Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 38412, Pages 1–11
DOI 10.1155/ASP/2006/38412

Speech Source Separation in Convolutive Environments Using Space-Time-Frequency Analysis

Shlomo Dubnov,[1] Joseph Tabrikian,[2] and Miki Arnon-Targan[2]

[1] *CALIT 2, University of California, San Diego, CA 92093, USA*
[2] *Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel*

Received 10 February 2005; Revised 28 September 2005; Accepted 4 October 2005

We propose a new method for speech source separation that is based on directionally-disjoint estimation of the transfer functions between microphones and sources at different frequencies and at multiple times. The spatial transfer functions are estimated from eigenvectors of the microphones' correlation matrix. Smoothing and association of transfer function parameters across different frequencies are performed by simultaneous extended Kalman filtering of the amplitude and phase estimates. This approach allows transfer function estimation even if the number of sources is greater than the number of microphones, and it can operate for both wideband and narrowband sources. The performance of the proposed method was studied via simulations and the results show good performance.

Copyright © 2006 Shlomo Dubnov et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Many audio communication and entertainment applications deal with acoustic signals that contain combinations of several acoustic sources in a mixture that overlaps in time and frequency. In the recent years, there has been a growing interest in methods that are capable of separating audio signals from microphone arrays using blind source separation (BSS) techniques [1]. In contrast to most of the research works in BSS that assume multiple microphones, the audio data in most practical situations is limited to stereo recordings. Moreover, the majority of the potential applications of BSS in the audio realm consider separation of simultaneous audio sources in reverberant or echo environments, such as a room or inside a vehicle. These applications deal with convolutive mixtures [2] that often contain long impulse responses that are difficult to estimate or invert.

In this paper, we consider a simpler but still practical and largely overlooked situation of mixtures that contain a combination of source signals in weak reverberation environments, such as speech or music recorded with close microphones. The main mixing effect in such a case is direct path delay and possibly a small combination of multipath delays that can be described by convolution with a relatively short impulse response. Recently, several works proposed separation of multiple signals when additional assumptions are imposed on the signals in the time-frequency (TF) domain. In [3, 4] an assumption that each source occupies separate regions in short-time Fourier transform (STFT) representation using an analysis window $W(t)$ (so-called W-disjoint assumption) was considered. In [5] a source separation method is proposed using so-called single-source autoterms of a spatial ambiguity function. In the W-disjoint case the amplitude and delay estimation of the mixing parameters of each source is performed based on the ratio of the STFTs of signals between the two microphones. Since the disjoint assumption appears to be too strict for many real-world situations, several improvements have been reported that only allow an approximate disjoint situation [6]. The basic idea in such a case is to use some sort of a detection function that allows one to determine the TF areas where each source is present alone (we will refer to such an area as a *single-source TF cell*, or *single-TF* for short) and use only these areas for separation. Detection of single-source autoterms is based on detecting points that have only one nonzero diagonal entry in the spatial time-frequency distribution (STFD). The STFD generalizes the TF distribution for the case of vector signals. It can be shown that under a linear data model, the spatial TF distribution has a structure similar to that of the spatial correlation matrix that is usually used in array signal processing. The benefits of the spatial TF methods is that they directly exploit the nonstationary

property of the signals for purposes of detecting and separating the individual sources. Recent reported results of BSS using various single-TF detection functions show excellent performance for instantaneous mixtures.

In this paper, we propose a new method for source separation in the echoic or slightly reverberant case that is based on estimating and clustering the spatial signatures (transfer functions) between the microphones and the sources at different frequencies and at multiple times. The transfer functions for each source-microphone pair are derived from eigenvectors of correlation matrices between the microphone signals at each frequency, and are determined through a selection and clustering process that creates disjoint sets of eigenvector candidates for every frequency at multiple times. This requires solving the permutation problem [7], that is, association of transfer function values across different frequencies into a single transfer function. Smoothing and association are achieved by simultaneous Kalman filtering of the noisy amplitude and phase estimates along different frequencies for each source. This differs from association methods that assume smoothness of spectra of the separated signals, rather than smoothness of the transfer functions. Even when notches in room response occur due to signal reflections, these are relatively rare compared to the inherent sparseness of the source signals, which is inherent in the W-disjoint assumption.

Our approach allows estimation of the transfer functions between each source and every microphone, and is capable of operating for both wideband and narrowband sources. The proposed method can be used for approximate signal separation in undercomplete cases (more than two sources in a stereo recording) using filtering or time-frequency masking [8], in a manner similar to that of the W-disjoint situation.

This paper is structured in the following manner: in the next section, we review some recent state-of-the-art algorithms for BSS, specifically considering the nonstationary methods of independent component analysis (ICA) and the W-disjoint approaches. Section 3 presents our model and the details of the proposed algorithm. Specifically, we will describe the TF analysis and representation and its associated eigenvector analysis of the correlation matrices at different frequencies and multiple times. Then, we proceed to derive a criterion for identification of the single-source TF cells and clustering the spatial transfer functions. Details of the extended Kalman filter (EKF) tracking, smoothing, and across-frequency association of the transfer function amplitudes and phases conclude this section. The performance of the proposed method for source separation is demonstrated in Section 5. Finally, our conclusions are presented in Section 6.

2. BACKGROUND

The problem of multiple-acoustic-source separation using multiple microphones has been intensively investigated during the last decade, mostly based on independent component analysis (ICA) methods. These methods, largely driven by advances in machine learning research, treat the separation issue broadly as a density estimation problem. A common assumption in ICA-based methods is that the sources have a particular statistical behavior, such that the sources are random stationary statistically independent signals. Using this assumption, ICA attempts to linearly recombine the measured signals so as to achieve output signals that are as independent as possible.

The acoustic mixing problem can be described by the equation

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

where $\mathbf{s}(t) \in \mathbf{R}^M$ denotes the vector of M source signals, $\mathbf{x}(t) \in \mathbf{R}^N$ denotes the vector of N microphone signals, and $\mathbf{A}$ stands for the mixing matrix with constant coefficients A_{nm} describing the amplitude scaling between source m and microphone n. Naturally, this formulation describes only an instantaneous mixture with no delays or convolution effects. In a multipath environment, each source m couples with sensor n through a linear time-invariant system. Using discrete time t and τ, and assuming impulse responses not exceeding length L, the microphone signals are

$$x_n(t) = \sum_{m=1}^{M} \sum_{\tau=1}^{L} A_{nm}(\tau) s_m(t-\tau). \tag{2}$$

Note that the mixing is now a matrix convolution between the source signals and the microphones, where $A_{nm}(\cdot)$ represents the impulse response between source n and microphone m. We can rewrite this equation by applying the discrete Fourier transform (DFT):

$$\widetilde{\mathbf{x}}(\omega) = \widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{s}}(\omega), \tag{3}$$

where $\widetilde{\ }$ denotes the DFT of the signal. This notation assumes that either the signals and the mixing impulse responses are of short duration (shorter than the DFT length), or that an overlap-add formulation of the convolution process is assumed, which allows infinite duration for $\mathbf{s}(t)$ and $\mathbf{x}(t)$, but requires a short duration of the $A_{nm}(\cdot)$ responses. From now on we will consider the convolutive problem by assuming separate instantaneous mixing problems $\widetilde{\mathbf{x}}(\omega) = \widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{s}}(\omega)$ at every frequency ω. The aim of the convolutive BSS is to find filters $W_{mn}(t)$ that when applied to $\mathbf{x}(t)$ result in new signals $\mathbf{y}(t)$ that are approximately independent. In the frequency-domain formulation we have

$$\widetilde{\mathbf{y}}(\omega) = \widetilde{\mathbf{W}}(\omega)\widetilde{\mathbf{x}}(\omega), \tag{4}$$

so that $\mathbf{y}(t)$ corresponds to the original sources $\mathbf{s}(t)$, up to some allowed transformation such as permutation, that is, not knowing which source $s_m(t)$ appears in which output $y_{m'}(t)$, and amplitude scaling (relative volume).

This problem can be reformulated in statistical terms as follows: for each frequency, given a multivariate distribution

of vectors $\widetilde{\mathbf{x}} = (\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_N)^T$, whose coordinates or components correspond to the signals at the N microphones, we seek to find a matrix $\widetilde{\mathbf{W}}$ and vector $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_M)^T$, whose components are "as independent as possible." Saying so, it is assumed that there exists a multivariate process with independent components $\mathbf{s}$, which correspond to the actual independent acoustic sources, such as speakers or musical instruments, and a matrix $\widetilde{\mathbf{A}} = \widetilde{\mathbf{W}}^{-1}$ that corresponds to the mixing condition (up to permutation and scaling), so that $\widetilde{\mathbf{x}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{s}}$. Note that here and in the following we will at times drop the frequency parameter ω from the problem formulation.

Since the problem consists of finding an inverse matrix to the model $\widetilde{\mathbf{x}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{s}}$, any solution of this problem is possible only by using some prior information of $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{s}}$. Considering a pairwise independence assumption, the relevant criterion can be described by considering the following:

$$\forall t, k, l, \tau, i \neq j : E\left[s_i^k(t)s_j^l(t+\tau)\right] = E\left[s_i^k(t)\right]E\left[s_j^m(t+\tau)\right]. \tag{5}$$

The parameterization of different ICA approaches can be written now as different conditions on the parameters of the independence assumption. For stationary signals, the time indices are irrelevant and higher-order statistical criteria in the form of independence conditions with $k, l > 1$ must be considered. For stationary colored signals, it has been shown that decorrelation of multiple times t for $k = l = 1$ allows recovery of the sources in the case of an instantaneous mixture, but is insufficient for the general convolutive case. For nonstationary signals, decorrelation at multiple times, t, can be used (for $k = l = 1$) to perform the separation.

The idea behind decorrelation at multiple times t is basically an extension of decorrelation at two time instances. In the case of nonmoving sources and microphones, the same linear model is assumed to be valid at different time instances with different signal statistics, with the same orthogonal separating matrix $\mathbf{W}$:

$$\mathbf{W}\mathbf{x}(t_j, \omega) = \mathbf{y}(t_j, \omega), \quad j = 1, \ldots, J, \tag{6}$$

where the additional index ω of $\mathbf{W}$ implies that we are dealing with multiple separation problems for different values of ω. The same formulation can be used without ω for a time-domain problem, which gives a solution to the instantaneous mixture problem. Considering autocorrelation statistics at time instances $t_1, \ldots, t_J$ we obtain J sets of matrix equations:

$$\mathbf{R}_{x,t_j} = \mathbf{W}^{-1}\Lambda_{y,t_j}\mathbf{W}^{-T}, \quad j = 1, \ldots, J, \tag{7}$$

where we assume that $\{\Lambda_{y,t_j}\}_{j=1}^J$ are diagonal since the components of $\mathbf{y}$ are independent. This problem can be solved using a simultaneous diagonalization of $\{\mathbf{R}_{x,t_j}\}_{j=1}^J$, without knowledge of the true covariances of $\mathbf{y}$ at different times. A crucial point in implementation of this method is that it works only when the eigenvalues of the matrices $\mathbf{R}_{x,t}$ are all distinct. This case corresponds in physical reality to sufficiently unequal powers of signals arriving from different directions, a situation that is likely to be violated in practical scenarios. Moreover, since the covariance matrices are estimated in practice from short time frames, the averaging time needs to correspond to the stationarity time. An additional difficulty occurs specifically for the TF representation: independence between two signals in a certain band around ω corresponds to independence between narrowband processes, which can be revealed at time scales that are significantly longer than the window size or the effective impulse response of the bandpass filter used for TF analysis. This inherently limits the possibility of averaging (taking multiple frames or snapshots of the signal in one time segment) without exceeding the stationarity interval of the signal. In the following we will show how our method solves the eigenvalue indeterminacy problem by choosing those time segments where only one significant eigenvalue occurs. Our "segmental" approach actually reduces the generalized (or multiple) eigenvalue problem into a set of disjoint eigenvalue problems that are solved separately for each source. The details of our algorithm will be described in the next section. In the following, we will consider the "directionally-disjoint" sources case in which the local covariance matrices $\mathbf{R}_{x,t_j}$ have a single large eigenvalue at sufficiently many time instances t_j. The precise definition and the amount of times that are sufficient for separation will be discussed later.

3. PROPOSED SOURCE SEPARATION METHOD

Consider an N-channel sensor signal $\mathbf{x}(t)$ that arises from M unknown scalar source signals $s_m(t)$, corrupted by zero-mean, white Gaussian additive noise. In a convolutive environment, the signals are received by the array after delays and reflections. We consider the case where each one of the sources has a different spatial transfer function. Therefore, the signal at the nth microphone is given by

$$x_n(t) = \sum_{m=1}^{M}\sum_{l=1}^{L} a_{nml}s_m(t - \tau_{nml}) + v_n(t), \quad n = 1, \ldots, N, \tag{8}$$

in which τ_{nml} and a_{nml} are the delay and gain of the lth path between source signal m and microphone n, and $v_n(t)$ denotes the zero-mean white Gaussian noise. The STFT of (8) gives

$$X_n(t, \omega) = \sum_{m=1}^{M} A_{nm}(\omega)S_m(t, \omega) + V_n(t, \omega), \quad n = 1, \ldots, N, \tag{9}$$

where $S_m(t, \omega)$ and $V_n(t, \omega)$ are the STFT of $s_m(t)$ and $v_n(t)$, respectively, and the transfer function between the mth signal to the nth sensor is defined as

$$A_{nm}(\omega) = \sum_{l=1}^{L} a_{nml}e^{-j\omega\tau_{nml}}. \tag{10}$$

In matrix notation, the model (9) can be written in the form

$$\mathbf{X}(t,\omega) = \mathbf{A}(\omega)\mathbf{S}(t,\omega) + \mathbf{V}(t,\omega). \tag{11}$$

Our goal here is to estimate the spatial transfer function matrix, $\mathbf{A}(\omega)$, and the signal vector, $\mathbf{s}(t)$, from the measurement vector $\mathbf{x}(t)$. For estimation of the signal vector, we will assume that the number of sources, M, is not greater than the number of sensors, N. This assumption is not required for estimation of the spatial transfer function matrix, $\mathbf{A}(\omega)$.

The proposed approach seeks time-frequency cells in which only one source is present. At these cells, it is possible to estimate the unstructured spatial transfer function matrix for the present source. Therefore, we will first identify the single-source TF cells and calculate the spatial transfer functions for the sources present in those cells. In the second stage, the spatial transfer functions are clustered using a Gaussian mixture model (GMM). The frequency-permutation problem is solved by considering the spatial transfer functions as a frequency-domain Markov model and applying an EKF to track it. Finally, the sources are separated by inverse filtering of the measurements using the estimated transfer function matrices.

The autocorrelation matrix at a given time-frequency cell is given by

$$\begin{aligned}\mathbf{R}_{\mathbf{x}}(t,\omega) &= E\left[\mathbf{X}(t,\omega)\mathbf{X}^H(t,\omega)\right] \\ &= \mathbf{A}(\omega)\mathbf{R}_{\mathbf{s}}(t,\omega)\mathbf{A}^H(\omega) + \mathbf{R}_{\mathbf{v}}(t,\omega),\end{aligned} \tag{12}$$

where $\mathbf{R}_{\mathbf{x}}$, $\mathbf{R}_{\mathbf{s}}$, and $\mathbf{R}_{\mathbf{v}}$ are the time-frequency spectra of the measurements, source signals, and sensor noises, respectively. We assume that the noise is stationary, and therefore its covariance matrix is independent of time t, that is, $\mathbf{R}_{\mathbf{v}}(t,\omega) = \mathbf{R}_{\mathbf{v}}(\omega)$. Furthermore, the noise spectrum is usually known, so (12) can be spatially prewhitened by left multiplying (11) by $\mathbf{R}_{\mathbf{v}}^{-1/2}(\omega)$. Thus, we can assume $\mathbf{R}_{\mathbf{v}}(\omega) = \sigma_v^2\mathbf{I}_N$ for all ω where $\mathbf{I}_N$ is the identity matrix of size N.

3.1. Identification of single-source TF cells

Each time-frequency window is tested in order to identify the time-frequency windows in which a single signal is present. In these cells, the unstructured spatial transfer function can be easily estimated. Consider a time segment consisting of T time cells in which the signals are stationary. Then, (12) becomes time-independent:

$$\mathbf{R}_{\mathbf{x}}(\omega) = \mathbf{A}(\omega)\mathbf{R}_{\mathbf{s}}(\omega)\mathbf{A}^H(\omega) + \sigma_v^2\mathbf{I}_N. \tag{13}$$

If only the mth source is present, (13) becomes

$$\mathbf{R}_{\mathbf{x}m}(\omega) = \mathbf{a}_m(\omega)\mathbf{a}_m^H(\omega)\sigma_{s_m}^2(\omega) + \sigma_v^2\mathbf{I}_N, \tag{14}$$

where $\mathbf{a}_m(\omega)$ is the mth column of the matrix $\mathbf{A}(\omega)$ and $\sigma_{s_m}^2(\omega)$ denotes the mth signal power spectrum. In this case, the rank of the (noiseless) signal covariance matrix is 1 and $\mathbf{a}_m(\omega)$ is proportional to the eigenvector of the autocorrelation matrix $\mathbf{R}_{\mathbf{x}m}(\omega)$ associated with the maximum eigenvalue: $\lambda_{1,m}(\omega) = \sigma_{s_m}^2(\omega)\|\mathbf{a}_m(\omega)\|^2 + \sigma_v^2$. This property allows us to derive a test for identification of the single-source segments and estimate the corresponding spatial transfer function $\mathbf{a}_m(\omega)$. We will denote the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathbf{R}_{\mathbf{x}}(\omega)$ by $\mathbf{u}(\omega)$, disregarding the source index m.

The three hypotheses for each time-frequency cell in a stationary segment, which indicate the number of active sources in this segment, are

$$\begin{aligned}H_0 &: \mathbf{X}(t,\omega) \sim N^{\mathrm{c}}(\mathbf{0}, \sigma_v^2\mathbf{I}_N), \\ H_1 &: \mathbf{X}(t,\omega) \sim N^{\mathrm{c}}\left[\mathbf{0}, \mathbf{u}(\omega)\mathbf{u}^H(\omega)\sigma_s^2(\omega) + \sigma_v^2\mathbf{I}_N\right], \\ H_2 &: \mathbf{X}(t,\omega) \sim N^{\mathrm{c}}\left[\mathbf{0}, \mathbf{R}_{\mathbf{x}}(\omega)\right],\end{aligned} \tag{15}$$

where H_0, H_1, H_2 indicate noise-only, single-source, and multiple-source hypotheses, respectively, with $\mathbf{X} \sim N^{\mathrm{c}}(\cdot,\cdot)$ denoting the complex Gaussian distribution. Under hypothesis H_0, the model parameters are known. Under hypothesis H_1, the vector $\mathbf{u}(\omega)$ is the normalized spatial transfer function of the present source in the segment (i.e., one of the columns of the matrix $\mathbf{A}(\omega)$) and $\sigma_s^2(\omega)$ represents the corresponding signal power spectrum. We assume that $\mathbf{u}(\omega)$ and $\sigma_s^2(\omega)$ are unknown. In hypothesis H_2, it is assumed that the data model is complex Gaussian-distributed and spatially colored with unknown covariance matrix $\mathbf{R}_{\mathbf{x}}(\omega)$, which represents the contribution of several mixed sources. Usually, the Gaussian distribution assumption for hypotheses H_1 and H_2 does not hold, and in fact leads to suboptimal solutions. However, this assumption enables obtaining a simple and meaningful result for source separation.

In order to identify the case of a single source, two tests are performed. In the first, the hypotheses H_0 and H_1 are tested, while in the second, hypotheses H_1 and H_2 are tested. A time-frequency cell is considered as a single-source cell if in both tests it is decided that a single source is present. These tests are carried out between hypotheses with unknown parameters, and therefore the generalized likelihood ratio test (GLRT) is employed, that is,

$$\begin{aligned}\mathrm{GLRT}_1 &= \max_{\mathbf{u},\sigma_s^2} \log f_{\mathbf{X}|H_1;\mathbf{u},\sigma_s^2} - \log f_{\mathbf{X}|H_0} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_1, \\ \mathrm{GLRT}_2 &= \max_{\mathbf{R}_{\mathbf{x}}} \log f_{\mathbf{X}|H_2;\mathbf{R}_{\mathbf{x}}} - \max_{\mathbf{u},\sigma_s^2} \log f_{\mathbf{X}|H_1;\mathbf{u},\sigma_s^2} \underset{H_1}{\overset{H_2}{\gtrless}} \gamma_2,\end{aligned} \tag{16}$$

where $f_{\mathbf{X}|H_0}$, $f_{\mathbf{X}|H_1;\mathbf{u},\sigma_s^2}$, and $f_{\mathbf{X}|H_2;\mathbf{R}_{\mathbf{x}}}$ denote the probability density functions (pdf's) of each time-frequency segment under the three hypotheses.

Now, we will derive the GLRTs for identification of single-source cells. Consider T independent samples of the data vectors $\mathbf{X}(\omega) \triangleq [\mathbf{X}(1,\omega),\ldots,\mathbf{X}(T,\omega)]$ for which the data vector is stationary. Then, under the three hypotheses described above, $\mathbf{X}(t,\omega)$ is complex Gaussian-distributed

$\mathbf{X}(t,\omega) \sim N^{c}[\mathbf{0}, \mathbf{R}_{\mathbf{x}}(\omega)]$. The model of $\mathbf{R}_{\mathbf{x}}(\omega)$ differs between the three hypotheses. The log-likelihood of the data $\mathbf{X}(\omega)$ under the joint model is

$$\begin{aligned}\log f_{\mathbf{X}|\mathbf{R}_{\mathbf{x}}} &= -T\log\left|\pi\mathbf{R}_{\mathbf{x}}(\omega)\right| - \sum_{t=1}^{T}\mathbf{X}^{H}(t,\omega)\mathbf{R}_{\mathbf{x}}^{-1}(\omega)\mathbf{X}(t,\omega)\\ &= -T\{\log\left|\pi\mathbf{R}_{\mathbf{x}}(\omega)\right| + \operatorname{tr}[\hat{\mathbf{R}}_{\mathbf{x}}(\omega)\mathbf{R}_{\mathbf{x}}^{-1}(\omega)]\},\end{aligned} \tag{17}$$

where $\hat{\mathbf{R}}_{\mathbf{x}}(\omega)$ is the sample covariance matrix $\hat{\mathbf{R}}_{\mathbf{x}}(\omega) \triangleq 1/T\sum_{t=1}^{T}\mathbf{X}(t,\omega)\mathbf{X}^{H}(t,\omega)$. For simplicity of notation, we will drop the dependence on frequency ω.

Under hypothesis H_0, $\mathbf{R}_{\mathbf{x}} = \sigma_v^2\mathbf{I}$, and therefore the log-likelihood from (17) becomes

$$\log f_{\mathbf{X}|H_0} = -T\left\{N\log\left(\pi\sigma_v^2\right) + \frac{1}{\sigma_v^2}\operatorname{tr}\left(\hat{\mathbf{R}}_{\mathbf{x}}\right)\right\}. \tag{18}$$

Under hypothesis H_1, $\mathbf{R}_{\mathbf{x}} = \sigma_s^2\mathbf{u}\mathbf{u}^H + \sigma_v^2\mathbf{I}_N$, for which the following equations are satisfied:

$$\begin{aligned}\mathbf{R}_{\mathbf{x}}^{-1} &= \frac{1}{\sigma_v^2}\left(\mathbf{I}_N - \frac{\mathrm{SNR}}{1+\mathrm{SNR}}\mathbf{u}\mathbf{u}^H\right),\\ \left|\mathbf{R}_{\mathbf{x}}\right| &= \sigma_v^{2N}(1+\mathrm{SNR}),\end{aligned} \tag{19}$$

where $\mathrm{SNR} \triangleq \sigma_s^2/\sigma_v^2$. Substitution of (19) into (17) yields

$$\begin{aligned}\log f_{\mathbf{X}|H_1,\mathbf{u},\sigma_s^2} &= -T\left\{\log\left[\left(\pi\sigma_v^2\right)^N(1+\mathrm{SNR})\right]\right.\\ &\qquad\left.+\frac{1}{\sigma_v^2}\operatorname{tr}\left[\hat{\mathbf{R}}_{\mathbf{x}}\left(\mathbf{I}_N - \frac{\mathrm{SNR}}{1+\mathrm{SNR}}\mathbf{u}\mathbf{u}^H\right)\right]\right\}\\ &= -T\left[N\log\left(\pi\sigma_v^2\right) + \frac{1}{\sigma_v^2}\operatorname{tr}\left(\hat{\mathbf{R}}_{\mathbf{x}}\right) + \log(1+\mathrm{SNR})\right.\\ &\qquad\left.-\frac{\mathrm{SNR}}{\sigma_v^2(1+\mathrm{SNR})}\mathbf{u}^H\hat{\mathbf{R}}_{\mathbf{x}}\mathbf{u}\right].\end{aligned} \tag{20}$$

Maximization of (20) with respect to σ_s^2 can be replaced by maximization with respect to SNR. This operation can be performed by calculating the derivative of (20) with respect to SNR and equating it to zero, resulting in $\widehat{\mathrm{SNR}}(\mathbf{u}) = \mathbf{u}^H\hat{\mathbf{R}}_{\mathbf{x}}\mathbf{u}/\sigma_v^2 - 1$ or $\hat{\sigma}_s^2(\mathbf{u}) = \mathbf{u}^H\hat{\mathbf{R}}_{\mathbf{x}}\mathbf{u} - \sigma_v^2$. Thus,

$$\begin{aligned}&\max_{\sigma_s^2}\log f_{\mathbf{X}|H_1,\mathbf{u},\sigma_s^2}\\ &\quad= -T\left[N\log\left(\pi\sigma_v^2\right) + \frac{1}{\sigma_v^2}\operatorname{tr}\left(\hat{\mathbf{R}}_{\mathbf{x}}\right) + 1 + \log\eta - \eta\right],\end{aligned} \tag{21}$$

where $\eta \triangleq \mathbf{u}^H\hat{\mathbf{R}}_{\mathbf{x}}\mathbf{u}/\sigma_v^2$. We seek to maximize (21) with respect to $\mathbf{u}$, where $\mathbf{u}$ is constrained to unity norm. Since (21) is monotonically increasing with η, for $\eta > 1$, then the log-likelihood is maximized when η is maximized. Let $\lambda_1 \geq \cdots \geq \lambda_N$ denote the eigenvalues of $\hat{\mathbf{R}}_{\mathbf{x}}$. Then, $\max_{\mathbf{u}}\mathbf{u}^H\hat{\mathbf{R}}_{\mathbf{x}}\mathbf{u} = \lambda_1$, and

$$\begin{aligned}\max_{\mathbf{u},\sigma_s^2}\log f_{\mathbf{X}|H_1,\mathbf{u},\sigma_s^2} &= -T\left[N\log\left(\pi\sigma_v^2\right) + 1 + \frac{1}{\sigma_v^2}\operatorname{tr}\left(\hat{\mathbf{R}}_{\mathbf{x}}\right)\right.\\ &\qquad\left.+\log\frac{\lambda_1}{\sigma_v^2} - \frac{\lambda_1}{\sigma_v^2}\right]\\ &= -T\left[N\log\left(\pi\sigma_v^2\right) + 1 + \sum_{i=2}^{N}\frac{\lambda_i}{\sigma_v^2} + \log\frac{\lambda_1}{\sigma_v^2}\right].\end{aligned} \tag{22}$$

Under hypothesis H_2, the matrix $\mathbf{R}_{\mathbf{x}}$ is unstructured and assumed to be unknown. Equation (17) is maximized for $\mathbf{R}_{\mathbf{x}} = \hat{\mathbf{R}}_{\mathbf{x}}$ [9]. The resulting log-likelihood under this hypothesis is

$$\begin{aligned}\max_{\mathbf{R}_{\mathbf{x}}}\log f_{\mathbf{X}|H_2,\mathbf{R}_{\mathbf{x}}} &= -T\left(\log\left|\pi\hat{\mathbf{R}}_{\mathbf{x}}\right| + N\right)\\ &= -T\left(N\log\pi + \sum_{i=1}^{N}\log\lambda_i + N\right).\end{aligned} \tag{23}$$

Now, the two GLRTs for decision between (H_0, H_1) and (H_1, H_2) can be derived by subtracting the corresponding log-likelihood functions:

$$\begin{aligned}\mathrm{GLRT}_1 &= \max_{\mathbf{u},\sigma_s^2}\log f_{\mathbf{X}|H_1;\mathbf{u},\sigma_s^2} - \log f_{\mathbf{X}|H_0} = T\left(\frac{\lambda_1}{\sigma_v^2} - \log\frac{\lambda_1}{\sigma_v^2} - 1\right) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_1',\\ \mathrm{GLRT}_2 &= \max_{\mathbf{R}_{\mathbf{x}}}\log f_{\mathbf{X}|H_2;\mathbf{R}_{\mathbf{x}}} - \max_{\mathbf{u},\sigma_s^2}\log f_{\mathbf{X}|H_1;\mathbf{u},\sigma_s^2} = T\left[\sum_{i=2}^{N}\left(\frac{\lambda_i}{\sigma_v^2} - \log\frac{\lambda_i}{\sigma_v^2}\right) - N + 1\right] \underset{H_1}{\overset{H_2}{\gtrless}} \gamma_2'.\end{aligned} \tag{24}$$

Finally, after dropping the constants, and modifying the thresholds accordingly, the two tests can be stated as

$$T_1 = \left(\frac{\lambda_1}{\sigma_v^2} - \log\frac{\lambda_1}{\sigma_v^2}\right) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_1,$$
$$T_2 = \sum_{i=2}^{N}\left(\frac{\lambda_i}{\sigma_v^2} - \log\frac{\lambda_i}{\sigma_v^2}\right) \underset{H_1}{\overset{H_2}{\gtrless}} \gamma_2. \tag{25}$$

The thresholds γ_1 and γ_2 in the two tests should be set according to the following considerations. Large values for γ_1 and small values for γ_2 will lead to missed detections of single-source TF cells, and therefore lead to a lack of data for calculation of the spatial transfer function. On the other hand, small values for γ_1 or large values for γ_2 will lead to false detections of single-source TF cells, which can cause erroneous estimation of the spatial transfer function. Generally, larger amounts of data will enable us to increase γ_1 and decrease γ_2.

In the case of stereo signals ($N = 2$), both tests could be expressed for $i = 1, 2$ and $\lambda_2 \geq \lambda_1 \geq \sigma_v^2$ as

$$T_i = \left(\frac{\lambda_i}{\sigma_v^2} - \log\frac{\lambda_i}{\sigma_v^2}\right) \underset{H_{i-1}}{\overset{H_i}{\gtrless}} \gamma_i. \tag{26}$$

3.2. Spatial transfer function estimation

In the TF cells that are identified to be single-source cells, the ML estimator for the normalized spatial transfer function of the present source at the given frequency ω is given by the eigenvector associated with the maximum eigenvalue of the autocorrelation matrix $\mathbf{R}_{\mathbf{x}m}$. It is important to note that a single amplitude-delay pair is sufficient to describe the spatial transform for a sufficiently narrow frequency band representation and assuming a linear spatial system. We can rewrite the model (11) for the case of two sources and two microphones as

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ a_1 e^{-j\omega\delta_1} & a_2 e^{-j\omega\delta_2} \end{bmatrix} \begin{bmatrix} S_1(\omega) \\ S_2(\omega) \end{bmatrix} \tag{27}$$

in which case, the mixing matrix column, corresponding to one of the sources, say source m, can be directly estimated from the eigenvector, $\mathbf{a}_m(\omega)$, associated with the maximum eigenvalue of the autocorrelation matrix $\mathbf{R}_{\mathbf{x}m}$ under hypothesis T_1, that is, a single-source m is present in this TF region. Thus,

$$a_m e^{-j\omega\delta_m} = \frac{a_{m,2}(\omega)}{a_{m,1}(\omega)}, \tag{28}$$

where $a_{m,i}$ denotes the ith component of $\mathbf{a}_m$, or more specifically

$$a_m = \left|\frac{a_{m,2}(\omega)}{a_{m,1}(\omega)}\right|,$$
$$\delta_m = \frac{1}{\omega}\Im\left[\log\frac{a_{m,2}(\omega)}{a_{m,1}(\omega)}\right], \tag{29}$$

where $\Im$ denotes taking the imaginary part.

Having different amplitude and delay values for each source at every frequency, we need to associate the different amplitude and delay values across frequency to their corresponding source. If we assume that the amplitude and delay are constant over different frequencies, occurring in the case of a direct path effect only, the association can be performed by clustering the amplitude and phase values around their mean value. In the case of multipath, the amplitude and delay values may differ across frequencies. Using smoothness considerations, one could try to associate the parameters across different frequencies by assuming proximity of parameter values across frequency bins for the same source. It should be also noted that smoothness of delay values requires unwrapping of the complex logarithm before dividing by ω. This is limited by spatial aliasing for high frequencies, that is, if the spacing d between the sensors is too large, the delay d/c where c is the speed of sound, might be larger than the maximum permissible delay $2\pi/\omega_s$, with ω_s denoting the sampling frequency. In other words, it might not be possible to uniquely solve the permutation problem if the delay between two microphones is more than one sample. Moreover, separate clustering or associating amplitude and delay parameters also looses information about the relations between the real and imaginary components of the spatial transfer function vector. In the following section, we will describe an optimal tracking and frequency association based on Kalman modeling, which addresses these problems assuming smoothness of the amplitude and phase of the spatial transfer function across frequency.

4. TRACKING AND FREQUENCY ASSOCIATION ALGORITHM

A common problem in frequency-domain convolutive BSS is that the mixing parameter estimation is performed separately for each frequency. In order to reconstruct the time signal, the frequency-separated channels must be combined together in a consistent manner, that is, one must insure that the different frequency components correspond to the same source. This problem is sometimes referred to as the frequency-permutation or association problem. In our method we perform the association in two steps. First, we reduce the number of points at every frequency by finding clusters of the points $a_{m,2}(\omega)/a_{m,1}(\omega)$ in the complex plane at different time segments. This clustering is performed using a two-dimensional GMM of the real and imaginary parts. The number of clusters is determined *a priori* according to the number of sources. When the number of sources is unknown, additional methods for determining the number of clusters may be considered. Next, association of the mixing

parameters across frequency is performed by operating separate EKFs on the cluster means, one for each source.

4.1. Gaussian mixture model and extended Kalman filter

The GMM assumes that the observations $\mathbf{z}$ are distributed according to the following density function

$$p_{\mathbf{z}}(\mathbf{z}) = \sum_{m=1}^{M} \pi_m N(\mathbf{z} \mid \Theta_m), \tag{30}$$

where π_m are the weights of the Gaussian distribution $N(\cdot \mid \Theta_m)$, and $\Theta_m = \{\mu_m, \Sigma_m\}$ are its mean and covariance matrix parameters, respectively. In our case, the observations, $\mathbf{z}$, are estimates of the real and imaginary parts of the transfer function over frequency (see previous section). The parameters of the GMM are obtained using an expectation-maximization (EM) procedure. The estimated mean and covariance matrix at each frequency are used for tracking the spatial transfer function.

An EKF is used for tracking and association of the transfer functions, whose mean and variance are estimated by the EM algorithm. The idea here is that the spatial transfer function between each source and microphone is smooth over frequency. Notches that occur in the transfer function due to signal reflections will be smoothed by the EKF, causing errors in the estimation (29), which color the signal but do not interfere with the association process since one of the sources in this case has small or zero amplitude. Therefore, the spatial transfer functions are modeled as first-order Markov sequences. It is natural to use the magnitude and phase of each spatial transfer function for the state vector, because in simple scenarios with no multipath, the absolute value of the transfer function is constant over frequency, while its phase linearly varies with frequency. Thus, the state vector of each EKF includes the magnitude (ρ), phase (α), and phase rate ($\dot{\alpha}$) of the transfer function. The presence of multipath causes deviation from this model, which can be represented by a noise model. Thus, the state vector dynamics across neighboring frequencies (frequency smoothness constraint) are modeled as

$$\begin{aligned} \boldsymbol{\phi}_k &= \begin{pmatrix} \rho_k \\ \alpha_k \\ \dot{\alpha}_k \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \rho_{k-1} \\ \alpha_{k-1} \\ \dot{\alpha}_{k-1} \end{pmatrix} + \mathbf{n}_{\phi k}, \\ \boldsymbol{\mu}_k &= \begin{pmatrix} \Re(\mathbf{a}_m(\omega_k)) \\ \Im(\mathbf{a}_m(\omega_k)) \end{pmatrix} = \begin{pmatrix} \rho_k \cos\alpha_k \\ \rho_k \sin\alpha_k \end{pmatrix} + \mathbf{n}_{\mu k}, \end{aligned} \tag{31}$$

in which the noise covariance of $\mathbf{n}_{\mu k}$ is taken from the abovementioned clustering algorithm, and the model noise covariance of $\mathbf{n}_{\phi k}$ is set according to the expected dynamics of the spatial transfer function.

For tracking the M transfer functions, M independent EKFs are implemented in parallel. At each frequency step, the data is associated with the EKFs according to the criterion of minimum-norm distance between the clustering estimates and the M Kalman predictions.

4.2. The separation algorithm

The various steps of the algorithm can be summarized as follows.

(i) Given a two-channel recording, perform a separate STFT analysis for every channel, resulting in the signal model (11).

(ii) Perform an eigenvalue analysis of the cross-channel correlation matrix at each frequency, as described in Section 3, where (12) and (26) determine the transfer function.

(iii) At each frequency, determine the cluster centers of the set of amplitude ratio measurements using the GMM.

(iv) Perform EKF tracking of the cluster means across frequency for each source to obtain an estimate of the mixing matrix as a function of frequency.

(v) If the mixing matrix is invertible, recover the signals by multiplying the STFT channels at each frequency by the inverse of the estimated mixing matrix. In case of more microphones than sources, the pseudoinverse of the mixing matrix should be used. In case of more sources than microphones, source separation can be approximately performed using time-frequency masking method of [8].

(vi) Perform an inverse STFT using the associated frequencies for each of the sources.

Since the mixing matrix can be determined only up to a scaling factor, we assume a unit relative magnitude for one of the sources and use the amplitude ratios to determine the mixing parameters of the remaining source. This problem of scale invariance may cause a "coloration" of the recovered signal (over frequency) and is one of the possible sources of error, being common to most convolutional source separation methods. Another typical problem is that the narrowband processing corresponds to circular convolution rather than the desired linear convolution. This effectively restricts the length of the impulse response between the microphones to be less than half of the analysis window length, or in frequency it restricts the spectral smoothness to that of the DFT length. Since speech sources are sparse in frequency (at least for the voiced segments), it is assumed that spectral peaks of speech harmonics would not be seriously influenced by spectral detail smaller than one FFT bin.

5. EXPERIMENTAL RESULTS

Separation experiments were carried out for simulated mixing conditions. We tested the proposed algorithm under different conditions, such as relative amplitudes of the sources, angles and amplitudes of the multipath reflections, and different types of sound sources.

In the first experiment, two female speakers were recorded by two microphones with 4.5 cm spacing. Figure 1 shows the measured versus smoothed spatial transfer functions for this difficult case of two female speaker sources of 20-second length, sampled at a rate of 8 kHz, with nearly equal amplitude mixing conditions. The separation

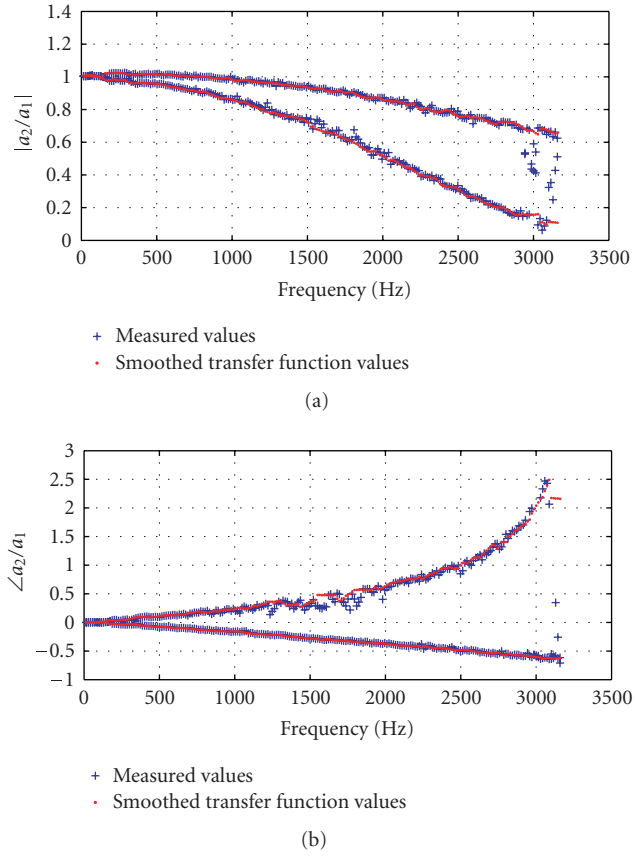

Figure 1: Amplitude and phase of two female speaker sources with nearly equal amplitude mixing conditions.

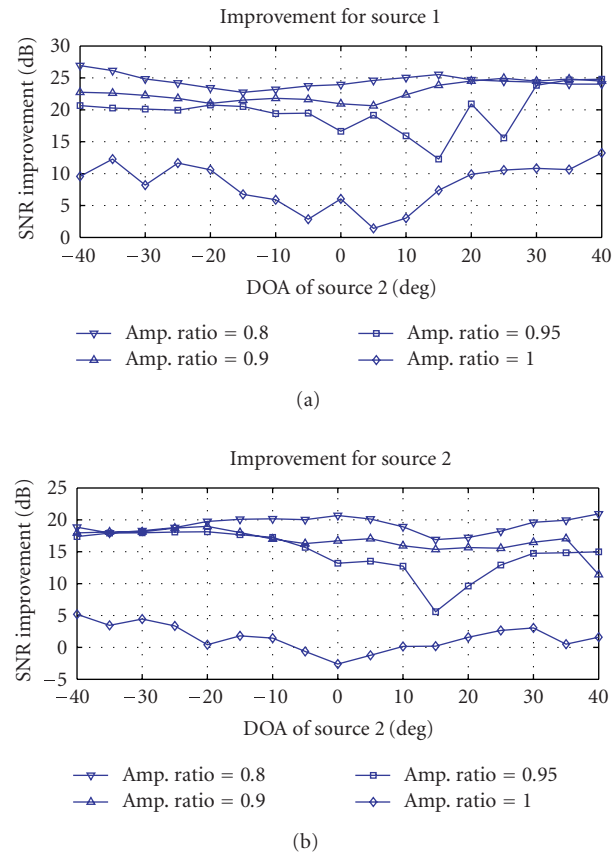

Figure 2: Improvement in SNR as a function of source angle for different relative amplitudes under weak multipath conditions.

is possible due to the different phase behavior of the signals, which is properly detected using the EKF tracking. The EKF parameters were set as follows. The system noise covariance matrix was set according to standard deviation (STD) of 0.1/sample in the transfer function amplitude and 0.1 rad/sample for phase. The measurement covariance matrices were set based on the results of the EM algorithm for GMM parameters estimation. The measurement STDs are in fact the widths of the Gaussians. The EKF parameters were also fixed in the following examples.

In Figure 2 the SNR improvement for different relative positions of the sources with different relative amplitudes is presented. The SNR improvement was calculated according to the method described in [10]. The separation quality of the mth source is evaluated by the ratio of the maximal energy output signal and sum of energies of the remaining output signals when only source m is present at the input. One of the sources was fixed at 0° while the other source was shifted from −40° to 40°. The amplitude ratio of the sources at the microphones varied from 0.8 to equal amplitude ratios. The multipath reflections occurred at constant angles of 60° and −40° with relative amplitudes of a few percent of the original. For equal amplitudes, we achieve up to 10 dB of improvement when the sources are 40° apart. The angle sensitivity disappears when sufficient amplitude difference exists between the sources. For an amplitude ratio of 0.8 (i.e., each microphone receives its main source at amplitude 1 and the interfering source at amplitude 0.8), we achieved 20–30 dB improvement. One should note that the above results contain weak multipath components. Even better improvement (50 dB or more) can be achieved for cases when no multipath is present.

The performance of the proposed method was tested also under strong multipath conditions. In this test, the two microphones measured signals from two sources. Each source signal arrives at the microphones through six different paths. The paths of the first source are from 0°, −5°, −10°, −20°, −30°, −40°, with strengths 0, −6, −7.5, −9, −11, and −13.5 dB. The paths of the second source are from 60°, 50°, 40°, 30°, 20°, with strengths −7.5, −9, −11, −13.5, and −17 dB, where the main path was at 0 dB with varying direction. The relative amplitude of the received paths at the microphones was randomly chosen between 0.67–0.86. Figure 3 shows the SNR improvement for both sources as a function of the main path direction for different relative amplitudes.

The proposed method was also tested for separation of three sources (female speakers) using three microphones. Figure 4 shows the SNR improvement results with different relative amplitudes as a function of the third source

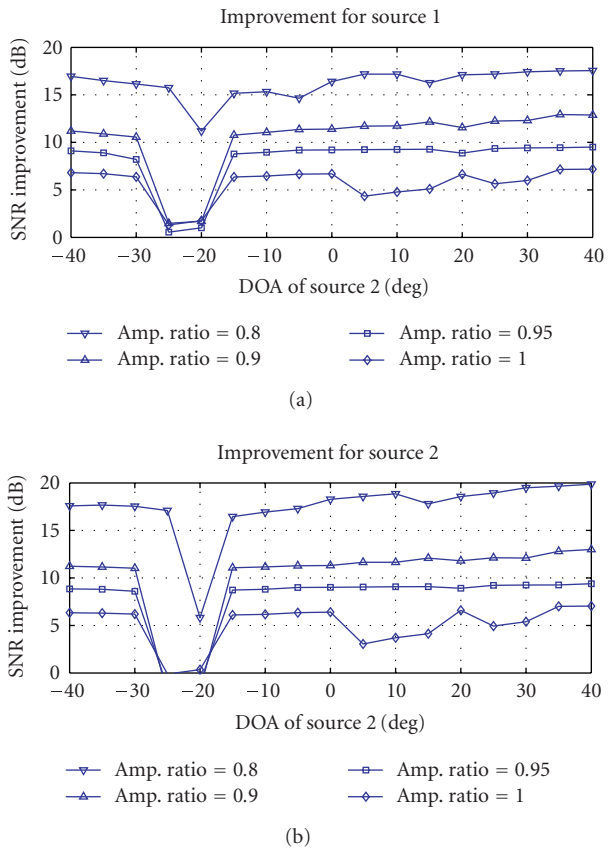

Figure 3: Improvement in SNR under strong multipath conditions as a function of source angle for different relative amplitudes.

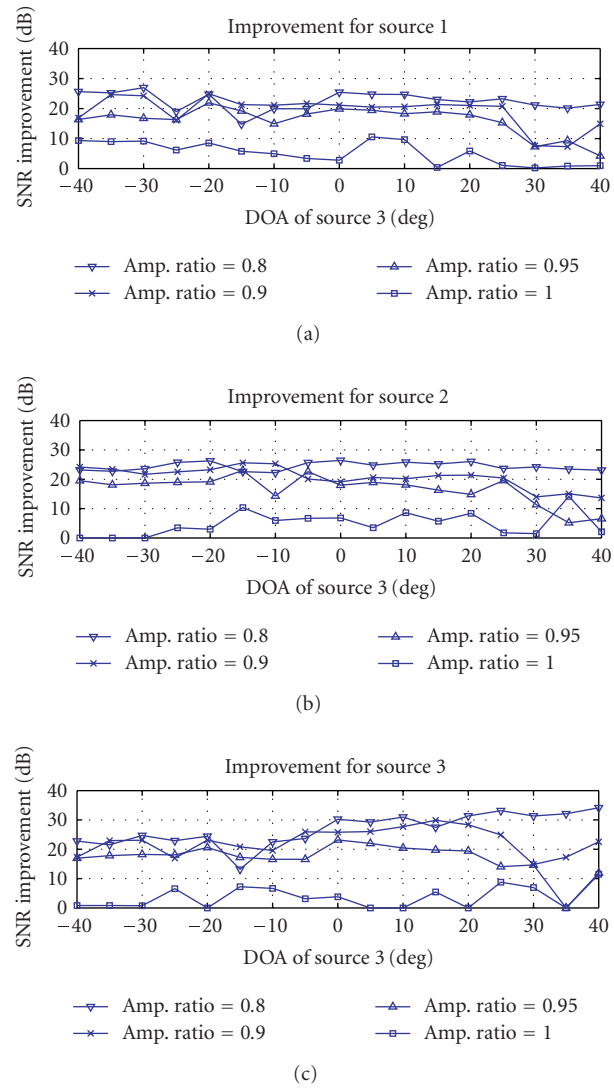

Figure 4: Improvement in SNR for the case of three microphones and three sources as a function of the third source angle for different relative amplitudes.

direction. The microphones were positioned within a linear, equally spaced (LES) array with 4.5 cm intersensor spacing. The performance in this case is slightly lower than the case of two microphones versus two sources, mainly because there are fewer TF cells in which a single source is present. Obviously, longer data can significantly improve the results in cases of multiple sources and multiple microphones.

As mentioned above, the proposed method is able to estimate the spatial transfer function in the case of more sources than sensors. Figure 5 shows the magnitude and phase of the true and estimated channel transfer functions of the three sources where only two microphones were used. The sources were located at -40°, -10°, and 30° with relative amplitudes of the different sources of 4, 2, and 0.5 between the microphones.

Figure 6 shows the amplitude of the spatial transfer function obtained by the inverse mixing matrix over frequency for the case of two sources located at 0° and 60°, without multipath. One can observe that the spatial pattern generated by the inverse of the estimated mixing matrix introduces a null in the direction of the interfering source. Figure 6(a) shows the null generated around 60° for recovering the source at 0°, while Figure 6(b) shows the null generated around 0° for recovering the source at 60°.

The proposed method for estimating the spatial transfer functions using the correlation matrix of the TF representation can be compared to the method for estimation of mixing and delay parameters from the STFT, as reported in [3, 8]. The basic assumption of that approach is the orthogonality of the "W-disjoint," which requires that part of TF the cells in the TF representation of the sources do not overlap. The derivation of the relative amplitude and delay parameters associated with source m being active at (t, ω) is done using

$$(a_m, \delta_m) = \left[\left| \frac{X_2(t,\omega)}{X_1(t,\omega)} \right|, \frac{1}{\omega} \angle \frac{X_2(t,\omega)}{X_1(t,\omega)} \right]. \tag{32}$$

Note that unlike the proposed method, in this case the mixing parameters are estimated directly from the STFT representation without taking into account the additive noise, which affects both amplitude and phase estimates. Using spatial correlation, it is possible to recover the relative amplitude

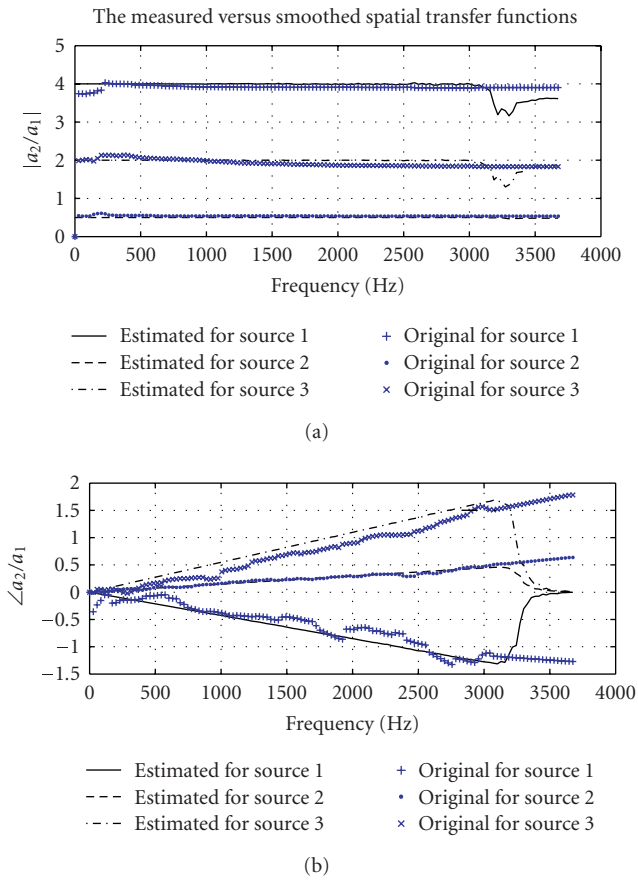

Figure 5: Channel transfer function estimation for three sources using two microphones.

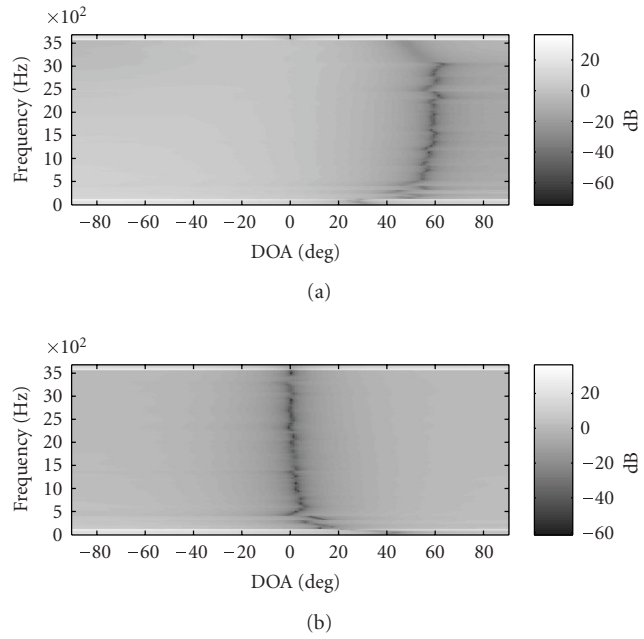

Figure 6: Spatial pattern obtained by the inverse of the mixing matrix for each frequency in the case of two sources at 0° and 60°.

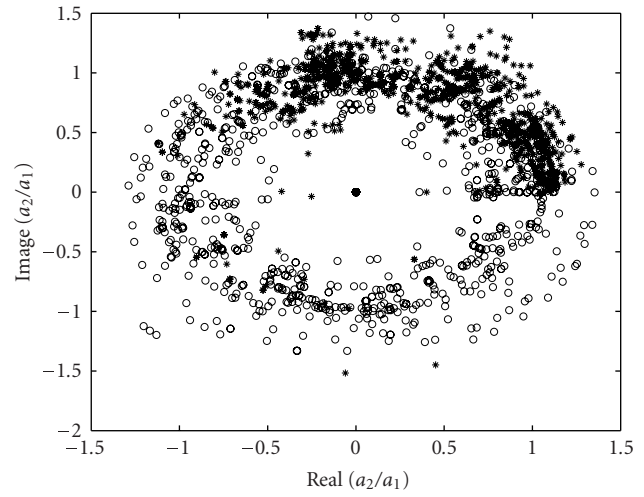

Figure 7: Distribution of the ratio of spatial transfer function values a_2/a_1 in the complex plane for two real sources (indicated by circles and asterisks) over different frequencies at TF points that have been detected as single-TFs.

and phase of the spatial transfer function for a single-source TF cell containing additive white noise. A central step in the W-disjoint approach is the clustering of the parameters in amplitude and delay space so as to identify separate sources in the mixtures. Usually this clustering step is performed under the assumption of constant amplitude and delay over frequency and is possible for speech signals when the sources are distinctly localized both in amplitude and delay. It should be noted that these methods can not handle multipath, that is, when more than one peak in the amplitude and delay space corresponds to a single source. Figure 7 shows the distribution of the ratio of spatial transfer function values a_2/a_1 in the complex plane for two real sources over different frequencies at TF points that have been detected as single-TFs. It can be seen from the figure that these values have significant overlap in amplitude and phase. It is evident that simple clustering can not separate these sources and more sophisticated methods are required.

6. CONCLUSIONS

In this paper, we presented a new method for speech source separation based on directionally-disjoint estimation of the transfer functions between microphones and sources at different frequencies and at multiple times. We assume that the mixed signals contain a combination of source signals in a reverberant environment, such as speech or music recorded with close microphones, where the mixing effect is a direct path delay in addition to a combination of weak multipath delays. The proposed algorithm detects the transfer functions in the frequency domain using eigenvector analysis of the

spatial correlation matrix at single-TF instances. The advantage of our approach is that it allows transfer function estimation even in difficult conditions where the amplitudes of the mixed signals are approximately equal, and it can operate for both wideband and narrowband sources. The current work successfully extends common BSS methods that use a single-TF detection criterion to the convolutive case. The paper formulates single-TF detection and transfer function permutation problems in a principled and optimal manner.

ACKNOWLEDGMENT

This work was partially supported by the Israeli Science Foundation (ISF).

REFERENCES

[1] K. Torkkola, "Blind separation for audio signals—are we there yet?" in *Proceedings of 1st International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '99)*, pp. 239–244, Aussois, France, January 1999.

[2] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[3] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 5, pp. 2985–2988, Istanbul, Turkey, June 2000.

[4] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[5] C. Fevotte and C. Doncarli, "Two contributions to blind source separation using time-frequency distributions," *IEEE Signal Processing Letters*, vol. 11, no. 3, pp. 386–389, 2004.

[6] Y. Deville, "Temporal and time-frequency correlation-based blind source separation methods," in *Proceedings of 4th International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '03)*, pp. 1059–1064, Nara, Japan, April 2003.

[7] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: investigation and solutions," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 1–13, 2005.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[9] A. Steinhardt, "Adaptive multisensor detection and estimation," in *Adaptive Radar Detection and Estimation*, S. Haykin and A. Steinhardt, Eds., pp. 91–160, John Wiley & Sons, New York, NY, USA, 1992.

[10] D. W. E. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proceedings of 1st International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '99)*, Aussois, France, January 1999.

Shlomo Dubnov received the Ph.D. degree in computer science from Hebrew University of Jerusalem, Jerusalem, Israel. He holds also a B.A. degree in music composition from the Rubin Academy of Music and Dance, Jerusalem. From 1996 to 1998, he worked as Invited Researcher at IRCAM, Centre Pompidou, Paris. During 1998–2003, he was a Senior Lecturer in the Department of Communication Systems Engineering at Ben-Gurion University of the Negev, Beer-Sheva, Israel. He is now an Associate Professor in the Department of Music and a Researcher at New Media Arts, CALIT2, University of California, San Diego.

Joseph Tabrikian received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Tel-Aviv University, Tel-Aviv, Israel, in 1986, 1992, and 1997, respectively. During 1996–1998, he was with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, as an Assistant Research Professor. He is now a Faculty Member in the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel. His research interests include statistical signal processing, source detection and localization, and speech and audio processing. He served as an Associate Editor of the IEEE Transactions on Signal Processing during 2001–2004.

Miki Arnon-Targan received the B.S. degree from the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 1998, where he is now studying towards the M.S. degree. He currently works as an Electrical Engineer in the Israel Aircraft Industries, Israel. His research interest lies in the fields of speech and statistical signal processing.