Enhanced Audio Source Separation in Room Acoustic Environments with Selected Binaural Cues

Namgook Cho, *Student Member*, IEEE and C.-C. Jay Kuo, *Fellow*, IEEE

***Abstract*--We study the problem of audio source separation from stereo-channel microphones in a room acoustic environment. Instead of using the whole binaural cues as done in DUET (Degenerate Unmixing Estimation Technique)-type methods, we propose a technique that selects a reliable subset of binaural cues by examining the phase determinacy and the sparse source conditions. We conduct experiments with a simulated room acoustic environment and show a significant performance gain of the proposed technique over the DUET-type methods.**

I. Introduction

Audio source separation is one of the emerging signal processing problems due to its rich applications in consumer electronics; *e.g.*, user-friendly interface to the infotainment system (*e.g.* the home media center) or handheld mobile devices via human voice of short sentences. It is often assumed in these applications that there is only one sound source (*i.e.*, speech) with background ambient noise [1]. However, this may not be the case practically. In this work, we examine the complex situation of multiple active sources under a room acoustic environment; *e.g.*, speech with background music.

DUET (Degenerate Unmixing Estimation Technique)-type methods have been proposed to solve the audio source separation problem [2]. They work well for a pair of microphones with a small spacing under the assumption that the delay between the microphones is not larger than one sample. However, the proposed technique fails to provide a satisfactory solution if the assumption is violated. It is worthwhile to point out that sounds could be acquired or recorded without meeting the microphone spacing constraint.

Here, we address the limitation of DUET-type methods and propose a new technique that selects a subset of binaural cues by examining the phase determinacy and the sparse source conditions. As compared with DUET-type methods that use the whole set of binaural cues, our solution can estimate mixing parameters more accurately and, therefore, achieve better results in audio source separation. The superior performance of the proposed technique has been extensively tested and representative examples are given in this work.

II. Problem Formulation

Consider a convolutive mixing model with N audio sources, denoted by $s_j(t), 1 \le j \le N$, and M microphones that yield linearly mixed signals, described by

$$x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t - \delta_{ij}), \quad i = 1, \cdots, M \tag{1}$$

where a_{ij} and δ_{ij} are scalar attenuation coefficients and time delay parameters, respectively, for the path from the jth source to the ith microphone. Without loss of generality, we set $\delta_{1j} = 0$. With stereo-channel mixtures, we can re-write the mixing model in the time-frequency domain as

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ a_{21} e^{-j\omega_0 l \delta_{21}} & \cdots & a_{2N} e^{-j\omega_0 l \delta_{2N}} \end{bmatrix} \begin{bmatrix} \hat{s}_1 \\ \vdots \\ \hat{s}_N \end{bmatrix} \tag{2}$$

where $\omega_0 = 2\pi / L$ and L is the length of the analysis window. Our objective is to recover unknown source signals s_j from observed mixtures x_i *only*, assuming $M (= 2) < N$ (*i.e.*, an underdetermined system of linear equations).

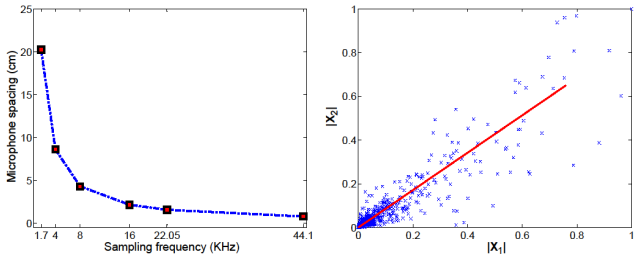

Figure 1. The minimum microphone spacing to avoid spatial aliasing as a function of sampling frequency (left) and the magnitude scatter plot at the frequency band 715.95 Hz, where three speech utterances are mixed under the echoic environment (RT_{60}=113 ms) [2] and the line represents the computed principal eigenvector (right),

Fig. 1 shows the minimum distance between microphones to prevent the occurrence of spatial aliasing as a function of sampling frequency. This condition can be met when free-standing microphones are close to one another. On the other hand, if the delay is larger than one sample [2], we are not able to obtain an accurate estimate of mixing parameters. In addition, some sound recordings may not satisfy the constraint of minimum microphone spacing, *e.g.*, KEMAR dummy head recording [3].

To address the problem, we propose a two-stage approach for convolutive audio source separation. Mixing parameters from mixtures are estimated in the short-time Fourier transform (STFT) domain. Then, we recover sources based on the estimated parameters and use the inverse STFT to

The authors are with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mails: namgookc@gmail.com and cckuo@sipi.usc.edu).

[2] RT_{60} is defined as the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound.

978-1-4244-4316-1/10/$25.00 ©2010 IEEE

reconstruct time-domain signals.

III. Proposed Technique

A. Estimation of mixing parameters

To avoid phase indeterminacy in (2), we choose proper TF-points that meet the following criterion: $\left|\omega_0 l \delta_{2j}\right| < \pi$. This is equivalent to the condition

$$\delta_{j\max} < \frac{\pi}{\omega_0 l} = \frac{L/2}{l}, \quad l = 1, \cdots, \frac{L}{2}, \tag{3}$$

where $\delta_{j\max} = \max_j \left|\delta_{2j}\right|$ and $\delta_{j\max}$ is the largest delay in the mixing system. Here, we estimate $\delta_{j\max}$ between stereo microphones using the GCC-PHAT method proposed in [4].

Acoustic signals, observed at microphones, usually contain not only the direct-path signals, but also attenuated and delayed replicas of the source signal due to reflections in a room. We use source sparsity to find TF-points where only the direct-path sound of a single source contributes to the mixtures. It is observed that sparse data points are aligned to a line in the magnitude mixture space. On the other hand, the non-sparse TF-points deviate from the oriented line. To determine sparsity of each TF-point, we measure a distance between the TF-point and the principal line of the scatter plot which is computed by the principal eigenvector of the covariance matrix, as shown in Fig. 1. With the selected TF-points, we obtain a subset of binaural cues, *i.e.*, attenuation ratio $\mathbf{x}_{\text{at}}$ and time delay $\mathbf{x}_{\text{d}}$. We obtain a smoothed 2D-histogram of selected binaural cues, and find peaks that correspond to sound sources to determine mixing parameters.

B. Recovery of source signals

After estimating mixing parameters, we use the minimum norm solution with l_p-norm criterion ($p \le 1$) [2] to solve the underdetermined system of linear equations in (2).

IV. Experimental Results

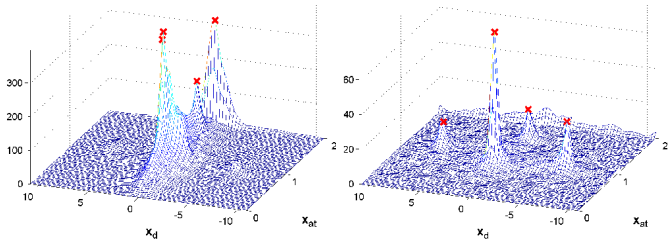

Figure 2. Comparison of histograms defined on the parameter space, where symbol "x" marks local peaks used to estimate mixing parameters over (left) the whole binaural cues and (right) the selection of binaural cues.

All sounds used in the experiments were downsampled to 16 KHz and had a length of 10 seconds. To measure the quality of reconstructed sounds with respect to the original one, the performance metrics suggested in [5] were used. Stereo recordings of several sources were simulated by convolving source signals with room impulse responses using the Roomsim toolbox [6].

Test Case 1: We compare smoothed histograms from a four-speech-source mixing example in Fig. 2, where the maximum time delay is larger than one sample ($\delta_{j\max}$ = 7.2 samples). The use of the whole TF-points yields spurious peaks and consequently, incorrect estimation of mixing parameters. These spurious points can be successfully eliminated using the proposed method with selected TF-points.

Test Case 2: Stereo mixtures of three speech utterances were simulated with $\delta_{j\max} > 1$. Table I shows the separation performance in a quiet-room environment with RT_{60}=44ms, where SDR, SIR and SAR denote the signal-to-distortion-ratio, the signal-to-interference-ratio and the signal-to-artifact-ratio, respectively. They are computed using the average for all extracted signals. A higher performance measure indicates a better reconstruction with less distortion. Table II reports the separation performance of the proposed technique in two different acoustic environments with RT_{60}=0 and 113 ms.

TABLE I: Comparison of two source separation algorithms

Minimum norm solution	The whole binaural cues			Selected binaural cues		
	SDR	SIR	SAR	SDR	SIR	SAR
p=1.0	-4.84	7.58	27.35	2.58	4.40	13.05
p=0.4	0.41	10.14	24.57	8.23	15.81	9.74

TABLE II: Performance of proposed technique

RT_{60}	0 ms			113 ms		
	SDR	SIR	SAR	SDR	SIR	SAR
p=1.0	2.26	4.06	12.16	-1.34	1.94	6.20
p=0.4	7.91	15.62	9.48	2.80	11.96	3.83

V. Conclusion

The limitation of DUET-type methods for convolutive audio source separation in room environments was identified. A new technique that selects proper binaural cues was proposed to offer a better audio source separation result.

References

[1] R. Flynn and E. Jones, "Robust distributed speech recognition using speech enhancement," *IEEE Trans. Consumer Electronic,* vol. 54, pp. 1267-1273, 2008.

[2] R. Saab *et al.*, "Underdetermined anechoic blind source separation via l^q-basis-pursuit with $q<1$," *IEEE Trans. Signal Process.*, vol. 55, pp. 4004-4017, 2007.

[3] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.,* vol. 97, pp. 3907-3908, 1995.

[4] C. H. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 24, pp. 320-327, 1976.

[5] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Process.,* vol. 14, pp. 1462-1469, 2006.

[6] K. P. D. Campbell and G. Brown, "A Matlab simulation of shoebox room acoustics for use in research and testing," *Computing and Inf. Syst. J.*, vol. 9, pp. 48-51, 2005.