14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

BLIND SOURCE SEPARATION FOR CONVOLUTIVE MIXTURES USING SPATIALLY RESAMPLED OBSERVATIONS

J.-F. Synnevåg and T. Dahl

Department of Informatics, University of Oslo
P.O. Box 1080, N-0316 Oslo, Norway

ABSTRACT

We propose a new technique for separation of sources from convolutive mixtures based on independent component analysis (ICA). The method allows coherent processing of all frequencies, in contrast to the traditional treatment of individual frequency bands. The use of an array enables resampling of the signals in such a way that all frequency bands are effectively transformed onto the centre frequency. Subsequent separation is performed "all-bands-in-one". After resampling, a single matrix describes the mixture, allowing use of standard ICA algorithms for source separation.

The technique is applied to the cocktail-party problem to obtain an initial estimate of the separating parameters, which may further be processed using cross-talk removal or filtering. Experiments with two sources of speech and a four element microphone array show that the mixing matrix found by ICA is close to the theoretically predicted, and that 15 dB separation of the sources is achieved.

1. INTRODUCTION

A vast number of techniques have been developed for blind source separation (BSS) and extraction (BSE) over the last decades, and the field of BSS and BSE spurs hundreds of paper every year (see [1] for a recent survey).

Many BSS techniques are based on the transition from time domain convolutive mixtures to frequency domain instantaneous mixtures. A well known problem with this approach is the permutation and scaling inconsistencies which leads to a ”re-mixing” of the sources when the frequency separated sources are transformed to the time-domain. As a consequence, numerous papers have been published that deal with this inconsistency, for a recent overview see [2, 3].

An interesting comment about the limitation of frequency based BSS was made by Araki et al [4], showing that such techniques are essentially limited by the performance of an adaptive beamformer. This is down to an argument that works across each and every frequency: The echoes from the echoic environment will appear as directional signals coming into the array, essentially taking the place as "new signals". For every frequency, it holds true that the limitations in the number of zeros that can be placed over the angular directions is limited, and that a tradeoff must be struck between the strength of the signal in the direction of the desire, and the various suppression levels that can be put on other directions. A direct consequence of Araki's statements, is that frequency-band based BSS is heavily overparamaterized: Since the optimal result is obtained by zero forcing in the same positions along all frequencies, separation could ideally be derived for all frequency bands by "extrapolating" the zero-forcing settings estimated for a single band. Assuming all sources to occupy the same frequency band, or that no source separation can be based on frequency contents alone, it follows that performing ICA on each and every bin is a potential waste of computer power, since analysis of the bands essentially outputs the same directional separation-information. Given the limitations of frequency-based BSS, post-filtering and cross-talk removal is necessary to improve separation.

Other scientists pursue the time-domain approach [5] to avoid dealing with this inconsistency, but such methods easily become very complex. Yet others use combined approaches, a recent method computes inversion filters in time domain while using a cost function in frequency domain [6]. Time-frequency signatures [7] is a powerful tool that may be used for separation even in cases where there are more sources than mixes.

We propose a technique which has the potential for utilizing advantages of both time-domain and frequency-domain BSS. By using spatial resampling along the array direction in the temporal frequency domain, every frequency band of the original signal is ”forced” onto the same spatial frequency. This enables an ICA-like representation of the BSS problem, avoiding the use of multiple frequency bins and the resulting permutation inconsistency. We present real-life experiments based on this approach and discuss its limitations.

It should be noted that we are no longer attempting the ”original” cocktail-party problem, but rather a modified problem with a lower number of estimated separation parameters than would be required for perfect separation in an echoic envirmonent. In [8], the authors propose a technique which incorporates knowledge of the microphone setup. However, the proposed techniques do no not involve the spatial resampling key step we propose here, and is hence quite different from the material in the present paper.

2. METHOD

2.1 Signal model

The use of independent component analysis for blind separation of independent sources requires observations of the form

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where $\mathbf{x} \in R^M$ a vector of observations in M mixes , $\mathbf{s} \in R^N$ is a vector containing samples from N independent sources and $\mathbf{A} \in R^{M \times N}$ is the mixing matrix describing how the sources are observed.

The time-domain model for the cocktail party is more complex. To build intuition around the problem, we first consider a simplified, anechoic model, with no distortion effects at the microphones. Each observation, $x_j(t)$, for $j = 1, \ldots, M$ can then be modeled as

$$x_j(t) = \sum_{i=1}^{N} s_i(t) * \frac{1}{r_{ij}} \delta(t - \tau_{ij}), \tag{2}$$

where $x_j(t)$ is the output of the jth microphone, s_i is independent component i (the ith speaker), $\delta(\cdot)$ is the Dirac delta-function, $*$ is the convolution operator, r_{ij} is the distance from speaker i to microphone j, τ_{ij} is the sound propagation delay from speaker i to microphone j, and N is the number of speakers. Whereas the ICA model in (1) does not capture the linear convolution required to describe the time-delays between the sources and observation points, the model (2) encompasses this possibility. The full convolutive BSS model, containing possible echoes from multiple directions as well as filtering and attenuation effects in space, time and linear equipment distortion is

$$x_j(t) = \sum_{i=1}^{N} s_i(t) * b_{ij}(t), \tag{3}$$

where $\{b_{ij}(t)\}$ is a set of FIR filters describing the contributions of each of the sources indexed by j on each of the mixes indexed by i. The traditional way to adapt the echoic cocktail party problem on to the form (1) is by transforming the observations to the frequency domain and treat each frequency independently, avoiding the convolution. Also, methods dealing purely with time-shifting and attenuation in the echo-free scenario [9] have been proposed. However, these do not take full advantage of the possibilities of array processing.

2.2 Frequency domain representation

Consider the model (2) in the frequency domain. The observations for one narrow frequency bin can be written as

$$x_j(\omega) = \sum_{i=1}^{N} a_{ij}(\theta) S_i(\omega), \tag{4}$$

where $S_i(\omega)$ is the amplitude of the ith source at frequency $\omega/2\pi$. Collecting terms from the M mixes into the vector $\mathbf{x}(\omega) = [x_1(\omega), x_2(\omega), ..., x_N(\omega)]^T$ and from all the mixes and the one source i into the vector $\mathbf{a}_i(\theta) = [a_{i1}(\theta), a_{i2}(\theta), \ldots, a_{iM}(\theta)]^T$, we can write

$$\mathbf{x}(\omega) = \sum_{i=1}^{N} \mathbf{a}_i(\theta) S_i(\omega) \tag{5}$$

where $\mathbf{a}_i(\theta)$ contains the time-delays between arrivals at the different sensors for the ith source. $\mathbf{a}_i(\theta)$ is known as the *steering vector* [10] in array processing. In matrix form, (5) can be written as

$$\mathbf{x}(\omega) = \mathbf{A}(\omega)\mathbf{S}(\omega), \tag{6}$$

where

$$\mathbf{A}(\omega) = [\mathbf{a}_1(\theta), \ldots, \mathbf{a}_M(\theta)] \tag{7}$$

and

$$\mathbf{S}(\omega) = [s_1(\omega), \ldots, s_N(\omega)]^T. \tag{8}$$

The problem (6) is now on the same form as (1), but each frequency has to be treated independently.

2.3 Transformation from convolutive to instantaneous mixtures

We propose to transform the convolutive model in (2) to the linear sum in (1) using a technique from array processing. The transformation requires an array of sensors with known geometry, as both the temporal and spatial frequency spectrum of the recorded wavefield must be captured. With this knowledge, spatial resampling is performed along the array direction for each temporal frequency band, effectively transforming the observations onto the ICA form.

2.3.1 *Spatial frequency: The wavenumber*

First, we explain the term *spatial frequency*, which is central in array signal processing. Imagine an acoustic wave measured in a single point in space and consider the *temporal* frequency of the signal. The question at hand is then how many times the signal oscillates within a given time span. Of course, a speech signal consists of many superpositioned waves oscillating with different periods, giving a whole spectrum of frequencies. Moving on to *spatial frequency*, the question is slightly different: If the wavefield is observed, not in a single point, but along a directive line segment in space, how many times does the wave oscillate within the line segment? This situation is illustrated in figure 1, which shows a narrowband, plane wave propagating in the xy-plane. The solid lines are the wave-fronts, meaning the lines of constant phase of the travelling wave, and the arrow indicates the direction of propagation. The number of periods of the wave fitting into a line segment of limited length in the xy-plane gives a measure of the spatial frequency in the direction of the line. Clearly, the spatial frequency will vary with the direction of the segment. In this example, if we look towards the wavefront, the spatial frequency will be higher than if the segment is placed along the x-axis. To be able to measure the frequency along a spatial dimension, we need access to data sampled along a line in space. This requires the use of an array.

The spatial frequency along the direction of propagation is called the *wavenumber* and is given by

$$k = \frac{\omega}{c}, \tag{9}$$

where c is the propagation velocity of the medium. The *wavenumber vector*, $\vec{k}$, contains the spatial frequencies along each spatial dimension of the wavefield, and satisfies the relation

$$|\vec{k}| = k. \qquad (10)$$

By using a linear array of sensors, we sample one spatial dimension of the wavefield, and can estimate the wavenumber component in that direction. In figure 1 an array is located along the x-axis. The spatial frequency along this dimension is given by

$$k_x = k\sin(\theta), \qquad (11)$$

where θ is the propagation angle, defined clockwise with respect to the y-axis. For the remainder of the paper we will refer to the wavenumber component along the array dimension simply as the wavenumber.

We denote the wavefield along the x-axis $z(x,t)$. By placing a microphone every d meters we can describe the sampled wavefield as

$$y_m(n) = z(md, nT), \qquad (12)$$

where m is the sensor number, n is the temporal sample number, and T is the sampling interval. Similar to estimating the temporal frequency of the sampled signal using the discrete Fourier transform as

$$Y_m(\omega) = \sum_{n=0}^{L-1} y_m(n) e^{-j\omega nT}, \qquad (13)$$

where L is the number of temporal samples, we can estimate the wavenumbers along the array direction as

$$Y(k_x) = \sum_{m=0}^{M-1} y_m(n) e^{-jk_x md}. \qquad (14)$$

The wavenumber-frequency response is a summation over both time and space,

$$Y(k_x, \omega) = \sum_{m=0}^{M-1} \sum_{n=0}^{L-1} y_m(n) e^{-j\omega nT} e^{-jk_x md}. \qquad (15)$$

For narrowband waves, ω is fixed, and k_x will change as a function of propagation direction of the wave. If the temporal frequency of the signal is known, the propagation angle can be found using (11). Figure 2 (a) and (b) shows two examples of sinusoidal waves sampled in space and time by an eight channel linear array. Figure (c) shows spatial samples at a selected point in time. For the wave in figure (a) which propagates in a direction orthogonal to the array, the amplitude is the same on all channels, because the signal arrives at the same time on all sensors. The wavenumber is then zero. For the signal propagating with non-zero angle, a sinusoidal pattern is evident over the array, leading to a non-zero wavenumber. Figure (d) shows the estimated wavenumbers for the two examples. The corresponding angle of arrival is shown on the top axis.

Broadband waves are described by superpositioning narrowband waves with different temporal frequencies. Broadband waves correspond to lines in wavenumber-frequency space. Figure 3 (a) shows the estimated wavenumber-frequency spectra of three broadband waves propagating across a uniform linear array. For non-zero angles of arrival, the wavenumber increases linearly with frequency, where the slope depends on the angle of arrival. All temporal frequencies of signals originating perpendicular to the array appear with zero wavenumber. Going back to the frequency domain model of the cocktail-party problem, the steering vector modeling the observed signal is simply $\mathbf{a}(\theta) = [1\ 1\ \ldots\ 1]^T$ for all frequencies, hence no time-delays are required to describe the observations. For signals arriving at an angle, the steering vector is frequency-dependent and given by $\mathbf{a}(\theta) = [1\ e^{jk_x d}\ \ldots\ e^{jk_x(M-1)d}]^T$. If the wavenumber was constant regardless of frequency for any incidence angle, the steering vector would be identical for all frequencies, and a single mixing matrix would describe the observations of several sources. The problem would then be of the form (1). Spatial resampling [11] is a means to achieve that.

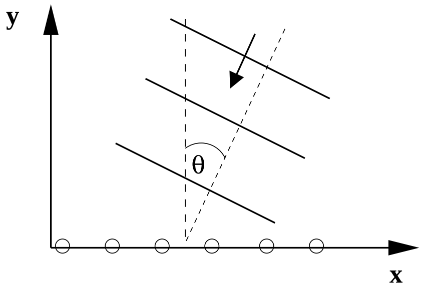
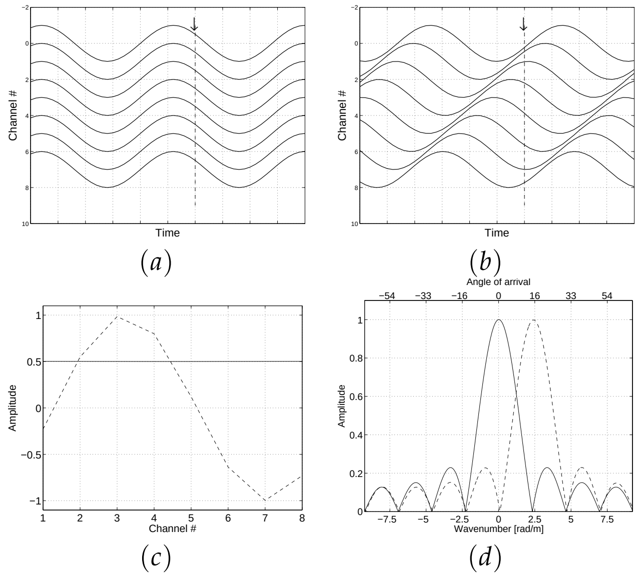


Figure 1: Illustration of a plane wave propagating in the xy-plane.

2.3.2 *Spatial resampling*

The goal of spatial resampling is to force all temporal frequency components of a wave to appear with the same wavenumber. We choose a center frequency, f_c, where the corresponding wavenumber will appear for all frequencies. For $f > f_c$ the spatial "sampling rate" is increased by a factor f/f_c by interpolating between samples. Each sample corresponds to a point in space, given by the sensor location. The original number of samples are kept, effectively reducing the size of the aperture as we are throwing away samples at the edges. For frequencies $f < f_c$, we decrease the sampling rate. Since we must keep the original number of samples, the latter case introduces zeros at the edges as information is missing outside the original aperture. Figure 3 (b) shows the corresponding frequency-wavenumber plot of figure 3 (a) after spatial resampling. We see that all frequency components appear with the same wavenumber. After the resampling step, the observations are transformed back to the time domain. Each temporal frequency component of the observations are now described by identical steering vectors, and we have a set of observations on ICA form (1).



(*a*) (*b*)

(*c*) (*d*)

Figure 2: Simulated time-series from an 8-channel linear array recording a plane wave propagating with (a) 0° and (b) 17° angle. (c) Spatial samples at the time instance given by the arrows. (d) The estimated wavenumber of the propagating waves. The top axis shows the corresponding propagation angles.

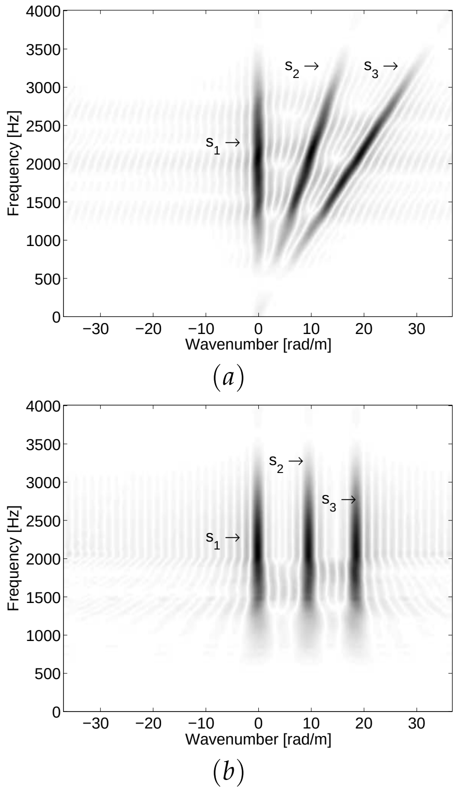
3. RESULTS

We have evaluated the performance of the method experimentally with a linear array of four microphones in an anechoic chamber. The experimental setup is shown in figure 4. The microphones were separated by 30 cm. Two speakers were placed in front of the array, each transmitting a different speech signal. The distance between the speakers was 1 m, giving propagation angles of 0° and approximately 17° for the sources, with reference to the center of the array. The recorded signals were bandpass filtered, passing frequencies between 300 Hz and 3 kHz. We resampled the observations spatially, using 500 Hz as resampling frequency. The sources were then separated with the JADE algorithm [12].

We evaluated the performance of the separation using the signal-to-interference ratio (SIR) for each of the estimated sources, given by the power in the desired signal to the power in the interfering signals. The SIR for source 1 was estimated as

$$S\hat{I}R(\hat{s}_1) = \left(\frac{E\left(\hat{s}_1(t)s_1(t)\right)}{E\left(\hat{s}_1(t)s_2(t)\right)}\right)^2 \qquad (16)$$

where $\hat{s}_1$ is estimate of source 1, $s_1(t)$ is the true source 1 and $s_2(t)$ is the true "interfering" source. Both $s_1(t)$ and $s_2(t)$ were normalized to unit variance before evaluation. This measure assumes that the estimate of source 1 contains scaled versions of $s_1(t)$ and $s_2(t)$, which is a simplification as no filtering effects are taken into account. The equivalent measure was calculated for source 2.



(*a*)

(*b*)

Figure 3: Illustration of the effect of spatial resampling. The s_i represent three different broadband waves propagating across a linear array in different directions. The figures show frequency-wavenumber plots (a) before and (b) after spatial resampling.

To evaluate the performance of the method, we first looked at how well the estimated mixing matrix compared to the theoretically predicted. Note that the theoretical mixing matrix relates to the resampled observations, hence the columns of this matrix is the steering vectors in (5) with $\theta = 0°$ and $\theta = 17°$ for frequency $f_c = 500$ Hz. It is most instructive to look at the wavenumber response of the steering vectors, as its peak corresponds to the propagation angle. Figure 5 shows the wavenumber response of the first column of the theoretical mixing matrix and the corresponding vector found by ICA. Note that for this, and the remaining plots, wavenumber has been translated to propagation angle using (11). The dashed vertical lines shows the true propagation angles. We see that the peak of the steering vector found by ICA corresponds to the true propagation angle for source 2.

Rather than looking at the mixing matrix, it is more instructive to study the unmixing matrix, as its rows correspond to spatial unmixing filters for each source. In the directions of interfering signals the response should be zero for perfect separation. Figure 6 shows the wavenumber response of the second row of the unmixing matrix found by ICA, together with the response of the theoretical unmixing filter for the second source. We see that there is a positive gain in the direc-

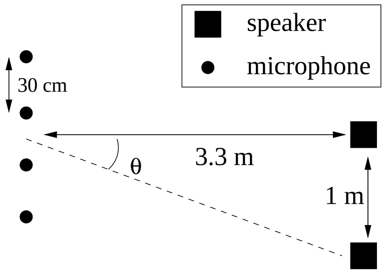


Figure 4: Experimental setup.

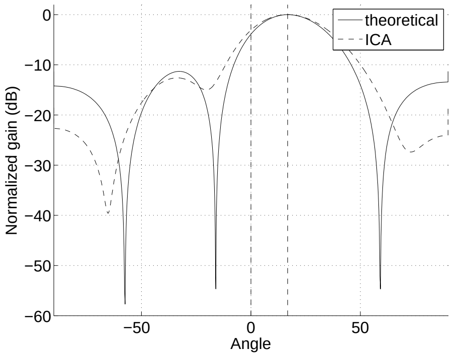


Figure 5: Wavenumber responses of the theoretical steering vector (solid) and the steering vector found by ICA (dashed) for source 2 ($f = 500$ Hz).

tion of source 2, and that a zero is placed in the direction of source 1. The resulting average signal-to-interference ratio was approximately 15 dB for the two sources.

4. DISCUSSION

The success of source separation with ICA on spatially resampled observations demands that no significant correlations is introduced between the original independent sources during resampling. The temporal frequency contents of the original signals are affected by the transformation, depending on resampling frequency and array geometry. Closely spaced sources may be correlated after the transformation and thereby indistinguishable with ICA.

The present method finds unmixing filters in the spatial domain, not exploiting temporal correlations in the observations. As a consequence, echos have to be treated as new "independent" sources, and we are limited by the number of zeros that can be forced in the wavenumber response. In the echoic scenario we need as many sensors as there are sources and echos for perfect separation.

5. CONCLUSION

We have presented a new method for blind source separation of convolutive mixtures based on independent component analysis. The method treats all frequency bands simultaneously and thereby avoids the permutation problem of frequency domain methods. We have demonstrated the method on experimental data from the anechoic cocktail-party problem, and shown good separation performance.

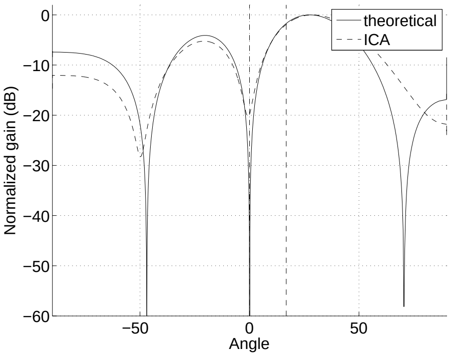


Figure 6: Wavenumber responses of the unmixing filters for source 2, theoretical (solid) and ICA (dashed). The dashed vertical lines indicate the propagation angles of each source.

REFERENCES

[1] S.T.Rickard P.D. O'Grady, B.A. Pearlmutter. Survey of sparse and non-sparse methods in source separation. *International Journal of Imaging Systems and Technology*, Special Issue on Blind Source Separation and Deconvolution in Imaging and Image Processing:18–33, July 2005.

[2] Nikolaos Mitianoudis and Mike E. Davies. Audio source separation: Solutions and problems. *International Journal of Adaptive Control and Signal Processing*, 18:299–314, 2004.

[3] Hiroshi Sawada, Ryo Mukai, Shoko Araki, and Shoji Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 12:530–538, September 2004.

[4] Shoko Araki, Ryo Mukai, Shoji Makino, Tsuyoki Nishikawa, and Hiroshi Saruwatari. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing*, 11(2):109–116, March 2003.

[5] S.C.Douglas and A.Cichocki. Convergence analysis of local algorithms for blind decorrelation. *NIPS96 Workshop, Blind Signal Processing and Their Applications*, 1996.

[6] F. Yin T.Mei, J.Xi and Z.Yang. A half-frequency domain approach for convolutive source separation based on the kullback-leibler divergence. *Eight International Symposium on Signal Processing and its Applications (ISSPA-2005)*, 1:25–28, August 2005.

[7] B.Barkat and K.Abed-Meraim. Algorithms for blind components separation and extraction from the time-frequency distribution of their mixture. *EURASIP Journal on Applied Signal Processing*, 13:2025–2033, 2004.

[8] L.C.Parra and C.V. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362, September 2002.

[9] Justinian Rosca, NingPing Fan, and Radu Balan. Real-time audio source separation by delay and attenuation compensation in the time domain. *Proc. of the 3rd ICA and BSS Conference, San Diego, CA*, December 2001.

[10] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, July 1996.

[11] Jeffrey Krolik and David Swingler. The performance of minimax spatial resampling filters for focusing wide-band arrays. *IEEE Transactions on Signal Processing*, 39(8):1899–1903, August 1991.

[12] Jean-François Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1):161–164, January 1996.