Separation of Mixed Audio Signals by Source Localization and Binary Masking with Hilbert Spectrum

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

[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
{molla, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract. The Hilbert transformation together with empirical mode decomposition (EMD) produces Hilbert spectrum (HS) which is a fine-resolution time-frequency (TF) representation of any nonlinear and non-stationary signal. A method of audio signal separation from stereo mixtures based on the spatial location of the sources is presented in this paper. The TF representation of the audio signal is obtained by HS. 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

When the audio recording is performed using two microphones in an adverse acoustical environment, time difference (TD) and intensity difference (ID) are introduced between the mixed signals. Those are termed as interaural differences and used to localize the audio sources in spatial domain. Such source localization method is used in [1], [2], [3] to separate the individual audio source from two mixtures. In [2], the mixtures are produced by convoluting monaural signal with measured head related transfer functions (HRTFs) [4], whereas two microphones are used for recording in [1], [3]. The use HRTF introduces TD and ID in the mixture signals. The TD is the main localization cue at low frequencies and ID dominates the high frequency range. The partition between these two ranges of frequency depends on the spacing between the microphones [2]. To cover the entire frequency range, the TD and ID are jointly used in localization. The TF masks are used to segregate the individual sources in TF domain. The principal assumption of the masking based separation in TF domain is that, the audio sources are disjoint orthogonal i.e. not more than one source is active at any TF point [1]. The short-time Fourier tranform (STFT) is an usual approch to represent the time domain signal in TF domain [1], [2], [3]. 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 degrades the disjoint orthogonality of the audio sources and hence the separation efficiency by using masking method in TF domain.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 641 – 648, 2006.

In this paper a novel technique to separate audio sources stereo mixtures based on spatial localization is described. The separation efficiency can be improved by maximizing the resolution and minimizing the cross-spectral energy terms in TF space. The proposed separation method employs HS as the TF representation. HS does not include noticeable amount of cross-spectral energy terms. 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. Based on the TD and ID between two mixtures, the TF spaces (HSs of two mixtures) are clustered in TD-ID space to localize the audio sources. The TF space of each source is segregated by binary masking method [3], and the time domain signals are recostructed by applying the inverse transformations. The HS has better TF resolutions as well as less cross-spectral energy than STFT and hence more suitable for source disjoint orthogonality consideration.

Regarding the arrangement of this paper, the EMD and HS are illustrated in section 2, the source localization and separation methods are described in sections 3 and 4 respectively. The concept of disjoint orthogonality is presented in section 5. The experimental results are shown in section 6 and finally some concluding remarks are included in section 7.

2 The Modification of EMD and Hilbert Spectrum

The EMD represents the mixture signal as a collection of oscillatory basis components $C_m(t)$ termed as intrinsic mode functions (IMFs) containing some basic properties [5, 6]. The decomposition process can also be considered as dyadic filter-bank as proved by analysis of white noise [6], [7]. 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 residue $r_0(t)=s(t)$ and index of IMF $m=1$
b) (i) set $g_0(t)=r_{m-1}(t)$ and $i=1$
   (ii) Find the extrema (minima and maxima) of $g_{i-1}(t)$
   (iii) Compute upper and lower envelopes $h_{i-1}(t)$ and $l_{i-1}(t)$
   (iv) Find mean envelope $\mu_{i-1}(t)=[h_{i-1}(t)+l_{i-1}(t)]/2$
   (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 satisfying the above mentioned two basic conditions. If so, the m^{th} IMF $C_m(t)=g_i(t)$and update residue $r_m(t)=r_{m-1}(t)-C_m(t)$
c) Repeat step (b) with the index of IMF $m=m+1$

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

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

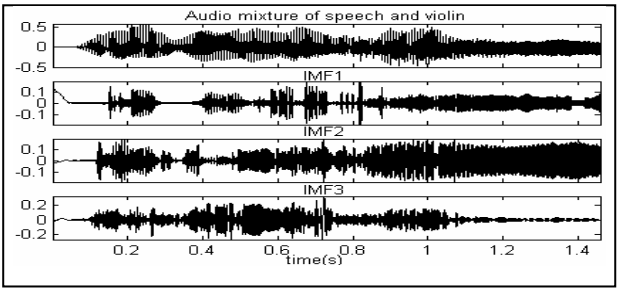

Fig. 1. The *EMD* of an audio mixture (speech and flute sound) showing first three IMFs out of 14

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. A band-limited (80Hz-4kHz) audio mixture signal and the decomposed IMF components are shown in Figure 1.

The IMFs computed by the basic EMD include energy at frequencies that cannot be associated with the original data. This phenomenon obviously includes unwanted signal energy in HS and hence degrades the separation performance. 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 the decomposition. The analyzing signal $s(t)$ is first 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 23 IMFs whereas, original one produces 14 IMFs from the same signal of Fig. 1. All the subsequent operations (computing instantaneous frequency, constructing Hilbert spectrum) are performed on the modified EMD.

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 angle at the instant of the "analytic" version of the signal. Every IMF is a real valued signal. The discrete Hilbert transform (HT) denoted by $\hbar_d[.]$ is used to compute the analytic signal for an IMF. 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:

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

where $a_m(t)$ and $\theta_m(t)$ are instantaneous amplitude and phase respectively of the m^{th} IMF. The IF of m^{th} IMF is then computed by the derivative of the phase $\theta_m(t)$ as: $f_m(t) = \frac{d\tilde{\theta}(t)}{dt}$, where $\tilde{\theta}_m(t)$ represents the unwrapped version of $\theta_m(t)$. The median smoothing filter is used to tackle the discontinuities of IF computed by discrete time derivative of the phase vector.

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 $H(\omega,t)$ or simply Hilbert spectrum (HS). This process first normalizes the IF vectors of all IMFs between 0 to 0.5. Each IF vector is multiplied by the scaling factor η=0.5/(IF_{max}-IF_{min}), where IF_{max}=Max(f_1, f_2,…,f_m,….,f_M) and IF_{min}=Min(f_1, f_2,…,f_m,….,f_M). The bin spacing of the HS is 0.5/B, where B is the number of desired frequency bins selected arbitrarily. Each element H(ω,t) is defined as the weighted sum of the instantaneous amplitudes of all the IMFs at ω^{th} frequency bin,

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

where the weight factor $w_m^{(\omega)}(t)$ takes 1 if $\eta \times f_m(t)$ falls within ω^{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. 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 sapling rate 16kHz).

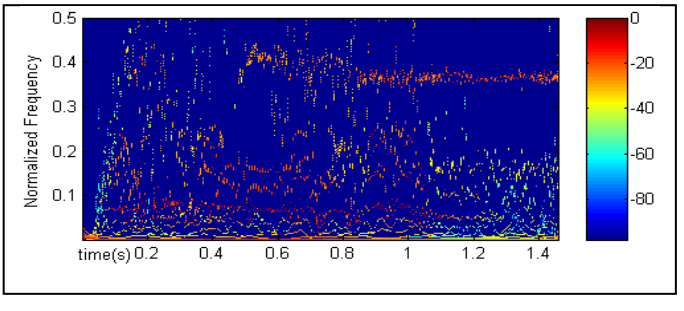

Fig. 2. Hilbert spectrum with 256 frequency bins. The amplitude is in dB.

3 Source Localization

The audio sources are localized in TD-ID space. There is a one-to-one mapping between the azimuth location and a region 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 $H_L(\omega,t)$ and $H_R(\omega,t)$ are the Hilbert spectrum of binaural mixtures $x_l(t)$ and $x_r(t)$ respectively, the TD and ID can easily be computed as [1], [3]:

$$TD(\omega,t) = \frac{1}{\omega}[\tilde{\phi}_L(\omega,t) - \tilde{\phi}_R(\omega,t)]$$
$$ID(\omega,t) = 20\log\left(\frac{|H_R(\omega,t)|}{|H_L(\omega,t)|}\right) \quad (4)$$

where $\tilde{\phi}_L(\omega,t)$ and $\tilde{\phi}_R(\omega,t)$ are the unwrapped phases corresponding to H_L and H_R respectively. The difference between the phase terms remains within $(-\pi,\pi)$. The intensity (energy) and phase information are smoothed in TF space by average filtering with the time frame of length 1ms. It improves the stability of the instantaneous energy and phase response computed by the discrete derivative of the analytic signals.

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. Fig. 3 shows the TD-ID space localization of three sources placed at 50°, 90° and 110° azimuths. 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.

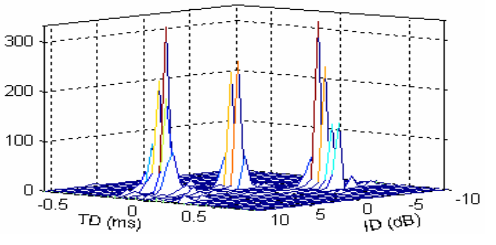

Fig. 3. TD-ID space localization of three sources

4 Source Separation

The individual source placed at different azimuth locations has 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 yield the TF representation of the original source. If δ_n and κ_n are the set of TD and ID respectively representing the rectangle of the peak region of the n^{th} source in ψ(TD, ID), its TF mask can be computed as:

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

The binary mask nullifies TF points of interfering sources. The HS of the n^{th} source can be computed as: $H^{(n)}(\omega,t) = M^{(n)}(\omega,t)H_L(\omega,t)$ or $H^{(n)}(\omega,t) = M^{(n)}(\omega,t)H_R(\omega,t)$.

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: $s^{(n)}(t) = \sum_{\omega} H^{(n)}(\omega, t) \cdot \cos[\phi(\omega, t)]$. where $\phi(\omega,t)$ is the phase matrix of H_L (or H_R). The phase matrix is saved during the construction of HS to be used in re-synthesis.

5 Disjoint Orthogonality in TF Space

If $Y_1(\omega,t)$ and $Y_2(\omega,t)$ are the TF representation of the signals $y_1(t)$ and $y_2(t)$ respectively, the disjoint orthogonality assumption can be stated as: $Y_1(\omega,t)Y_2(\omega,t) = 0; \forall \omega, t$. In order to better measure of a signal at a particular time and frequency (ω, t), it is natural to desire that Δ_t and Δ_ω be as narrow as possible. In STFT based TF representation Δ_t and Δ_ω has to satisfy an uncertainty inequality $\Delta_t \Delta_\omega \geq 0.5$ which is the trade-off of the selection of TF resolution. The Hilbert spectrum has better time-frequency resolution and improved disjoint orthogonality (DO) of audio sources in TF space. The signal to interference ratio (*SIR*) is used as basis to measure the DO. The *SIR* for the n^{th} source signal is,

$$SIR_n = \sum_{\omega} \sum_{t} \frac{X_n(\omega,t)}{Y_n(\omega,t)}; Y_n(\omega,t) \neq 0 \qquad (6)$$

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

where N is the number of audio signal considered to be disjoint orthogonal, $X_n(\omega,t)$ is the TF representation (using STFT or HS) of the n^{th} signal. The dimension of *TF* representation using STFT and HS may be different, hence the DO is defined in percentage computed over the entire TF space. Finally the average disjoint orthogonality (*ADO*) is the average of all *SIR*s of individual signal as: $ADO = \frac{1}{N} \sum_{n=1}^{N} SIR_n$ .The same process is applied to measure $ADO \in (0, 1)$ for STFT and HS based TF representation of the audio signals. Some experimental results are presented to compare STFT and HS as the TF representation tools of audio signals in terms of disjoint orthogonality.

6 Experimental Results

The separation efficiency of the proposed algorithm is evaluated by separating the signals from two mixtures of three audio sources: speech of two male persons (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 was 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(70^o), sm2(100^o), sf1(140^o)}, m2{sm1(50^o), sm2(80^o), sf1(90^o)}, m3{sm1(130^o), sm2(90^o), sf1(150^o)}. 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:

$$OSSR = \left| \frac{1}{T} \sum_{t=1}^{T} \log 10 \left(\frac{\sum_{i=1}^{w} s_{original}^2 (t+i)}{\sum_{i=1}^{w} s_{separated}^2 (t+i)} \right) \right| \tag{7}$$

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 sources are placed closely. The separation accuracy is better for larger apart angle between the sources. The separation efficiency is compared for two types of TF representations: HS and STFT. Also the efficiency is compared for two types of stereo mixtures: using HRTF and recorded by two microphones (Mic2). It is noticed that the HS based TF representation improves the separation performance than STFT. Although, HRTF based mixing system has better separation efficiency, it is less applicable in real world applications.

The individual audio signal is projected to TF space using HS and STFT separately to produce some experimental results of DO. Fig. 4(a) shows the comparison between HS and STFT (using Hamming and Hanning window with 60% overlapping) in terms of ADO as a function of the number of frequency bins, whereas Fig. 4(b) presents the comparison as a function of window overlapping. The ADO of the audio signals of HS is better than that of STFT based TF representation. It is obvious to produce better source separation by the proposed method with HS as the TF representation.

Table 1. Experimental results of proposed separation algorithm

Mixture	TF	OSSR of sm1		OSSR of sm2		OSSR of sf1	
		Mic2	HRTF	Mic2	HRTF	Mic2	HRTF
m1	HS	0.0472	0.0401	0.0532	0.0491	0.0642	0.0586
	STFT	0.0781	0.0617	0.0743	0.0687	0.0831	0.0817
m2	HS	0.0519	0.0485	0.0882	0.0803	0.0817	0.0783
	STFT	0.0827	0.0758	0.1035	0.0975	0.1106	0.1047
m3	HS	0.0817	0.0737	0.0534	0.0478	0.0784	0.0711
	STFT	0.1073	0.0902	0.0903	0.0817	0.1012	0.0983

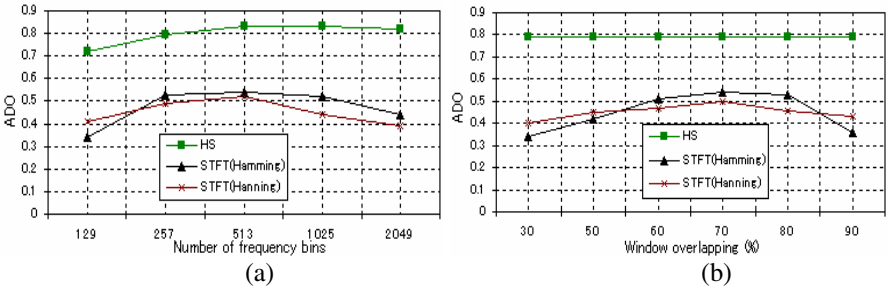

Fig. 4. ADO of HS and STFT as a function of (a)frequency bins, (b) window overlapping

6 Conclusions

We have presented a method of separating mixed audio signals by localizing the sources in TD-ID space. It is assumed that the sources are disjoint orthogonal and the separation is obtained by estimating the binary masks for individual source signal in TF space. The use of HS as the TF representation improves the separation efficiency for both of the mixtures using HRTF as well as microphone pair. 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. The robust localization and separation of moving sources are the main concern as the future works.

References

1. Yilmaz, O., Rickard, S.:Blind Separation of Speech Mixtures via Time-Frequency Masking, IEEE Transactions on Signal Processing, Vol. 52, No. 7, pages 1830-1847, July 2004.
2. Roman, N., Wang, D., Brown, G. J.: Speech segregation based on sound localization. Acost. Soc. of America, 114(4): 2236-2252, 2003.
3. Baeck, M., Zolzer, U.: Real-Time Implementation of Source Separation Algorithm. DAFx-03, London, UK, 2003.
4. http://sound.media.mit.edu/KEMAR.html
5. Huang, N.E, 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. Flandrin, P., Rilling, G., Goncalves, P.: Emperical Mode Decomposition as a filter bank. IEEE Sig. Proc. Letter, 2003.
7. Wu, B. Z., Huang, N. E.: A study of the characteristics of white noise using the empirical mode decomposition method. Proc. R. Soc. Lond. A (460), pp: 1597-1611, 2004.