Localization Based Audio Source Separation by Sub-band Beamforming

Md. Khademul Islam Molla[†], Keikichi Hirose[‡] and Nobuaki Minematsu[†]
[†]Graduate School of Frontier Sciences, [‡]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— **In this paper, a localization based approach of audio signal separation from binary mixtures is carried out. The audio sources are localized in the spatial domain (azimuth plane) using the delay and amplitude variation cues between two microphones' signals. A coherence based technique is introduced here to localize the audio sources in adverse acoustical environment. The mixture signals are decomposed into a desired number of sub-bands with empirical mode decomposition (EMD) which is a data adaptive filtering scheme suitable for nonlinear and non-stationary signals. Data independent minimum variance beamforming is employed to separate the component sources in underdetermined condition (more sources than sensors). The experimental results of the proposed algorithm show noticeable separation efficiency. It is also found that the sub-band implementation improves the performance compared with and full-band approach.**

Index Terms— **Empirical mode decomposition, signal coherence, source localization, spatial beamforming.**

I. Introduction

THE separation of mixed audio signals has many potential applications including robust speech recognition, music transcription, speaker separation from recorded meeting and video conferencing, robotics. The present research trend is to reduce the number of mixture (microphones) signals. The separation of audio sources in underdetermined case remains problematic. The localization based approach is proposed in the paper to separate the audio sources from binary mixtures. Two cues, the time difference (TD) and the intensity difference (ID) between two microphones are employed to localize and separate concurrent audio signals from binary mixtures. The effects of TD and ID depend on the signal frequency as well as the spacing between the microphones. The source localization ability is dominated by TD and ID in lower and higher frequency ranges respectively. The TD is gradually substituted by ID with increasing of frequency.

The models of localization based audio source separation form binary mixtures have been proposed in [1, 2, 3, 4]. In all the algorithms the mixtures are produced by using measured head related transfer function (HRTF). The advantage of localization based separation is that the separation efficiency is independent of the content of the individual signals. The performance only depends on the spatial location of the sources. In [1] the azimuth of the source is considered as the direction of arrival (DOA) and the location dependent weighted filter is used in separation. In [2], only one source has been taken into account to be segregated from interfering sound. A supervised learning based ratio mask is used in separation and hence, the performance depends on priori knowledge about the sources. In [3], the authors only consider two sources in separation whereas the separation of source in underdetermined condition is a challenging task. Visual cue is employed in [4] as the primary support of source localization.

This paper presents a technique to detect, discriminate and separate individual audio sources from two mixtures using the binaural localization cues and adaptive beamforming. In a multi-source audio environment, the localization ability may degrade due to the interference, diffraction and resonance effects of the signals around the region close to the pair of microphones [5]. To reduce such unwanted effects in localization, the coherent frequency components of the mixed signals are used to compute the localization cues (TD and ID). The TD is represented here as the phase difference (PD). The multi-band approach of adaptive beamforming scheme is applied to segregate the localized source signals. The multi-band decomposition of the mixed signals is performed by using empirical mode decomposition (EMD) [6]. The gain factor of the beamformer is controlled by the location based parameters.

Regarding the arrangement of this paper, source localization method is illustrated in section two, the multi-band implementation with EMD is presented in section three and sub-band based beamforming approach is described in section four. The experimental results and discussion are presented in section five and finally section six includes some concluding remarks.

II. Spatial Localization Of The Sources

The proposed algorithm mainly consists of two steps: to localize the sources in terms of azimuth angle and to separate the localized sources by employing sub-band beamforming method. A microphone pair is used to capture the multi source audio signals. A priori map of PD and ID between the microphones are computed in an anechoic room based on different azimuth locations. The azimuth localization cue is defined by combining the PD and ID for individual azimuth locations. The localization is performed by comparing priori map with the azimuth cue calculated from the mixed signals.

All the interfering sources sometimes introduce an error in localization by producing a source location neither the real one [7]. In this method, the spectra of coherent frequency are used to compute the localization cues (IP and ID) to resolve such error. The coherence between two mixed signals $x_1(t)$ and $x_2(t)$ is defined as:

$$\zeta_{12}(\omega) = \frac{|P_{12}(\omega)|^2}{P_1(\omega)P_2(\omega)} \qquad (1)$$

where $P_{12}(\omega)$ is the cross power spectra of $x_1(t)$ and $x_2(t)$, $P_1(\omega)$ and $P_2(\omega)$ are the power spectra of $x_1(t)$ and $x_2(t)$ respectively. The signals are normalized prior to computing the coherence function. It is noted that $\zeta_{12}(\omega) \in (0,1)$. The frequency component with $\zeta_{12}(\omega) > 0.9$ is termed here as coherent frequencies and denoted by ω_c.

Consider $X_1(\omega)$ and $X_2(\omega)$ are the short time Fourier spectrum (512 point FFT, 30ms Hamming window with 20ms overlapping) of first and second microphones' mixtures $x_1(t)$ and $x_2(t)$ respectively. Then PD and ID can be calculated at the coherent frequency ω_c as:

$$\rho(\omega_c) = [\phi_1(\omega_c) - \phi_2(\omega_c)] \qquad (2)$$

$$\kappa(\omega_c) = 20\log\left(\frac{|X_1(\omega_c)|}{|X_2(\omega_c)|}\right) \qquad (3)$$

where $\phi_1(\omega_c)$, $\phi_2(\omega_c)$ are the unwrapped phase vectors of $X_1(\omega)$ and $X_2(\omega)$ respectively at coherent frequency ω_c.

The effects of PD and ID in source localization task are not linear with frequency. It depends also on the spacing between the microphones. In this experiment the spacing is 10cm. At low frequencies (<1.7 kHz), there is little ID information, but the PD is the dominant cue of localization [2]. At high frequencies (>1.7 kHz), there is ambiguity in PD, and the ID resolves such localization ambiguity. The proposed combined azimuth cue is defined as:

$$a_{\rho\kappa}(\omega_c) = [1-\beta(\omega_c)]\alpha_\rho(\omega_c)\rho(\omega_c) + \beta(\omega_c)\alpha_\kappa(\omega_c)\kappa(\omega_c) \qquad (4)$$

where $\alpha_\rho(\omega_c)$ and $\alpha_\kappa(\omega_c)$ are frequency dependent normalization factors of $\rho(\omega_c)$ and $\kappa(\omega_c)$ respectively. Since PD and ID are measured in different scales, it is required to normalize prior to combine them. $\beta(\omega_c)$ is the forgetting factor of azimuth cue. $\beta(\omega_c)$ =0 up to 1.5 kHz, 1 for frequency greater than 2.0 kHz and it is increased gradually between the frequency range 1.5 kHz to 2.0 kHz. A priori map $\chi(\omega_c,\varphi)$ of the proposed azimuth cue $a_{\rho\kappa}(\omega_c)$ is computed using the transfer functions between the microphones and the source placed at individual azimuth locations.

The DOA information of the source placed at azimuth φ is calculated as:

$$\delta_a(\varphi) = \sum_{\omega_c} \eta_e(\omega_c)\cdot\psi(\omega_c,\varphi) \qquad (5)$$

where $\eta_e(\omega_c)$ stands for weighting factor based on energy. The energy term of the mixture spectrogram is normalized. The function $\psi(\omega_c,\varphi)$ represents how close the derived azimuth cue $a_{\rho\kappa}(\omega_c)$ to the priori map $\chi(\omega_c,\varphi)$ and it is expressed as:

$$\psi(\omega_c,\varphi) = \exp(-|\chi(\omega_c,\varphi) - a_{\rho\kappa}(\omega_c)|) \qquad (6)$$

The resultant DOA function $\delta_a(\varphi)$ produces the peaks at some azimuth angle φ representing the source azimuth positions. The DOA represented by the function $\delta_a(\varphi)$ of the sources placed at azimuth angles 20°, 50° and 110° is shown in Fig 1. It shows a comparison of azimuth localization with and without coherent frequency. It is observed that the use of coherent frequencies performs better localization.

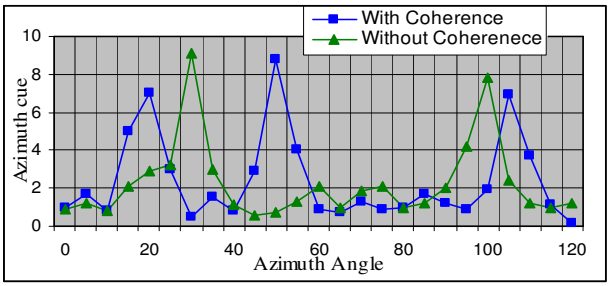

Fig 1: Three source localization at 20°, 50° and 110° azimuths

III. Multi-band Decomposition with EMD

The multi-band representation scheme of the mixture signals is implemented by using empirical mode decomposition (EMD). It is specifically designed to analyze the non-linear and non-stationary properties of a time domain signals [6]. The principle of the EMD technique is to decompose any signal into a sum of the oscillatory components called intrinsic mode functions (IMFs). 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. For any time series $s(t)$ the EMD algorithm can be expressed as follows:

a) Initialize the residual $r_0(t)=s(t)$ and index of IMF $j=1$
b) (i) set $h_0=r_{j-1}$ and $i=1$
(ii) Identify the extrema (minima and maxima) of $h_{i-1}(t)$
(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 $h_i(t)=h_{i-1}(t)-\mu_{i-1}(t)$ and $i=i+1$
(vi) Repeat steps (ii)-(v) until $h_i(t)$ being an IMF. If so, the j^{th} IMF $m_j(t)=h_i(t)$
c) Update residual $r_j(t)=r_{j-1}(t)-m_j(t)$
d) Repeat steps (b) with the index of IMF $j=j+1$

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

$$s(t) = \sum_{j=1}^{n} m_j(t) + r_n(t) \qquad (7)$$

where n is the number of IMF components and r_n is the final residue. Another way to explain how EMD works is that it filters the highest frequency oscillation that remains in the signal. Thus locally, each IMF contains lower frequency oscillation than the one extracted just before.

The IMF components are interpreted as the basis vectors representing the data. The EMD is also interpreted as dyadic filer-bank [8]. In this application, the IMFs are used in sub-band filtering. The multi-band decomposition is implemented in time domain by grouping the IMFs in order. Each group of IMFs corresponds a band passed signal. Conventionally, the filtering is carried out in frequency domain. Any frequency domain (e.g. Fourier) filtering method applied on nonlinear and non-stationary signal (e.g. speech) eliminates some of the harmonics, which will cause the deformation of the wave forms of the fundamental modes [9]. The signal of b^{th} band can be represented as the

summation of the selected IMF components. Then $s_b(t)$ can be defined as:

$$s_b(t) = \sum_{j=p}^{q} m_j(t) \qquad (8)$$

where the indices p and q depend on the number of desired sub-bands and fractional energy of the individual IMF. The advantage of this time-space filtering is that the resulting band passed signals preserve the full nonlinearity and non-stationary in physical space. An audio stream (mixture of speech and flute sound) and its three-band decomposition using EMD are shown in Fig 2. It should be noted that the the original signal can be reconstructed by summing up the sub-band signals with a negligible error (of the order 10^{-14}).

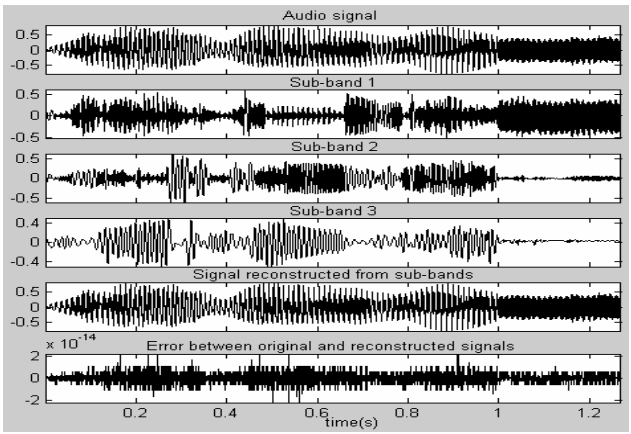

Fig 2: Three band decomposition of audio mixture

IV. Separation By Sub-band Beamforming

The beamformer performs as a spatial filtering to separate signals that may have overlapping frequency content but originated from different spatial locations [10]. A Sub-band implementation of linearly constrained minimum variance beamforming (LCMVB) is applied here with location based constrained. It passes the signal of desired location with specified gain factor while minimizing the contribution to the output due to interfering signals and noise arriving from other directions [11]. The major problem of beamforming is spatial aliasing which occurs for $(\lambda_\omega/2) \le v$, where v is the distance between the microphones λ_ω is the signal wavelength [12]. The spatial aliasing problem is resolved here by using a function $\xi[\kappa(\omega)]$ derived from the intensity difference between the sensors. Then the difference term $\Delta\phi(\omega)=[\phi_1(\omega)-\phi_2(\omega)]$ is represented as $\Delta\phi(\omega)=[\rho_\omega/\pi]\pm k$ where k=0, 1, 2… and $\rho_\omega \in (0,\pi)$.

Let $w^H=[w_1^*, w_2^*]$ be the sensor weight vector (superscript H represents the Hermitian transpose). $d(\varphi)=[1, e^{j2\pi\omega\tau(\varphi)}]^H$ is the sensor response vector when the source is placed at the azimuth angle φ. One sensor is considered as the reference and $\tau(\varphi)$ is the time delay (corresponding to ρ_ω) between the sensors and it is numerically computed from the location based transfer functions. The beamformer response $w^H d(\varphi)=g$ tells that at a specific temporal frequency ω, the signal of the source located at azimuth φ is passed to the output with gain g. Minimization of contribution to the output from the interference is accomplished by choosing the weights to minimize the output variance $E\{|y|^2\}=w^H R_x w$. The LCMVB problem of choosing the weights is thus written

$$\min_w w^H R_x w \quad subject\ to \quad d^H(\varphi)w = g \qquad (9)$$

The method of Lagrange multipliers can be used to solve Eq. (9) resulting in

$$w = g\frac{R_x^{-1}d(\varphi)}{d^H(\varphi)R_x^{-1}d(\varphi)} \qquad (10)$$

where $R_x = E\{XX^H\}$ is the data covariance matrix and

$$X=[X_1^{(\xi)}(\omega), X_2^{(\xi)}(\omega)]^H \qquad (11)$$

where $X_1^{(\xi)}(\omega)$ and $X_2^{(\xi)}(\omega)$ represent the phase-modified version by applying resolving function $\xi[\kappa(\omega)]$ between the original mixture spectra $X_1(\omega)$ and $X_2(\omega)$ respectively. The single linear constraint in Eq. (9) can easily be generalized for multiple constraints as:

$$\begin{bmatrix} d_1^H(\varphi_1) \\ d_2^H(\varphi_2) \end{bmatrix} w = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \qquad (12)$$

The Eq. (12) tells that it is desired to pass the sources at azimuth angle φ_1 and φ_2 with specified gain g_1 and g_2 respectively. It can be defined as $C^H w=G$ with constraint vector C_{2xL} and gain vector G of length L. Then the analytical solution for the optimum weight is:

$$w = R_x^{-1}C[C^{-1}R_x^{-1}C]^{-1}G \qquad (13)$$

With L constraints there are only 2-L degree of freedom (DOF) to minimize the variance. In this application two constraints (d_1 for target source and d_2 for nearest interfering source) are used. No DOF is available and hence a data independent beamformer is obtained. The gain vector G is set to as $G=[1, e^{-\gamma}]^T$, where γ is the Euclidian distance in 2D space between the target and the nearest interfering source. The evidence to use such constraints is that the target source is mostly affected by the nearest interfering signal.

The spectrum of a source at frequency ω is separated by projecting original mixture spectra $X_1(\omega)$ and $X_2(\omega)$ on to the weight vector $w(\omega)$. Obtaining the entire frequency spectrum, inverse FFT is applied to get back the short time windowed portion source signal in time domain. The same process is repeated for the number of sources indicated by the number of peaks in azimuth localization space. The above described beamforming technique is performed on each of the decomposed sub-band signals. Then the overall separation of a source signal is performed by summing up its corresponding resultant sub-bands.

V. Experimental Results And Discussion

The performance of the proposed algorithm is evaluated by separating the signals from binaural 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 azimuth angles from 0^o to 180^o with 5^o resolution are selected to compute the priori map of azimuth cue. The sampling rate of all the recording was set to 16 kHz 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(60°), sm2(100°), sf1(130°)}, m2{sm1(70°), sm2(120°), sf1(110°)}, m3{sm1(130°), sm2(40°), 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:

$$OSSR = \left| \frac{1}{T} \sum_{t=1}^{T} \log 10 \left(\frac{\sum_{i=1}^{w} s^2_{original}(t+i)}{\sum_{i=1}^{w} s^2_{separated}(t+i)} \right) \right| \quad (14)$$

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. In the case for zero energy in a particular window, no OSSR measurement is performed. If the two signals are same, OSSR=*0* and any other value is a measure of their dissimilarity. Table 1 shows the average OSSR of each signal for every mixture. Smaller value of OSSR indicates better separation. It is observed that the separation efficiency is degraded when the apart angle between the sources become smaller i.e. the sources are placed closely. Also the separation efficiency is compared between full-band (FB) and sub-band (SB) approaches of spatial beamforming. It is noticed that the sub-band based technique produces better separation than the full-band approach. The comparative evaluation of the performances of sub-band implementation using EMD (SB_{EMD}) and Fourier transform based filtering (SB_{FT}) is also presented in Table 1.

Table 1: Experimental results of our proposed algorithm

Mixtures		OSSR of sm1	OSSR of sm2	OSSR of sf1
m1	SB_{EMD}	0.132	0.145	0.201
	SB_{FT}	0.193	0.189	0.278
	FB	0.341	0.312	0.395
m2	SB_{EMD}	0.143	0.942	0.971
	SB_{FT}	0.169	0.937	0.992
	FB	0.247	0.983	1.031
m3	SB_{EMD}	0.867	0.194	0.861
	SB_{FT}	0.891	0.219	0.872
	FB	0.974	0.372	0.983

The evidence of improving the separation efficiency by using multi-band beamforming is that the band-limited or even the partial band-limited noise corruption does not affect the overall separation. The data adaptive time domain filtering using EMD plays a great support to the improvement of the separation efficiency with sub-band implementation of the spatial beamforming approach. The separation performance is better when EMD is employed to implement the multi-band decomposition than Fourier based method. In a multi-source audio environment, there occurs diffraction, interference etc. producing some band limited noise signals which affect the separation performance. Such type of noise does not spread over the entire frequency band during beamforming. In [1], separation is performed based on PD and ID independently (not combined) in azimuth localization. They have not localized the sources in spatial domain. A weighting filter is proposed to separate the target source only based on the values of PD and ID. There is a possibility of introducing noises from other interfering sources.

Usual beamforming approach to separate the sources from convolutive mixtures [10] used azimuth as DOA and to define the array response vector. The phase delay between the sensors is used as the main parameter to produce the response vectors that usually includes some ambiguous source with the higher frequency components of the mixture signals. The proposed algorithm makes use of a frequency dependent function to resolve such ambiguity and hence improves both localization and separation efficiency.

VI. Conclusions

We have proposed a model of localization based concurrent audio source separation from the binaural mixtures even in underdetermined situation. A strong source localization method is introduced based on priori map of interaural cues computed from the transfer functions between the source locations and microphones. The improvement in separation performance is achieved with the proposed sub-band beamforming technique. Another superiority of the algorithm is to introduce a potential function to resolve the spatial aliasing that improves the robustness of localization and beamforming scheme. To investigate the performance of the proposed method in noisy environment and to implement the localization based separation of concurrent moving sources are the main consideration as future work.

References

[1] H. Nakashima, Y. Chisaki, T. Usagawa and M. Ebata, "Frequency domain binaural model based on interaural phase and level differences", *Journal of Acoustic Sc. & Tech.,* Vol. 24, No. 4, pp. 172-178, April 2003.

[2] N. Roman, D. Wang and G. J. Brown, "Speech segregation based on sound localization", *J. Acoust. Soc. Am. 114(4)*, pp. 2236-2252, October 2003.

[3] M. S. Pedersen, L. K. Hansen, U. Kjems, and K. B. Rasmussen, "Semi-blind Source Separation Using Head-Related Transfer Functions," *ICASSP04*, 2004.

[4] H. G. Okuno, K. Nakadai, T. Lourens and H. Kitanp, "Separating three simultaneous speeches with two microphones by integrating auditory and visual processing", *Eurospeech03*, 2003.

[5] C. Faller and J. Merimaa, "Source localization in complex listening situation: Selection of binaural cues based on interaural coherence", *J. Acoust. Soc. Am.,* 116(5), November 2004.

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

[7] S. Rickard and O. Yilmaz, "On the W-Disjoint Orthogonality of Speech", *ICASSP'02*, May, 2002.

[8] P. Flandrin, G. Rilling, P. Goncalves, "Emperical Mode Decomposition as a filter bank" *IEEE Sig. Proc. Letter*, Vol. 11, No. 2, pp: 112-114, Feb 2004.

[9] N. E. Huang, et al., "Application of Hilbert-Huang transform to non-stationary financial time series analysis", *Applied Stochastic Model in Business and Industry*, 19:245-268, 2003.

[10] L. C. Para and C. V. Alvino, "Geometric Source Separation: Merging Convolutive Source Separation with Geometric Beamforming", *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 6, pp. 352-362, September 2003.

[11] B. D. Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering", *IEEE ASSP Magazine*, April 1988.

[12] A. Flaig and G. R. Arce, "Nearfield Spot-Beamforming with Distributed Arrays", Proc. of IEEE Int. Workshop on Sensor Array and Multichannel Signal Processing, 2000.