Audio Source Separation by Source Localization with Hilbert Spectrum

Md. Khademul Islam Molla[1]
[1]Graduate School of Frontier Sciences
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Email: molla@gavo.t.u-tokyo.ac.jp

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

***Abstract*—This paper presents a technique to separate the audio signals from their binaural mixtures based on localizing the sources in the space of interaural differences. Two interaural differences *ITD* (interaural time difference) and *ILD* (interaural level difference) are used as the principal cues to localize and segregate the sources. Hilbert spectrum is employed to decompose the mixture signals into time-frequency (*T-F*) space. The sources of the mixtures are considered as disjoint orthogonal in the *T-F* space. Hilbert spectrum has a better *T-F* resolution than Fourier based method and hence it produces a better disjoint orthogonality of the sources. The separation efficiency as presented in experimental results using our proposed algorithm is noticeable in this research area.**

I. INTRODUCTION

The separation of mixed audio signals has many potential applications including robust speech recognition, music trascription, speaker separation from recorded meeting and video conferencing. The present research trend is to reduce the number of mixture signals. The researches on single mixture source separation [1,2] produce some application specific results. They have the limitations in robustness and it is very difficult to separate more than two sources using single mixture. The cocktail-party effect is a crucial situation for humanoid robotics to segregate and recognize a particular sound. In such situation human has the ability to keep the attention to a single audio source in an adverse acoustical condition. The location of the source helps human auditory to be separated from the interfering sounds. In multi-source listening situation, human exploits spatial characteristics of source signals by the mechanism of binaural hearing.

The human localization ability is simulated for speech segregation in [3,4]. They have only focused on the separation of one target source (speech). The location based separation is also applied in [5, 10] using *FFT* based time-frequency (*T-F*) representation cosidering the sources as disjoint orthogonal. Whereas *FFT* based method (*STFT* short-time Fourier transform) is only acceptable for disjoint orthogonality consideration of speech signal but not well suited for all types audio signals.

This paper presents a technique to detect, discriminate and separate individual audio source from their binaural mixtures using some spatial localization cues. In human audition, *ITD* and *ILD* are introduced between two ears' binaural signals. The *ITD* is the main localization cue at low frequency (<1.5kHz) and *ILD* dominates the high frequency range [3]. Measured head related transfer functions (*HRTF*s) introduce natural combination of *ITD* and *ILD* in binaural mixture. The empirical mode decompostion (*EMD*) together with Hilbert transform [2, 6] is used for T-F representation of the binaural mixture signals. The *T-F* is clustered in *ITD/ILD* space to localize the source signals. The sources properly localized in *ITD/ILD* space are segregated and recostructed using some reverse transformations. *EMD* based *T-F* represenation has better *T-F* resolution and hence more suitable for source disjoint orthogonality consideration.

The proposed method can blindly separate the sources (at fixed position) from the mixture of more than two sources. The separation model is described in scetion 2, section 3 presents the derivation of source disjoint orthogonality, some experimental results are presented in section 4 and finally section 5 contains some discussion and conclutions

II. PROPOSED MODEL DESCRIPTION

The schematic diagram of proposed binaural separation model is shown in Figure 1. It consists of two basic stages: (1) source localization in *ITD/ILD* space and (2) source separation and reconstruction. Each stage is described in detail in the following sub-sections. The inputs to the system are the monaural signals presented at different but fixed locations. The binaural signals are obtained by convoluting the monaural signals with measured *HRTFs* (from a KEMAR dummy head of MIT media laboratory under anechoic condition) corresponding to the direction of incidence [6]. The mixtures can be defined as the convolutive sum with *HRTF*s:

0-7803-8834-8/05/$20.00 ©2005 IEEE.

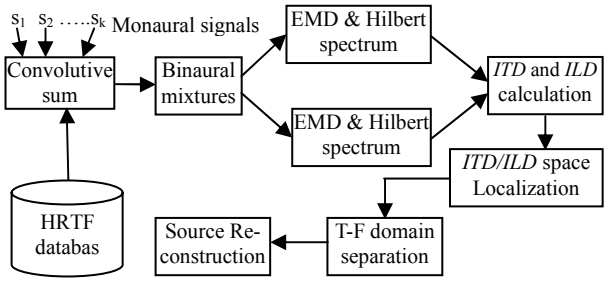

Figure 1. Schematic diagram of the proposed method

$$x_j(t) = \sum_k h_{jk} * s_k \qquad (1)$$

where j =l, r (denoting left and right respectively), s_k is the k^{th} source and h_{jk} denotes the measured *HRTF* at the k^{th} source position. We have proposed a localization based source separation method considering source disjoint orthogonality with Hilbert spectrum which is a fine resolution T-F representation.

A. Hilbert Spectrum

The principle of the *EMD* method is to decompose a time domain signal into a sum of oscillatory functions called intrinsic mode functions (*IMFs*) [2]. 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 adapts a global requirement to a local one, and is necessary to ensure that the instantaneous frequency will not have redundant fluctuations as induced by asymmetric waveforms [6]. 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. There exist many algorithmic approaches of *EMD* [7]. At the end of *EMD*, any signal $x_r(t)$ can be represented as:

$$x_r(t) = \sum_{i=1}^{n} imf_i(t) + r_n(t) \qquad (2)$$

where n is the number of *IMF*s and $r_n(t)$ is the final residue signal. Every *IMF* is a real valued signal. Analytic signal method is used to calculate the instantaneous frequency (*IF*) of the *IMF*s. The analytic signal corresponding to i^{th} *IMF* is defined as: $imf_i(t) + jH[imf_i(t)] = a_i(t)e^{j\theta_i(t)}$. where *H[]* is the Hilbert transform operator, $a_i(t)$ and $\theta_i(t)$ are instantaneous amplitude and phase respectively. The instantaneous frequency $\omega_i(t)$ can easily be computed as the change of $\theta_i(t)$ with respect to time t as: $\omega_i(t)=d\theta_i(t)/dt$.

Hilbert spectrum $H(\omega,t)$ describes the joint distribution of signal amplitude as a function of frequency and time. To build $H(\omega,t)$, the *IF* 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 *IF*, we represent the time-frequency plane as the triplet $\{t, \omega_i(t), a_i(t)\}$ where $a_i(t)$ is the amplitude of the analytic signal associated to $imf_i(t)$. $H(\omega,t)$ can produce the instantaneous (even at every sampling time) spectra of nonlinear and non-stationary signals.

B. Source Localization in ITD/ILD Space

Hilbert spectrum (*HS*) is a fine resolution time-frequency representation. 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 *ITD* and *ILD* can easily be computed with a simple division [5, 8] as:

$$\{ITD(\omega,t_f), ILD(\omega,t_f)\} = \left[\frac{1}{\omega}\angle\frac{H_L(\omega,t_f)}{H_R(\omega,t_f)}, 20\log\left|\frac{H_R(\omega,t_f)}{H_L(\omega,t_f)}\right|\right] \qquad (3)$$

The average relative energy and phase are calculated within the time frame t_f (1ms length with 0.5ms overlapping) yielding the *ITD* and *ILD*.

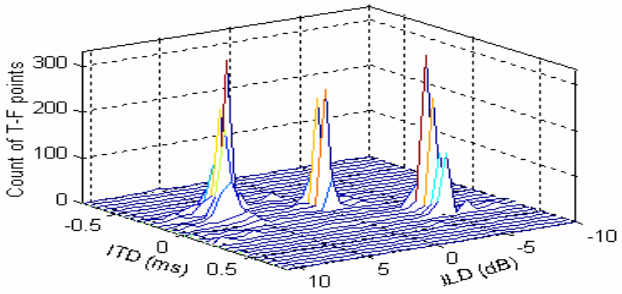

Figure 2. *ITD/ILD* Space Localization of three sources

The values of computed *ITD* and *ILD* are quantized into the discrete levels (50 levels) and the histogram *h(ITD, ILD)* is constructed by mapping T-F points into quantized *ITD/ILD* space. In *h(ITD, ILD)*, we observe that each source is properly localized at specific region in *ITD/ILD* space. Figure 2 shows the *ITD/ILD* space localization of three sources (two speech signals and flute sound located at –40°, 30° and 0° azimuths respectively with all at 0° elevation). It is clearly noticeable that the strong peaks correspond to distinct active sources.

C. Source Separation and Re-Synthesis

Some further processing is necessary to smooth the histogram such that every source produces only one peak with some surrounding points. The individual source signal is separated by deriving corresponding masking function for *T-F* domain. A binary mask (by nullifying *T-F* points of interfering sources) is computed to collect the *T-F* points (from H_L or H_R) corresponding to each peak region (for every source) in joint *ITD-ILD* space. If $M_k(\omega,t)$ be the binary mask of the k^{th} source, the *HS* of k^{th} source $H_k(\omega,t)= M_k(\omega,t)H_L(\omega,t)$ or $H_k(\omega,t)= M_k(\omega,t)H_R(\omega,t)$.

During the Hilbert transform the real part of the signal remains unchanged. The time domain signal of k^{th} source is reconstructed by filtering out the imaginary part from the *HS* and summing over frequency bins as [2]:

$$s_k(t) = \sum_{\omega} H_k(\omega,t) \cdot \cos[\phi(\omega,t)] \qquad (4)$$

where $\phi(\omega,t)$ 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.

III. DISJOINT ORTHOGONALITY OF THE SOURCES

In order to better measure 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 Fourier based *T-F* representation Δ_t and Δ_ω has to satisfy an uncertainty inequality $\Delta_t\Delta_\omega \geq \frac{1}{2}$ which is the trade-off of the selection of time-frequency resolution in *STFT*. The simple definition of disjoint orthogonality of audio sources says that not more than one source signal is active at the same time and with same frequency. This is a very hard definition to comply with the audio signals. Some assumption relaxes this definition as in [5, 10]. They have called two functions $f_1(t)$ and $f_2(t)$ as *w*-disjoint orthogonal if for a given window function $w(t)$, the supports of the windowed Fourier transforms with $w(t)$ of $f_1(t)$ and $f_2(t)$ are disjoint. If $F_1(\omega,t)$ and $F_2(\omega,t)$ are the windowed Fourier transform of the signals $f_1(t)$ and $f_2(t)$, the W-disjoint Orthogonality assumption can be stated as: $F_1(\omega,t)F_2(\omega,t) = 0; \forall \omega, t$.

We are not considering the window function to measure the disjoint orthogonality (*DO*) here as no window function is required in computing the Hilbert spectrum. Hence we are simply calling it disjoint orthogonality by dropping the *w* term. The signal to interference ratio (*SIR*) is used as basis to measure the *DO*. The SIR for the j^{th} source signal is,

$$SIR_j = \sum_{\omega}\sum_{t} \frac{X_j(\omega,t)}{Y_j(\omega,t)}; Y_j(\omega,t) \neq 0$$
$$Y_j(\omega,t) = \sum_{\substack{i=1 \\ i\neq j}}^{N} X_i(\omega,t) \qquad (5)$$

where N is the number of audio signal considered to be disjoint orthogonal, $X_j(\omega,t)$ is the *T-F* representation (using *STFT* or Hilbert spectrum) of the j^{th} signal. The dimension of *T-F* representation using *STFT* and *HS* may be different, hence the *DO* is defined as the percentage over the whole *T-F* region. It is achieved by dividing the SIR_j with the total number of *T-F* points used to calculate SIR_j. Finally the average disjoint orthogonality (*ADO*) is the average of all *SIR*s of individual signal as: $ADO = \frac{1}{N}\sum_{j=1}^{N} SIR_j$.

The same process is applied to measure *ADO* (between 0 to 1) for *STFT* and *HS* based *T-F* representation of the experimental audio signals. We have presented some experimental results to compare *STFT* and Hilbert spectrum as the *T-F* representation tools of audio signals in terms of source disjoint orthogonality.

IV. EXPERIMENTAL RESULTS

We have used the binaural mixtures of three audio streams of two male speech (sp1 and sp2) and flute sound (ft) to test the efficiency of our algorithm. Each monaural recording is upsampled to 44.1 kHz to match with the *HRTF*. The binaural mixtures are obtained using equation (1) and then down-sampled to 16kHz. Placing the sound sources at various locations (azimuth, elevation) produces different binaural mixtures for the test purpose. Such three mixtures are produced as: m1{sp1(-40°, 0°), sp2(30°, 0°), ft(0°, 0°)}, m2{sp1(20°, 10°), sp2(0°, 10°), ft(-10°,10°)}, m3{sp1(40°, 20°), sp2(30°, 20°), ft(-20°, 20°)} and the separation result is presented bellow. The origin (0° azimuth, 0° elevation) is considered at the front of the listener.

The average value of short time energy ratio between original and separated signal is used here as the criterion to measure the separation efficiency. It is termed as *OSSR* (original to separated signal ratio) [2] and defined by

$$OSSR = \frac{1}{T}\sum_{t=1}^{T} \log 10 \left(\frac{\sum_{i=1}^{w} s^2_{original}(t+i)}{\sum_{i=1}^{w} s^2_{separated}(t+i)} \right) \qquad (6)$$

where $s_{original}$ and $s_{separated}$ are the original and separated signal respectively, w is frame length (10 ms) and T is the number of frames. In the case for zero energy in a particular window, no *OSSR* measurement is performed. If the two signals are similar, *OSSR=0* 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. It also presents comparative experimental results of separation using *HS* and *STFT* though *STFT* depends on many factors.

Table 1: The experimental results of audio source separation using *HS* (with 257 frequency bins) and *STFT* (30ms Hamming window, 20ms overlapping and 512 point *FFT*).

Mixtures	T-F	OSSR of sp1	OSSR of sp2	OSSR of ft
m1	*HS*	-0.0271	0.0213	0.0264
	STFT	0.0621	-0.0721	-0.0531
m2	*HS*	0.0211	-0.0851	-0.0872
	STFT	0.0824	0.1202	0.1182
m3	*HS*	0.0941	-0.0832	0.0225
	STFT	-0.1261	0.1092	-0.0821

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. It is suggested to keep apart angle among the locations of the sources not less than 10° (both azimuth and elevation).

Each one of the three audio signals is converted to *T-F* space using *HS* and *STFT* separately to produce some experimental results of *DO*. Figure 3 shows *ADO* of *HS* and *STFT* (using Hamming and Hanning window with 60% overlapping) as a function of the number of frequency bins. Figure 4 presents the comparison between *HS* and *STFT* as a function of window overlapping.

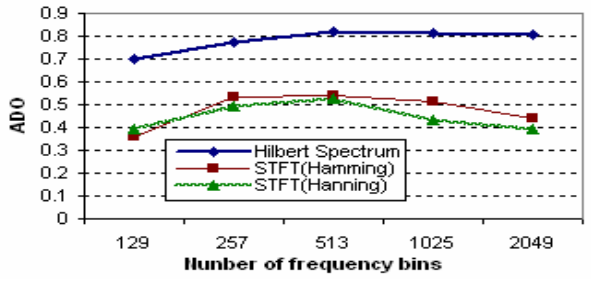

Figure 3. ADO of HS and STFT as a function of frequency bins

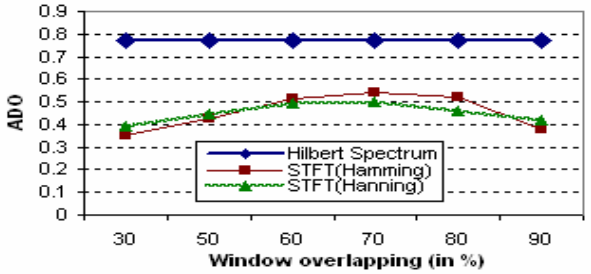

Figure 4. ADO of HS and STFT as a function of window overlapping

The variation of *ADO* of *HS* is very small (only for the number of frequency bins), whereas many factors affect the *ADO* of *STFT*. It is also a crucial decision to determine the suitable parameters for better source separation with *ADO* consideration. Hence the *ADO* of the audio signals of *HS* is better than that of *STFT* based *T-F* representation. It is obvious to produce better source separation efficiency by the proposed method with *HS* as the *T-F* representation.

V. Discussion and Conclusions

We have proposed an audio source separation technique by localization of the sources in the domain of binaural cues. *HRTF* is employed to introduce the binaural cues of human auditory system. It is considered that the sources are disjoint orthogonal [5, 10] in their *T-F* domain. The better the *DO* of the sources produces better separation on the basis of *DO* consideration. Hilbert spectrum is used for *T-F* representation of the binaural mixture signals. In [5] it is argued that the source disjointness depends on many factors including window size, window type, number of overlapping samples and also for the number of *FFT* points to produce the spectrogram. Hilbert spectrum is not affected by any of the mentioned factors as presented in experimental results. Another potential issue of the improvement of *DO* of *HS* is that some crossed terms are introduced in *STFT* with window overlapping and *HS* is free of such scenario. In [3] the authors have used the binaural cues to estimate the ideal binary mask to separate speech signal from interfering noise. The source with higher contribution at any *T-F* point is considered as the target signal (speech). They proposed a training based energy ratio function measuring the relative strength between the target source and the acoustic interference at each T-F point. The individual filter output of the filterbank (with 128 of gammatone filters) is divided into 20ms time frame with 10ms overlap that correspond the T-F unit. This consideration is hard to be used for separating more signals individually. In our system *ITD/ILD* space localization is used to separate more than two sources. Being independent of signal content the localization cues can be used to segregate sequence of voiced and unvoiced components originating from the same location. The separation performance does not depend on the spectral nature of the target sources.

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. Hence it serves as the potential *T-F* representation for the consideration of source disjoint orthogonality. The robust analysis of disjoint orthogonality of various audio sources and the separation of moving sources are the main concern as the future works of this research.

References

[1] Sam T. Roweis, "One Microphone Source Separation", Neural Information Processing Systems, pp. 793-799, 2000.

[2] Md. Khademul Islam Molla, Keikichi Hirose, Nobuaki Minematsu, "Audio Source Separation from the Mixture using Empirical Mode Decomposition with Independent Subspace Analysis", To appear in the Proc. of ICSLP2004.

[3] Nicoleta Roman, Deliang Wang, Guy J. Brown, " Speech segregation based on sound localization", Acos. Soc. of America, 114(4): 2236-2252, 2003

[4] Johannes Nix, Volker Hohmann, "Enhancing sound sources by use of binaural spatial cues", Workshop on Consistent & Reliable Acoustic Cues (CRAC'01), 2001.

[5] Matthias Baeck, Udo Zolzer, "Real-Time Implementation of Source Separation Algorithm", DAFx-03, London, UK, 2003.

[6] N. E. Huang etl., "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.

[7] P. Flandrin, G. Rilling, P. Goncalves, "Emperical Mode Decomposition as a filter bank", IEEE Sig. Proc. Letter, (in press), 2003.

[8] S. Srinivasan, N. Roman, D. Wang, "On binary and ratio time-frequency masks for robust speech recognition", To appear in the Proc. of ICSLP2004.

[9] http://sound.media.mit.edu/KEMAR.html

[10] Scott Rickard, Ozgur Yilmaz, "On the approximate W-disjoint orthogonality of speech", ICASSP2002, Orlando, Florida, USA, 2002