ms here. In the case for zero energy in a particular window, no OSSR measurement is performed.

This OSSR calculates the relative short-term energy level between those two signals. It is used here to measure the difference between two signals in terms of short-term energy level. If the two signals are similar, the OSSR produces 0 value and any other value (positive or negative) is a measure of their dissimilarity. Table 1 shows the average OSSR of each signal for every mixture. Smaller deviation of OSSR from 0 indicates the higher degree of separation.

Table 1: The experimental results of our proposed algorithm. Each value represents relative short-term energy level between original and after separation of that signal.

Mixtures	OSSR of Sig1	OSSR of Sig2
m1	-0.2403	-0.0570
m2	-0.3009	0.0128
m3	-0.4320	-0.1570
m4	0.3766	0.1620

Table 2: The separation performance and audio quality of the re-synthesized signals by human evaluation (by hearing). The performance score is between 1 (for min) and 5 (for max).

Mixtures	Sig1		Sig2	
	Separation performance	Audio quality	Separation performance	Audio quality
m1	4.75	4.25	4.5	4.25
m2	3.75	4.0	3.75	3.75
m3	3.50	3.50	4.0	3.75
m4	4.0	3.75	4.75	4.25

Besides the quantitative measure of separation efficiency, hearing by human is perhaps the best way to ensure the separation performance and the audio quality (how much the separated signal is clear in hearing relative to the original one) of the re-synthesized signal. We have asked five people who are directly related to speech processing research to evaluate the separation performance and the audio quality of the separated signals. They have scored the performance between 1 and 5 for highest and lowest performance respectively. The average results of human evaluation are presented in Table 2. Based on quantitative and human evaluation, it is observed that the separation efficiency is noticeable in this research area.

5. Discussion and Conclusions

In this paper a method for single mixture audio source separation using EMD and ICA is presented. EMD has many uses for wave data analysis and recently it is using as a filter-bank analysis. We have technically employed it for audio source separation. The specialty of the Hilbert spectrum is that the time resolution can be as precise as the sampling period and the frequency resolution depends on the choice (it should not be the power of 2 as in Fourier method) up to Nyquist frequency. When the type of analyzing signal is known the required time frequency resolution can be defined in prior. However, if we donít know anything about the signals inside the mixture, the better time-frequency resolution obviously performs better in separation. The separation efficiency is presented as the average amount of signal to mixture energy. Also the separation performance and the audio quality of the separated signals are evaluated by hearing. The experimental results are sound good.

Additional post-processing is also necessary to improve the audio quality of the extracted signals. An enhancement is expecting by implementing the EMD process as a perceptually tuned filter bank instead of simple EMD. The automatic detection of number of sources in a given mixture stream, their proper separation and the improvement of the robustness of separation process will be the main concern for our future works.

6. References

[1] Casey M.A. and A. Westner, ìSeparation of Mixed Audio Sources by Independent Subspace Analysisî, *International Computer Music Conference*, 2000.

[2] Christian Uhle, Christian Dittmar and Thomas Sporer, ìExtraction of Drum Tracks from Polyphonic Music using Independent Subspace Analysisî, *ICA2003*, Nara, Japan, April 2003.

[3] Jourjine, S. Rickard and O. Yilmaz, ìBlind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixturesî, *ICASSP*, 2000.

[4] N. Roman, D. Wang and G. J. Brown, ìSpeech Segregation Based on Sound Localizationî *Journal Acoustic Society of America*, 114(4): 2236-2252, 2003.

[5] Sam T. Roweis, ìOne Microphone Source Separationî, *NIPS*, pp. 793-799, 2000.

[6] N.E Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung and H.H. Liu, ìThe empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysisî, *Proc. Roy. Soc. London A*, Vol. 454: 903-995, 1998.

[7] P. Flandrin, G. Rilling and P. Goncalves, ìEmperical Mode Decomposition as a filter bankî, *IEEE Sig. Proc. Letter,* (in press), 2003.

[8] Ivan Magrin-Chagnolleau and Richard G. Baraniuk, ìEmpirical mode decomposition based frequency attributesî, *Proceedings of the 69th SEG Meeting*, Texas, USA, 1999.

[9] A. Hyv%inen and E. Oja, ìIndependent Component Analysis: Algorithms and Applicationsî, *Neural Networks*, 13(4-5): 411-430, 2000.

[10] J.F. Cardoso and A. Souloumiac, ìBlind beamforming for nonGaussian signalsî, *IEEE Proceedings*, 140(6): 362-370, 1993.