UNDERDETERMINED AUDIO SOURCE SEPARATION FROM ANECHOIC MIXTURES WITH LONG TIME DELAY

Namgook Cho and C.-C. Jay Kuo

Ming Hsieh Department of Electrical Engineering and Signal and Image Processing Institute
University of Southern California, Los Angeles, CA 90089-2564, USA

ABSTRACT

We propose a technique to separate audio sources from their anechoic mixtures with long delay in an underdetermined setting (*i.e.*, the number of audio sensors is smaller than that of sources). It consists of two stages: 1) to estimate anechoic mixing parameters of attenuation and arrival delay and 2) to recover original audio sources based on estimated mixing parameters. When delay is longer than one sample, previous algorithms perform poorly. To address this shortcoming, we estimate the maximum delay and use it to find a proper frequency range that produces no phase ambiguity. Then, we determine mixing parameters with time-frequency points in this range. Finally, mathematical tools are used to solve the underdetermined linear system to recover original audio sources. The effectiveness of the proposed technique on various mixing scenarios with noisy observation of mixtures and different types of sounds is demonstrated by experimental results.

***Index Terms*—** Audio source separation, underdetermined mixing, delay estimation, sparse representation, multichannel audio.

1. INTRODUCTION

Audio source separation, which aims to estimate the original sources given acoustic mixtures of those sources, is one of the emerging research topics in recent years due to its many potential applications; *e.g.*, human voice extraction in noisy background with music and/or environmental sounds. In this work, we examine the underdetermined anechoic audio source separation problem, where the number of sources is greater than the number of mixtures, from noisy observations of anechoic mixtures of various sounds.

Recently, under the source sparsity assumption in a transform domain, the Sparse Component Analysis (SCA) was proposed to solve the audio source separation problem. The SCA-based method exploits clustering in the scatter plot and uses ℓ_1-norm optimization to solve the underdetermined source separation problem with limited success. Another approach was developed based on the ratios of time-frequency (TF) transforms of observations, which leads to the DUET-type methods [1,2]. One essential requirement of these methods is that sources must be strongly sparse in the analysis domain. However, it is difficult to estimate the mixing matrix accurately with clustering algorithms when the sources are not sufficiently sparse. Moreover, the sparsity condition can be violated in reality, *e.g.*, noise and the high degree of overlapping of sources.

Another shortcoming of the DUET-type methods is that they cannot yield an accurate mixing matrix if the arrival delay between multiple sources is longer than one sample due to the phase unwrapping ambiguity. Some solutions were proposed in [3] to solve the problem by finding the TF region where *only* one source is dominant. However, they still need the assumption of strong source sparsity (or low degree of source overlapping in the TF domain).

It is desirable to develop an approach that does not rely on the assumption of source sparsity and noise-free mixing. For example, we may consider mixtures of music and/or environmental sounds (rather than only speech signals) that have a wide range of spectral and temporal characteristics and overlapping of source signals in the TF domain is strong. Here, we propose a new method to estimate the anechoic mixing model for underdetermined audio source separation. Being inspired by human audition in interaural time difference, we estimate the maximum delay and use it to find a proper frequency range that produces no phase ambiguity. Using TF points in this range, we can identify a parameter space of attenuation ratios and time delay and determine mixing parameters accordingly. Then, several mathematical tools are used to solve the underdetermined linear system so as to separate original audio sources. Experimental results demonstrate the effectiveness of the proposed technique on various mixing scenarios using noisy observation of mixtures and different types of sounds.

The rest of this paper is organized as follows. The problem of underdetermined audio source separation from anechoic mixtures is formulated in Sec. 2. The proposed solution is described in Sec. 3. Experimental results are presented and discussed in Sec. 4. Finally, concluding remarks and future research work are given in Sec. 5.

2. PROBLEM FORMULATION

Consider an anechoic mixing model with N audio sources, denoted by $s_j(t)$, $1 \le j \le N$, and M audio sensors (or microphones) that yield linearly mixed signals. This mixing process can be 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 the scalar attenuation coefficient and the time delay parameter, respectively, for the path from the jth source to the ith microphone. Without loss of generality, we set $\delta_{1j} = 0$ and scale sources with $\sum_{i=1}^{M} |a_{ij}|^2 = 1$, for $j = 1, \cdots, N$. In this work, we assume $M < N$, *i.e.*, the mixing is underdetermined and the number of sources N is known *a priori*. The goal is to recover unknown source signals from observed mixtures only.

Instead of solving the problem in the time domain, we apply the time-frequency transformation to mixture signals. By using the short-time Fourier transform (STFT) with a fixed window function,

E-mails: namgookc@usc.edu and cckuo@sipi.usc.edu

we can re-write the mixing model in Eq. (1) as

$$\hat{\mathbf{x}}[k,l] = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ a_{21}e^{-j\omega_0 l\delta_{21}} & \cdots & a_{2N}e^{-j\omega_0 l\delta_{2N}} \\ \vdots & \ddots & \vdots \\ a_{M1}e^{-j\omega_0 l\delta_{M1}} & \cdots & a_{MN}e^{-j\omega_0 l\delta_{MN}} \end{bmatrix} \hat{\mathbf{s}}[k,l], \quad (2)$$

where $\hat{\mathbf{x}} = [\hat{x}_1, \cdots, \hat{x}_M]^T$ and $\hat{\mathbf{s}} = [\hat{s}_1, \cdots, \hat{s}_N]^T$ are the STFT of mixtures and sources, respectively. The use of STFT has several advantages. First, the convolutive mixtures in (1) reduces to instantaneous ones in each TF point $[k,l]$. Second, we can exploit the sparsity of source components, which plays a key role in our work. Generally, sources are not sparse in the time domain.

3. PROPOSED SOLUTION

To solve the problem formulated in Eq. (1), we employ a two-stage approach for underdetermined anechoic audio source separation. First, we apply the short-time Fourier transform (STFT) to mixtures and estimate mixing parameters, a_{ij} and δ_{ij}, from mixtures. Then, we recover sources based on the estimated parameters and use the inverse STFT to reconstruct time-domain signals. These two modules are described in detail in the following two subsections.

3.1. Estimation of Mixing Parameters

In this subsection, we propose a parameter estimation technique that is able to efficiently extract good features with no phase ambiguity under the assumption of source sparsity. The algorithm is motivated by human audition for source localization. At low frequencies, since sound's wavelength is much longer than the human head diameter, the phase difference between signals received by two ears can be estimated with no ambiguity. In contrast, there can be several cycles of shift in high frequencies, which results in phase ambiguity of interaural time difference. Thus, our goal is to find the frequency range that produces no phase ambiguity, and then use TF points located in this range to construct the feature space for parameter estimation.

For example, to avoid phase indeterminacy in Eq. (2) with stereophonic observations (*i.e.*, $M = 2$), we should choose good TF points $[k,l]$ that meet the following criterion:

$$|\omega_0 l\delta_{2j}| < \pi, \quad (3)$$

where $\omega_0 = 2\pi/L$ and L is the length of the analysis window. Let $\delta_{jmax} = \max_j |\delta_{2j}|$, where δ_{jmax} is the largest delay in the mixing system. Clearly, condition (3) is guaranteed for all j if $\omega_0 l\delta_{jmax} < \pi$. This is equivalent to the condition

$$\delta_{jmax} < \frac{\pi}{\omega_0 l} = \frac{L/2}{l}, \quad l = 1, \cdots, \frac{L}{2}. \quad (4)$$

Thus, we can determine the frequency range that satisfies the phase determinacy condition based on the condition in (4).

To estimate the maximum time delay between multiple channels, we use the GCC-PHAT method proposed in [4] to compute the inverse Fourier transform of the weighted cross-power spectrum of the mixture signals. It is widely acknowledged that GCC-PHAT is more immune to reverberation and able to provide consistent performance when the characteristics of the source signal change over time. With GCC-PHAT, the time delay estimate can be obtained mathematically as

$$\hat{\delta}_{jmax} = \arg\max_m \Psi_{\text{PHAT}}[m], \quad (5)$$

where

$$\Psi_{\text{PHAT}}[m] = \text{FT}^{-1}\{S_{\mathbf{x}}/|S_{\mathbf{x}}|\}$$

and $S_{\mathbf{x}}$ is the cross-power spectrum of the mixture signals.

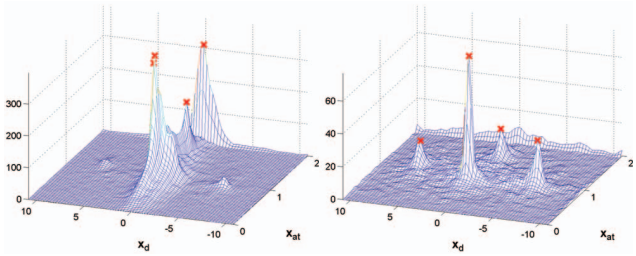
Next, using TF points that satisfy the phase determinacy condition, we can define a parameter space of the attenuation ratio and time delay. Suppose that only a particular source j is significantly different from zero at TF point $[k,l]$. Then, the mixing model in Eq. (2) becomes

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \begin{bmatrix} a_{1j} \\ a_{2j}\, e^{-j\omega_0 l\delta_{2j}} \end{bmatrix} \hat{s}_j.$$

Thus, at TF point $[k,l]$, the attenuation ratio $\mathbf{x}_{at}$ and time delay $\mathbf{x}_d$ can be written as

$$\mathbf{x}_{at} = \frac{a_{2j}}{a_{1j}} = \left|\frac{\hat{x}_2[k,l]}{\hat{x}_1[k,l]}\right| \text{ and } \mathbf{x}_d = -\frac{1}{\omega_0 l}\angle\frac{\hat{x}_2[k,l]}{\hat{x}_1[k,l]}. \quad (6)$$

To determine mixing parameters, we rely on features computed by Eq. (6) and construct a smoothed histogram in the two-dimensional (2D) parameter space $(\mathbf{x}_{at}, \mathbf{x}_d)$. After that, we determine the mixing parameters by locating peaks in the 2D histogram.



(a) (b)

Fig. 1. Comparison of histograms defined on parameter space $(\mathbf{x}_{at}, \mathbf{x}_d)$, where symbol "x" marks local peaks used to estimate mixing parameters: (a) the full-frequency method and (b) the partial-frequency method.

We compare histograms from a four-speech-source synthetic mixing example in Fig. 1, which are computed by two methods (*i.e.* the full-frequency and partial-frequency methods), where the maximum time delay is larger than one sample (δ_{jmax} = 7.2 samples in this example). The use of the whole frequency range yields spurious peaks and, consequently, incorrect estimation of mixing parameters as shown in Fig. 1 (a). These spurious points can however be successfully eliminated as shown in Fig. 1 (b) using TF points in a partial frequency range that meet the condition specified by Eq. (4). Each peak location labeled by symbol "x" corresponds to one pair of mixing parameters. Since $\hat{\delta}_{jmax}$ was computed as 7 samples by GCC-PHAT in this example, the frequency interval $(0, 1120]$ Hz was used to confine the parameter space. In contrast, if δ_{jmax} is less than one sample, both *partial-* and *full-*frequency methods can estimate the mixing parameters successfully.

It is worthwhile to emphasize our contribution along this line. The original goal of Eq. (5) is not to estimate time delay parameters δ_{ij} between multiple audio sources, but the maximum time delay δ_{jmax} among them. Here, we obtain good results for time delay estimation based on the restriction on the frequency range of our interest introduced by (4). In contrast, the traditinal method, which uses the whole frequency range, yields spurious points in the parameter space due to phase ambiguities.

3.2. Estimation of Audio Sources

Based on estimated mixing parameters, the mixing model in Eq. (2) can be written as

$$\hat{\mathbf{x}}[k,l] = \widehat{A}[l]\,\hat{\mathbf{s}}[k,l], \tag{7}$$

where $\widehat{A}[l] \in \mathbb{C}^{M\times N}$ is the estimated mixing matrix. Our goal is to find estimates $\hat{\mathbf{s}}$ of the original sources. However, unmixed signals cannot be directly obtained since the mixing matrix given in Eq. (7) is underdetermined. That is, at each TF point, the mixing model has more unknowns (N) than constraints (M). Several mathematical techniques have been proposed to solve the underdetermined system of linear equations. Among them, the sparsity of the source vector has been widely and successfully exploited. For example, the minimum norm solution with ℓ_1-norm or ℓ_p-norm criterion ($p < 1$).

To find the sparsest $\hat{\mathbf{s}}[k,l]$ at each TF point, Eq. (7) can be formulated as the following optimization problem:

$$\min_{\hat{\mathbf{s}}} \|\hat{\mathbf{s}}\|_p \quad \text{subject to} \quad \widehat{A}\,\hat{\mathbf{s}} = \hat{\mathbf{x}}, \tag{8}$$

where $0 < p \leq 1$. It has been known in [2] that the N-dimensional vector $\hat{\mathbf{s}}$ that solves Eq. (8) contains at least $N - M$ zeros if the columns of $\widehat{A}$ are normalized. Based on this result, it is possible to employ a combinatorial algorithm (CA) to solve Eq. (8). That is, one can find set $\mathcal{A}$ that contains all $M \times M$ invertible submatrices from $\widehat{A}$ and chooses the one that offers a solution with the minimum norm as

$$\min \|B^{-1}\,\hat{\mathbf{x}}\|_p, \qquad B \in \mathcal{A}, \tag{9}$$

where $B_{M\times M} = [\mathbf{a}_{i1}, \cdots, \mathbf{a}_{iM}]$ and the possible number of B is equal to C_N^M. To solve the problem in Eq. (9), Winter *et al.* [5] employed the ℓ_1-norm constraint while Saab *et al.* [2] showed experimentally that the separation performance can be improved when one uses ℓ_p-norm with $p < 1$.

In this work, we also employ FOCUSS algorithm [6] which is not based on the combinatorial optimization as described above, but on iteratively re-weighted norm minimization of the source vector. It is observed that the FOCUSS algorithm admits a sparser solution (*i.e.*, more than $N - M$ components in vector $\hat{\mathbf{s}}$ can be zero).

4. EXPERIMENTAL RESULTS

The performance of the proposed solution technique is evaluated in this section. To measure the quality of reconstructed sounds with respect to the original one, the performance metrics suggested in [7] were used, including the source-to-distortion-ratio (SDR), the source-to-interference-ratio (SIR), and the source-to-artifact-ratio (SAR). A higher performance measure indicates a better reconstruction with less distortion.

Test signals were chosen from several excerpts of audio sounds: male/female speech utterances, recordings of musical instruments and environmental sounds. All sounds used in the experiments were downsampled to 16,000 Hz and had a length of 10 seconds. The frame size L of the Hanning window was set to 512 samples and the shifting interval of the frame was 256. We examined underdetermined mixtures with $N = 3$ and $M = 2$.

Stereo recordings of several sources were simulated by convolving the source signals with room impulse response using the Roomsim toolbox [8]. The positions of the omnidirectional microphones and loudspeakers are illustrated in Fig. 2. In the configuration, the maximum time delay between microphones is larger than one sample.

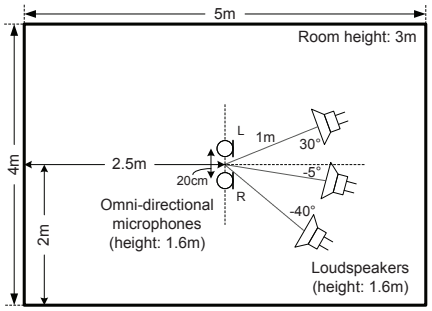


Fig. 2. The configuration of loudspeakers and microphones in a room, where recordings were simulated with an absorption coefficient of 0.9 for room's surface.

4.1. Estimation of Mixing Parameters

We consider two settings in the generation of simulated room recording data:

- Case A: mixtures of three speech utterances;
- Case B: mixtures of one speech utterance and two musical sounds (flute and acoustic guitar).

In Fig. 3, we present smoothed histograms in the parameter space in these two settings. Figs. 3 (a) and (b) are obtained from Case A while Figs. 3 (c) and (d) are obtained from Case B.

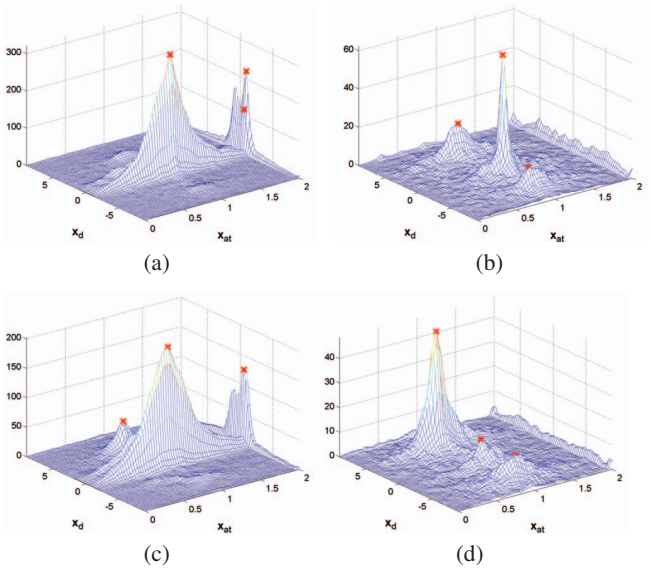


Fig. 3. Smoothed histograms for the mixtures of three sources computed from simulated room recordings: (a) Case A with full-frequency data, (b) Case A with partial-frequency data, (c) Case B with full-frequency data, and (d) Case B with partial-frequency data.

In Figs. 3 (a) and (c), the conventional approach proposed in [1, 2] was employed, where data in the whole frequency range were used in parameter estimation. The approach yields spurious points and peaks in the parameter space. In contrast, we used the partial frequency range that meets the condition in Eq. (4) to construct good

features. The results are plotted in Figs. 3 (b) and (d) for comparison. It is clear that peaks of histograms in Figs. 3 (b) and (d) can be easily located. It is observed that poorly estimated mixing parameters degrade the audio source separation performance significantly in the source recovery stage.

4.2. Reconstruction of Audio Sources

We conducted audio source separation experiments with simulated room recordings of several sources based on the room configuration in Fig. 2. Table 1 shows the separation performance in terms of SDR, SIR, and SAR, where each performance metric is computed based on the average for all extracted signals. The combinatorial algorithm (CA) with $p = 0.4$ yields the best SDR and SIR performance. To understand the sensitivity of paramter p, we plot the SDR and SIR performance curves as a function of p in Fig. 4. We see that any choice of $0.1 \leq p \leq 0.9$ provides equally good performance.

For comparison, the conventional approach proposed in [1, 2] fails to separate source signals due to their incorrect estimation of mixing parameters when $\delta_{jmax} \geq 1$; for example, –3.79, –3.37 and –4.39 dB were obtained on SDRs of CA ($p = 1$), CA ($p = 0.4$), and FOCUSS ($p = 0.4$), respectively, for the example of speech+piano+guitar mixtures in the table.

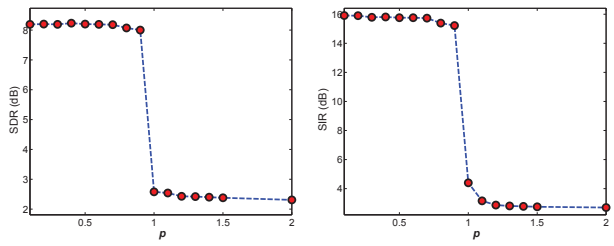
Table 1. Source separation example with stereo mixtures of three sources with $\delta_{jmax} \geq 1$.

Mixtures	speech + piano + guitar			speech + cello + train		
	SDR	SIR	SAR	SDR	SIR	SAR
$\text{CA}_{p=1.0}$	5.05	10.68	8.39	2.30	7.25	6.83
$\text{CA}_{p=0.4}$	6.08	13.24	8.13	4.73	11.00	7.53
$\text{FOCUSS}_{p=0.4}$	3.55	12.40	4.54	0.82	2.95	5.74

Finally, we tested our algorithm with noisy observation of anechoic recordings to understand the effect of noise on the sparsity assumption. In the past, source separation methods were proposed under the assumption that the effect of noise on the mixtures is negligible, *e.g.,* [1, 2, 5] . However, noise and/or overlapping sources do affect the accuracy of mixing model recovery as well as audio source estimation. Fig. 5 shows the separation performance in terms of SDRs as a function of the additive white Gaussian noise level (*i.e.*, the Source-to-Noise-Ratio). The performance degrades as the SNR value decreases although the proposed algorithm estimated mixing parameters successfully. Note that FOCUSS yields slightly robust performance for SDRs since it can have a sparser solution than the combinatorial algorithms. For results of FOCUSS, our own listening tests indicate that there is little interference from sources other than extracted one, but there exist some artificial distortion, known as the "musical noise" artifact, due to forced zeros in the FT points.

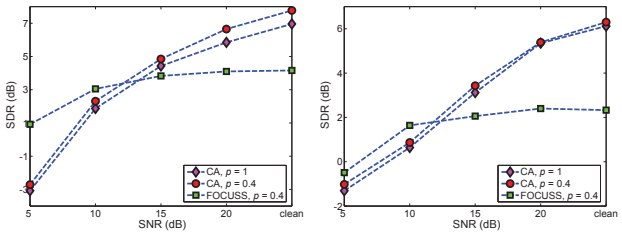
5. CONCLUSION AND FUTURE WORK

We examined main assumptions and limitations of SCA-based methods for underdetermined anechoic audio source separation and presented a new technique to overcome their limitations. The proposed technique provides good performance under noisy observations with different source types, mixing conditions and longer delay (with the delay time larger than one sample). In the future, we would like to extend the proposed framework to an array of microphones and use the beamforming technique to improve the performance furthermore.



(a) (b)

Fig. 4. The averaged values of (a) SDR and (b) SIR as a function of p in CA with three estimated speech sources where $\delta_{jmax} \geq 1$.



(a) (b)

Fig. 5. The averaged values of SDR as a function of SNR with three estimated sources where $\delta_{jmax} \geq 1$: (a) mixtures of speech, piano and guitar, and (b) mixtures of guitar, piano and cello.

6. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, pp. 1830–1847, 2004.

[2] R. Saab, O. Yilmaz, M. J. McKeown, and R. Abugharbieh, "Underdetermined anechoic blind source separation via ℓ^q-basis-pursuit with $q < 1$," *IEEE Trans. Signal Process.*, vol. 55, pp. 4004–4017, 2007.

[3] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear anechoic mixture," in *IEEE Int. Conf. Audio, Speech, Signal Process.*, 2007, pp. 745–748.

[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] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP J. Adv. in Signal Process.*, 2007.

[6] I. F. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, pp. 600–616, 1997.

[7] 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.

[8] 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.