LOCALIZATION BASED SEPARATION OF MIXED AUDIO SIGNALS WITH BINARY MASKING OF HILBERT SPECTRUM

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

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

ABSTRACT

This paper presents a method of audio signal separation from stereo mixtures using binary masking in time-frequency (TF) domain based on the spatial location of the audio sources. The TF representation of audio signal is obtained by Hilbert spectrum (HS). The Hilbert transformation together with empirical mode decomposition (EMD) produces HS which is a fine-resolution TF representation of any nonlinear and non-stationary signal. The sources are localized in the space of time and intensity differences between two microphones' signals. The separation is performed by masking the target signal in TF domain considering that the sources are disjoint orthogonal. The experimental results of the proposed method show a noticeable improvement of separation efficiency.

1. INTRODUCTION

The multi-source audio environment is a crucial situation for humanoid robotics to segregate and recognize a particular sound. Such acoustical situation requires the technique to separate the audio sources from stereo mixtures as implemented in [1, 2, 3, 4]. When the sources are located at different spatial locations, the time difference (TD) and intensity difference (ID) are introduced between the mixed signals recorded by two microphones. The separation is performed by time-frequency (TF) masking based on the TD-ID space localization of the sources [3, 4]. It is assumed that the sources are disjoint orthogonal i.e. only one source is active at any point in TF space. The short-time Fourier transform (STFT) is employed in TF representation of the mixture signals. The STFT based TF representation includes a remarkable amount of cross-spectral energy due to the harmonic assumption and window overlapping. The both time and frequency resolution can not be extended independently. Those two limitations of STFT based TF representation degrades the disjoint orthogonality of the audio sources and hence the separation efficiency by using masking method in TF domain. The separation efficiency can be improved by maximizing the resolution and minimizing the cross-spectral energy terms in TF space.

This paper presents a technique to the individual audio sources from stereo mixtures based on their spatial locations. The HS which does not include noticeable amount of cross-spectral energy terms is employed here in TF representation. The empirical mode decomposition (EMD), a new technique for nonlinear and non-stationary time series analysis [5] and Hilbert transformation are employed together to derive HS. The EMD decomposes the mixture signal as a collection of some oscillatory basis components termed as intrinsic mode functions (IMFs) containing some basic properties [5, 6]. It can also be considered as dyadic filter-bank as being proved by the analysis of white noise [6, 7]. A modified version of the original EMD method is introduced here. Instantaneous frequency of each real valued IMF is calculated by applying Hilbert transform. The Hilbert Spectrum (*HS*) of the mixture signals are constructed by properly arranging the frequency responses of the individual IMF along time and frequency axes. Based on the TD and ID between two mixtures, the TF spaces (HSs of two mixtures) are clustered to localize the audio sources in TD-ID space. The TF space of each source is segregated by binary masking method [4], and the time domain signals are recostructed by applying the inverse transformations. The HS has better TF resolution as well as less cross-spectral energy and hence more suitable for disjoint orthogonality assumption of audio sources.

Regarding the arrangement of this paper, the basics of EMD method and the Hilbert spectrum are described in Scetion 2, sources localization and separation methods are illustrated in Section 3 and 4 repectively. The measure of disjoint orthogonality is described in Section 5. The experimental results are presented in Section 6 and finally the conclutions are included in Section 7.

2. THE EMD AND HILBERT SPECTRUM

The empirical mode decomposition (EMD) technique decomposes any signal $s(t)$ into a set of band-limited functions $C_m(t)$ called intrinsic mode functions (IMFs). Each IMF should satisfy two basic conditions: (i) in the whole data set, the number of extrema and the number of zero crossing must be the same or differ at most by one, (ii) the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is always zero. There exist many approaches of computing EMD [6]. The following algorithm is adopted here to decompose the signal $s(t)$ into a set of IMF components.

a) Initialize the residual $r_0(t)=s(t)$ and index of IMF $m=1$
b) Compute m^{th} IMF
c) (i) set $g_0=r_{m-1}$ and $i=1$
(ii) Identify the extrema (minima and maxima) of $g_{i-1}(t)$
(iii) Compute upper and lower envelopes $u_{i-1}(t)$ and $l_{i-1}(t)$
(iv) Find mean envelope $\mu_{i-1}(t)=[u_{i-1}(t)+l_{i-1}(t)]/2$

(v) Update $g_i(t)=g_{i-1}(t)-\mu_{i-1}(t)$ and $i=i+1$

(vi) Repeat steps (ii)-(v) until $g_i(t)$ being an IMF. If so, the m^{th} IMF $C_m(t)=g_i(t)$

d) Update residual $r_m(t)=r_{m-1}(t)-C_m(t)$

e) Repeat steps (b) to (c) with the index of IMF $m=m+1$

At the end of the decomposition the signal s(t) is represented as:

$$s(t) = \sum_{m=1}^{M} C_m + r_M \qquad (1)$$

where M is the number of IMF components and r_M is the final residue. The r_M monotonously converges to a constant or takes a function with only one maxima and minima such that no more IMF can be derived. The mixed (speech and flute sound) audio signal and the decomposed IMF components are shown in Fig.1.

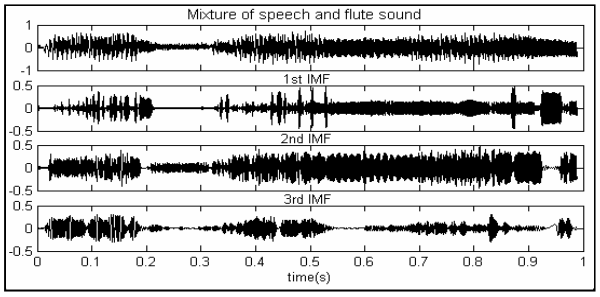

Fig. 1. The EMD of an audio mixture (speech and flute sound) showing first three IMFs out of 15

It is noticed that EMD yields a small number of modes (IMFs) that completely fall inside the frequency range of the original data. During the sifting process most of IMFs include signals at frequencies that cannot be associated with the data. To eliminate such unwanted signals, a band-pass filtering method is proposed to be included in the original EMD algorithm.

This attempt ensures to run every IMF inside the given frequency band. The proposed modification also increases the number of IMF components that improves the frequency resolution of derived HS. The analyzing signal $s(t)$ is passed through a zero phase band-pass filter (BPF). The same filter is included in step (vi) of the original algorithm. The procedure is as follows: first generate the IMF $C_m(t)$, filter it to yield the filtered IMF $\hat{C}_m(t)$ and compute the residue $\hat{r}_m(t)=r_{m-1}(t)-\hat{C}_m(t)$ to generate $\hat{C}_{m+1}(t)$. After completing the decomposition, the modified EMD can be represented by the same way as in Eq. (1). Experimentally it is found that the modified EMD generates 25 IMFs whereas, original one produces 15 IMFs from the same signal of Fig. 1.

2.1. Instantaneous Frequency

Instantaneous frequency (IF) represents signal's frequency at an instance, and is defined as the rate of change of the phase of the "analytic" version of the signal with respect to time. Every IMF is a real valued signal. The discrete Hilbert transform (HT) denoted by $H_d[.]$ is used to compute the analytic signal for an IMF. The HT provides a phase-shift of $\pm\pi/2$ to all frequency components, whilst leaving the magnitudes unchanged [5].Then the analytic version of the m^{th} IMF $\hat{C}_m(t)$ is defined as:

$$z_m(t)=\hat{C}_m(t)+jH_d[\hat{C}_m(t)]=a_m(t)e^{j\theta_m(t)} \qquad (2)$$

where $a_m(t)$ and $\theta_m(t)$ are instantaneous amplitude and phase respectively of the m^{th} IMF. The analytic signal is advantageous in determining the instantaneous quantities such as energy, phase and frequency. The IF of m^{th} IMF is then given as the derivative of the phase $\theta_m(t)$: $\omega_m(t)=\frac{d\tilde{\theta}_m(t)}{dt}$ where $\tilde{\theta}_m(t)$ represents the unwrapped version of instantaneous phase $\theta_m(t)$. The median smoothing is applied to reduce the discontinuities of the IF obtained by discrete time derivative.

2.2. Hilbert Spectrum

Hilbert Spectrum represents the distribution of the signal energy as a function of time and frequency. It is also designated as Hilbert amplitude spectrum $H(f,t)$ or simply Hilbert spectrum (HS). This process first normalizes all IF vectors between 0 to 0.5. Each IF vector $\omega_m(t)$ is multiplied by the scaling factor $\eta=0.5/(IF_{max}-IF_{min})$, where $IF_{max}=\text{Max}(\omega_1,\omega_1,\ldots,\omega_M)$ and $IF_{min}=\text{Min}(\omega_1,\ \omega_1,\ldots,\ \omega_M)$. The bin spacing of the HS is $0.5/B$, where B is the number of desired frequency bins. The overall HS is expressed as the superposition of the HS of individual IMF defined as: $H(f,t)=\sum H_m(f,t)$ for $m=1,\ 2,\ldots\ldots,M$ [6]. Hence, each element $H(f,t)$ of the overall HS is defined as the weighted sum of the instantaneous amplitudes of all the IMFs at f^{th} frequency bin,

$$H(f,t)=\sum_{m=1}^{M} a_m(t)w_m^{(f)}(t) \qquad (3)$$

where the weight factor $w_m^{(f)}(t)$ takes 1 if $\eta\times\omega_m(t)$ falls within f^{th} band, otherwise is 0. After computing the elements over the frequency bins, H represents the instantaneous signal spectrum in TF space as a 2D table.

It is noted that the time resolution of H is equal to the sampling rate and the frequency resolution can be chosen up to Nyquest limit. Fig. 2 represents the Hilbert spectrum of the audio signal shown in Fig. 1 using 256 frequency bins with 16kHz sapling rate.

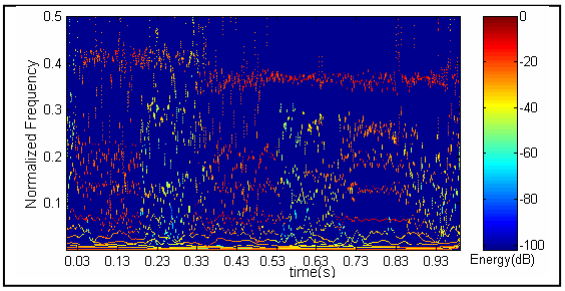

Fig. 2. Hilbert spectrum of audio mixture with 256 frequency bins

3. SOURCE LOCALIZATION

The audio signals placed at different azimuth (0^o to 180^o) locations are recorded by two omni-directional homogeneous

microphones. There is a one-to-one mapping between the azimuth location and a region in TD-ID space, and hence the sources are localized in TD-ID space. The TD and ID are computed from the relative phase and energy differences of the TF spaces of two mixtures. If $H_L(f,t)$ and $H_R(f,t)$ are the Hilbert spectrum of the mixtures $x_l(t)$ and $x_r(t)$ respectively, the TD and ID can easily be computed as [1, 4]:

$$TD(f,t) = \frac{1}{k}[\tilde{\phi}_L(f,t) - \tilde{\phi}_R(f,t)] \quad (4)$$

$$ID(f,t) = 20\log\left(\frac{|H_R(f,t)|}{|H_L(f,t)|}\right)$$

where $\tilde{\phi}_L(f,t)$ and $\tilde{\phi}_R(f,t)$ are the unwrapped phase matrices corresponding to $H_L(f,t)$ and $H_R(f,t)$ respectively. The differences between the unwrapped phase terms are required to remain within $(-\pi,\pi)$. The intensity (energy) and phase in TF space are calculated by averaging within the time frame of length 1ms with 50% overlapping. It improves the source localization performance.

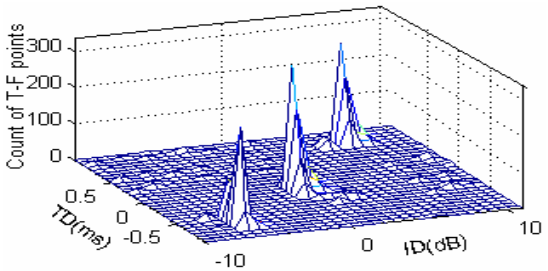

Fig. 3. TD-ID Space Localization of three sources

The values of TD and ID computed by Eq. (4) are quantized into discrete levels (50 levels). Then the histogram ψ(TD, ID) is constructed by mapping each TF point into quantized TD-ID space. In ψ(TD, ID), it is observed that each source is properly localized at specific region within TD-ID space. Fig. 3 shows the localization of three sources (two speech signals and flute sound located at 60°, 90° and 120° azimuths respectively). The three peaks (with some degree of spreading) correspond to distinct active sources. The histogram is weighted by the energy function in the TF space of the mixture. Some further processing is necessary to smooth the histogram such that its peaks are in one-to-one correspondence with TD and ID parameters of each source.

4. SOURCE SEPARATION

The individual sources placed at different azimuth locations are in one-to-one relation with the unique regions in the histogram ψ(TD, ID). Such a mapping allows to construct the TF mask corresponding to each region and it is used to mask H_L or H_R to produce the TF representation of the component source. If δ_n and κ_n are the set of TD and ID respectively representing the peak region of the n^{th} source in $\psi(TD, ID)$, its TF mask can be computed as:

$$M^{(n)}(f,t) = \begin{cases} 1: [(TD(f,t) \in \delta_n) and (ID(f,t) \in \kappa_n)] \\ 0: otherwise \end{cases}; \forall f,t \quad (5)$$

The binary mask nullifies TF points of interfering sources. Then the Hilbert spectrum of the n^{th} source can be computed as: $H^{(n)}(f,t) = M^{(n)}(f,t)H_L(f,t)$. During the Hilbert transform the real part of the signal remains unchanged. The time domain signal of n^{th} source is reconstructed by filtering out the imaginary part from the HS and summing over frequency bins as [8]:

$$s^{(n)}(t) = \sum_f H^{(n)}(f,t) \cdot \cos[\phi(f,t)] \quad (6)$$

where $\phi(f,t)$ is the phase matrix of H_L (or H_R). The phase matrix is saved during the construction of Hilbert spectrum to be used in re-synthesis.

5. DISJOINT ORTHOGONALITY IN TF SPACE

The simple definition of disjoint orthogonality of audio sources says that not more than one source is active at the same time and with same frequency. If $Y_1(f,t)$ and $Y_2(f,t)$ are the TF representation of the signals $y_1(t)$ and $y_2(t)$, the disjoint orthogonality assumption can be stated as: $Y_1(f,t)Y_2(f,t) = 0; \forall f,t$. In order to better measure a signal at a particular point (f,t), it is natural to desire that Δ_t and Δ_f be as narrow as possible. In STFT based TF representation Δ_t and Δ_f has to satisfy an uncertainty inequality $\Delta_t\Delta_f \geq 0.5$ which is the trade-off of the selection of time-frequency resolution. The Hilbert spectrum has better time-frequency resolution and improved disjoint orthogonality (DO) of audio signals in TF space.

The signal to interference ratio (SIR) is used as the basis to measure the DO. The SIR for the n^{th} source signal is,

$$SIR_n = \sum_f \sum_t \frac{X_n(f,t)}{Y_n(f,t)}; Y_n(f,t) \neq 0 \quad (7)$$

$$Y_n(f,t) = \sum_{\substack{i=1 \\ i \neq n}}^{N} X_i(f,t)$$

where N is the number of audio signal considered to be disjoint orthogonal, $X_n(f,t)$ is the TF representation (using STFT or HS) of the n^{th} signal. The dimension in TF space using STFT and HS may be different and hence the DO is defined as the percentage over the whole TF space. It is achieved by dividing the SIR_n with the total number of TF points used to calculate SIR_n. Finally the average disjoint orthogonality (ADO) is the average of all SIRs of individual signal as: $ADO = \frac{1}{N}\sum_{n=1}^{N} SIR_n$. The same process is applied to measure ADO (between 0 to 1) for STFT and HS based TF representation of the audio signals. Some experimental results are presented to compare STFT and Hilbert spectrum as the TF representation tools of audio signals in terms of ADO.

6. EXPERIMENTAL RESULTS

The separation efficiency of the proposed algorithm is evaluated by separating the signals from stereo mixtures of three audio sources: speech of two male speakers (sm1 and sm2) and speech of a female (sf1). The recording is performed in an anechoic room. The spacing between two microphones is 10cm placed at 1.5m distance from each source. The sources are placed at different azimuth locations (0^o to180^o). The sampling rate of all the recording is set to 16kHz with 16-bit amplitude resolution.

Three binaural mixtures (m1, m2 and m3) are produced by

arranging the sources at different azimuth locations as: m1{sm1(40°), sm2(80°), sf1(110°)}, m2{sm1(80°), sm2(130°), sf1(120°)}, m3{sm1(140°), sm2(50°), sf1(120°)}. The average value of short time energy ratio between original and separated signal is proposed as the criterion to measure the separation efficiency. It is termed as OSSR (original to separated signal ratio) and defined as:

$$OSSR = \left| \frac{1}{T} \sum_{t=1}^{T} \log 10 \left(\frac{\sum_{i=1}^{w} s^2_{original}(t+i)}{\sum_{i=1}^{w} s^2_{separated}(t+i)} \right) \right| \qquad (8)$$

where $s_{original}$ and $s_{separated}$ are the original and separated signal respectively, w is frame length (10 ms) and T is the data length. If the two signals are same, OSSR=0 and any other value is a measure of their dissimilarity. Smaller value of OSSR indicates better separation. Table 1 shows the average OSSR of each signal for every mixture. It is observed that the separation efficiency is degraded when the apart angle between the sources becomes smaller. Also the separation efficiency is compared for three types of TF representations: HS using basic EMD (HS_b), HS using modified EMD (HS_m) and STFT. It is noticed that the HS based TF representation improves the separation performance whereas, in most of the cases HS_m performs better than HS_b.

Table 1: The experimental results of proposed algorithm

Mixtures	TF	OSSR of sm1	OSSR of sm2	OSSR of sf1
m1	HS_b	0.0512	0.0573	0.0664
	HS_m	0.0498	0.0567	0.0621
	STFT	0.0761	0.0781	0.0812
m2	HS_b	0.0611	0.0972	0.0893
	HS_m	0.0584	0.0873	0.0891
	STFT	0.0874	0.1032	0.1095
m3	HS_b	0.0815	0.0526	0.0781
	HS_m	0.0809	0.0521	0.0779
	STFT	0.1107	0.0.886	0.0986

The separation efficiency depends only on the apart angle between the sources locations but not on the signal contents. The separation accuracy is better for larger apart angle between the sources.

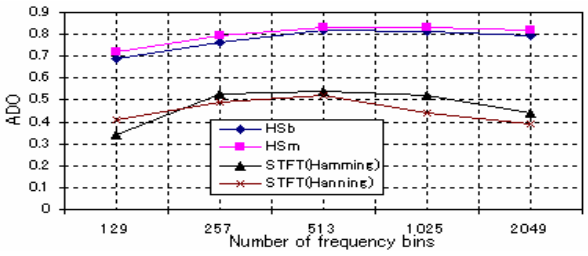

Fig. 4. ADO of HS and STFT as a function of frequency bins

Each one of the three audio signals is converted to TF space using HS (both HS_b and HS_m) and STFT separately to produce the experimental results of DO. Fig. 4 and 5 show ADO of HS and STFT (using Hamming and Hanning window with 60% overlapping) as a function of the number of frequency bins and window overlapping respectively. In both cases HS offers better ADO than STFT. The HS with the modification of the basic EMD i.e. HS_m improves the disjoint orthogonality of the audio sources in TF domain and hence the separation efficiency.

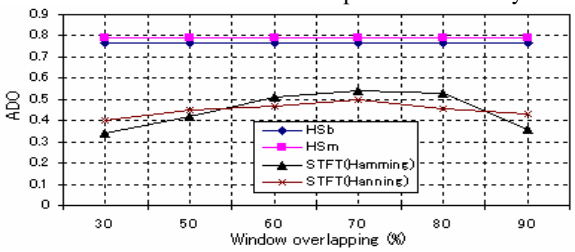

Fig. 5. ADO of HS and STFT as a function of window overlapping

7. CONCLUSIONS

A localization based method of audio source separation from stereo mixtures is proposed in this paper. The sources are localized in TD-ID space and separation is obtained by binary masking in TF domain. It is assumed that the sources are disjoint orthogonal in TF domain. The use of HS as the TF representation improves the disjoint orthogonality of the sources as well as the separation efficiency compared with STFT based method. The specialty of HS is that the time resolution can be as precise as the sampling period and the frequency resolution depends on the choice up to Nyquist frequency. Hence it serves as the potential TF representation for the consideration of disjoint orthogonality of audio sources. Also the modification of basic EMD method is proposed here to enhance the separation performance. The robust analysis of disjoint orthogonality of various audio sources and the separation of moving sources are the main concern as the future extension of this work.

8. REFERENCES

[1] N. Roman, D. Wang and G. J. Brown, "Speech segregation based on sound localization", *Acos. Soc. of America*, 114(4): 2236-2252, 2003

[2] J. Nix, and V. Hohmann, "Enhancing sound sources by use of binaural spatial cues", *Workshop on Consistent & Reliable Acoustic Cues* (CRAC'01), 2001.

[3] M. Baeck and U. Zolzer, "Real-Time Implementation of Source Separation Algorithm", DAFx-03, London, UK, 2003.

[4] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech", *ICASSP02*, Florida, USA, 2002.

[5] N. E. Huang et al, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis", *Proc. Roy. Soc. London* A, Vol. 454: 903-995, 1998.

[6] P. Flandrin, G. Rilling and P. Goncalves, "Emperical Mode Decomposition as a filter bank", *IEEE Sig. Proc. Letter*, 2003.

[7] B. Z. Wu, and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method", *Proc. R. Soc. London*, A (460), pp: 1597-1611, 2004.

[8] M. K. I. Molla, K. Hirose and N. Minematsu, "Audio Source Separation from the Mixture using Empirical Mode Decomposition with Independent Subspace Analysis", *ICSLP'04*, Oct, 2004.