Convolutive Blind Separation of Non-Stationary Sources

Lucas Parra, Clay Spence

Abstract— **Acoustic signals recorded simultaneously in a reverberant environment can be described as sums of differently convolved sources. The task of source separation is to identify the multiple channels and possibly to invert those in order to obtain estimates of the underlying sources. We tackle the problem by explicitly exploiting the non-stationarity of the acoustic sources. Changing cross-correlations at multiple times give a sufficient set of constraints for the unknown channels. A least squares optimization allows us to estimate a forward model, identifying thus the multi-path channel. In the same manner we can find an FIR backward model, which generates well separated model sources. Furthermore, for more than three channels we have sufficient conditions to estimate underlying additive sensor noise powers. We show good performance in real room environments and demonstrate the algorithm's utility for automatic speech recognition.**

Keywords— **blind source separation, non-stationary signals, multi-path channel, multiple decorrelation, frequency domain.**

I. Introduction

In recent years a growing number of researchers have published on the problem of blind source separation. The problem seems to be relevant in various application areas, e.g., speech enhancement with multiple microphones, crosstalk removal in multichannel communications, multipath channel identification and equalization, direction of arrival (DOA) estimation in sensor arrays, improvement over beamforming microphones for audio and passive sonar, and discovery of independent sources in various biological signals, such as EEG, MEG and others. Additionally, theoretical progress in our understanding of the importance of higher-order statistics in signal modeling has generated new techniques for addressing the problem of identifying statistically independent signals — a problem that lies at the heart of source separation. This development has been driven not only by the signal processing community but also by machine learning research that has treated the issue mainly as a density estimation task.

The basic problem is simply described. Assume d_s statistically independent sources $\mathbf{s}(t) = [s_1(t), ..., s_{d_s}(t)]^T$. These sources are convolved and mixed in a linear medium leading to d_x sensor signals $\mathbf{x}(t) = [x_1(t), ..., x_{d_x}(t)]^T$ that may include additional sensor noise $\mathbf{n}(t)$,

$$\mathbf{x}(t) = \sum_{\tau=0}^{P} \mathbf{A}(\tau)\mathbf{s}(t-\tau) + \mathbf{n}(t)\,. \tag{1}$$

How can one identify the $d_x d_s P$ coefficients of the channels $\mathbf{A}$ and how can one find an estimate $\hat{\mathbf{s}}(t)$ for the unknown sources?

Alternatively one may formulate an FIR inverse model $\mathbf{W}$,

$$\mathbf{u}(t) = \sum_{\tau=0}^{Q} \mathbf{W}(\tau)\mathbf{x}(t-\tau)\,, \tag{2}$$

and try to estimate $\mathbf{W}$ such that the model sources $\mathbf{u}(t) = [u_1(t), ..., u_{d_u}(t)]^T$ are statistically independent.

Early work in the signal processing community had suggested decorrelating the measured signals, i.e., diagonalizing measured correlations for multiple time delays [1], [2]. For an instantaneous mix ($P = 1$, or $Q = 1$), also referred to as the constant gain case, it has been shown that for non-white signals, decorrelation of multiple taps are sufficient to recover the sources [3], [4]. Early on however it became clear that for convolutive mixtures ($P > 1$) of wide-band signals this solution is not unique [5], and in fact may generate source estimates that are decorrelated but not statistically independent. As clearly pointed out by Weinstein et al. in [6] additional conditions are required. In order to find separated sources it seems one would have to capture more than second-order statistics, since, indeed, statistical independence requires that not only second but all higher cross moments vanish.

Comon [7], [8] formulated the problem of an instantaneous linear mix, clearly defining the term *independent component analysis*, and presented an algorithm that measures independence by capturing higher-order statistics of the signals. Previous work on DOA estimation had already suggested higher-order statistics [9], [10]. Cardoso [11] suggested that the eigenstructure of 4th order cumulants could be used for blind separation. Herault and Jutten [12] were the first to capture higher statistics by decorrelating nonlinear transformations of the signals. Pham [13] and later Bell and Sejnowski [14] presented a simple algorithmic architecture which in effect performs density estimation [15], [16] and is based on prior knowledge of the cumulative density function of the source signals. Amari et al. made modifications to the update equations to dramatically improve convergence and computational costs [17].

In the convolutive case Yellin and Weinstein [18] established conditions on higher order multi-tap cross moments

Manuscript submitted May 19, 1998; revised Dec. 23, 1998; accepted March 11, 1999. Patent pending. Sarnoff Corporation, CN-5300, Princeton, NJ 08543, E-mail: lparra@sarnoff.com, cspence@sarnoff.com.

that allow convolutive cross talk removal. Thi and Jutten [19] gave simpler iterative algorithms to estimate the forward model $\mathbf{A}$ based on multi-tap third and forth-order cross moments. Recently Shamsunder and Giannakis have suggested to use trispectra, i.e. a frequency representation of fourth order moments, to compute the forward model [20]. Although these algorithms extend to higher dimensions, these researchers have concentrated on the two-dimensional case since in two dimensions a multichannel FIR forward model can be inverted with a properly chosen architecture using the estimated forward filters. For higher dimensions, however, the problem of finding a stable approximation of the forward model remains unsolved.

In contrast, the density estimation approaches mentioned before generalize to the convolutive case by estimating an FIR backward model $\mathbf{W}$ that directly tries to generate independent model sources [21], [22]. They resemble equations obtained from multidimensional extensions of the Busgang blind equalization method [21]. Maximum likelihood density estimation derivations of this type of algorithm are given in [23], [24].

All these techniques are shown to work satisfactorily in computer simulations but perform poorly for recordings in real environments. One could speculate that the densities may not have the hypothesized structures, the higher-order statistics may lead to estimation instabilities, or the violation of the stationarity condition may cause problems.

An alternative approach to the statistical independence condition is to exploit the additional second-order information provided by non-stationary signals. While it is true that diagonalization of single-time cross-correlation (or cross-power spectrum) is not sufficient, additional information is obtained if one considers second-order statistics at multiple times. This has been used in the instantaneous case [25] and the equivalent problem of convolutive mixtures of narrow-band signals [26]. In the convolutive case with broad-band signals this idea has been touched on by Weinstein et al. in [6]. For non-stationary signals a set of second-order conditions can be specified that uniquely determine the parameters $\mathbf{A}$. No algorithm was given in [6] nor have there been any results reported on this approach to our knowledge. Ehlers and Schuster [27] recently proposed a related algorithm that attempts to solve for the frequency components of $\mathbf{A}$ by extending prior work of Molgedey and Schuster [28] on instantaneous mixtures into the frequency domain. However, they mistakenly confuse this idea with simple decorrelation of multiple taps in the time domain, which is known to be insufficient [5]. After our first presentation of this work in [29] we have found contemporary work by Principe [30] who suggests a similar approach for the time domain, and Murata et al. [31], and Kawamoto [32] for the frequency domain.

We take up this multiple decorrelation approach assuming non-stationary signals and use a least squares (LS) optimization to estimate $\mathbf{A}$ or $\mathbf{W}$ as well as signal and noise powers. As such, the algorithm makes no assumptions about the cumulative densities of the signals and limits itself to more robust second-order statistics. Unlike most previous work on source separation, we take additive sensor noise into account.

In Section II we present our approach for the instantaneous case and point out the differences between estimating the forward model $\mathbf{A}$ and the backward model $\mathbf{W}$. In addition to the source power one can estimate additive sensor noise powers. Computing estimates $\hat{\mathbf{s}}$ from a forward model $\mathbf{A}$ requires a further estimation step, in particular for the case of fewer sources than sensors, i.e. $d_s < d_x$. The least squares (LS), maximum likelihood (ML) or maximum a posteriori probability (MAP) estimates are given in Section II-C. In a backward model $\mathbf{W}$ the LS optimization gives the inverse of the mixture and we obtain model sources $\mathbf{u}$ directly. In Section III we carry over the concept of multiple decorrelation to the convolutive case by solving independent models for every frequency. We pay particular attention to the approximation of linear convolutions by circulant convolutions in Section III-A as well as the permutation issue in Section III-C. Since inverting a multichannel forward FIR model is in itself a challenging task we restrict ourself in the implementations to estimating the inverse model $\mathbf{W}$. Finally we report some encouraging results on real room recordings in and demonstrate the utility of the algorithm for automatic speech recognition in Section IV.

II. Instantaneous Mixture

As discussed in the previous section, for the instantaneous case a multitude of approaches have been proposed. We present it here in order to lay out some basic ideas, which will be used again in the convolutive case. Part of our treatment of additive sensor noise estimates goes beyond previous work.

A. Forward model estimation

For an instantaneous mixture, i.e. $P = 1$, the forward model (1) simplifies to

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t)\,. \tag{3}$$

We can formulate the covariance $R_x(t)$ of the measured signals at time t with the assumption of independent noise as

$$\begin{aligned} R_x(t) \equiv& \left\langle \mathbf{x}(t)\mathbf{x}^T(t) \right\rangle \\ =& \mathbf{A} \left\langle \mathbf{s}(t)\mathbf{s}^T(t) \right\rangle \mathbf{A}^T + \left\langle \mathbf{n}(t)\mathbf{n}^T(t) \right\rangle \\ \equiv& \mathbf{A}\Lambda_s(t)\mathbf{A}^T + \Lambda_n(t)\,. \end{aligned} \tag{4}$$

Since we assume uncorrelated sources at all times, we postulate diagonal covariance matrixes $\Lambda_s(t)$. We also assume uncorrelated noise at each sensor, i.e. diagonal $\Lambda_n(t)$.

Note that any scaling and permutation of the coordinates of $\Lambda_s(t)$ can be absorbed by $\mathbf{A}$. It is well known that the solution is therefore only specified up to an inherently arbitrary permutation and scaling. We are therefore free to choose the scaling of the coordinates in $\mathbf{s}$. For now we choose $A_{ii} = 1\,, i = 1, ..., d_s$, which places d_s conditions on our solutions.

For non-stationary signals, a set of K equations (4) for different times $t_1, ...t_K$ and the d_s scaling conditions give a total of $Kd_x(d_x+1)/2 + d_s$ constraints on $d_s d_x + d_s K + d_x K$ unknown parameters $\mathbf{A}, \Lambda_s(t_1), ..., \Lambda_s(t_K), \Lambda_n(t_1), ..., \Lambda_n(t_K)$.[1] Assuming all conditions are linearly independent we have sufficient conditions if

$$Kd_x(d_x+1)/2 + d_s \geq d_s d_x + d_s K + d_x K\,. \tag{5}$$

It is interesting to note that in the square case, $d_s = d_x$, there are not sufficient constraints to determine the additional noise parameters unless $d_x \geq 4$, no matter how many more times one considers.[2] If we assume zero additive noise, in principle $K = 2$ is sufficient to specify the solution up to arbitrary permutations.

In the square case, the solutions can be found by solving a non-symmetric eigenvalue problem, as outlined in [28]. The difficulty with such algebraic solutions, however, is that one does not have perfect estimates of $R_x(t)$. At best one can assume non-stationary signals and measure the sample estimates $\hat{R}_x(t)$ within some time interval. If we interpret the inaccuracy of that estimation as measurement error

$$E(k) \equiv \hat{R}_x(k) - \Lambda_n(k) - \mathbf{A}\Lambda_s(k)\mathbf{A}^T\,, \tag{6}$$

it is reasonable to estimate the unknown parameters by minimizing the total measurement error for a sufficiently large K,

$$\hat{\mathbf{A}}, \hat{\Lambda}_s, \hat{\Lambda}_n = \underset{\mathbf{A},\Lambda_s,\Lambda_n,A_{ii}=1}{\arg\min} \sum_{k=1}^{K} \|E(k)\|^2\,. \tag{7}$$

The matrix norm here is the sum of the absolute squared values of every coefficient. Note that $\|E(k)\|^2 = \mathrm{Tr}[E(k)E^H(k)]$. This represents a least squares (LS) estimation problem. To find the extrema of the LS cost $J = \sum_{k=1}^{K} \|E(k)\|^2$ in (7) let us compute the gradients with respects to its parameters[3]

$$\frac{\partial J}{\partial \mathbf{A}} = -4 \sum_{k=1}^{K} E(k)\mathbf{A}\Lambda_s(k) \tag{8}$$

$$\frac{\partial J}{\partial \Lambda_s(k)} = -2\,\mathrm{diag}\left[\mathbf{A}^T E(k)\mathbf{A}\right] \tag{9}$$

$$\frac{\partial J}{\partial \Lambda_n(k)} = -2\,\mathrm{diag}\left[E(k)\right] \tag{10}$$

We can find the minimum with respect to $\mathbf{A}$, and $\Lambda_s(k)$ with a gradient descent algorithm using the gradients (8), and (9). The optimal $\Lambda_n(k)$ for given $\mathbf{A}$ and $\Lambda_s(k)$ at every gradient step can be computed explicitly by setting the gradient in (10) to zero, which yields $\hat{\Lambda}_n(k) = \mathrm{diag}\left[\hat{R}_x(k) - \mathbf{A}\Lambda_s(k)\mathbf{A}^T\right]$.

B. Normalization conditions

In the previous section we proposed to fix the arbitrary scaling by setting the diagonal parameters $A_{ii} = 1$. For the non-square case this normalization may seem somewhat arbitrary. One could in such a case demand instead that $\|\mathbf{a}_j\| = 1, j = 1, ..., d_s$ with $\mathbf{a}_j = [A_{1j}, ..., A_{d_x j}]^T$. Instead of the gradients given in (8) one then has to consider their projections onto the hyper-planes defined by $\|\mathbf{a}_j\| = 1$. The projection operator for the jth column of $\partial J/\partial \mathbf{A}$ is

$$P_j^{(1)} = I - \mathbf{a}_j \mathbf{a}_j^T\,. \tag{11}$$

Or we can write a constraint gradient

$$\left.\frac{\partial J}{\partial \mathbf{A}}\right|_{\|\mathbf{a}_j\|=1} = \frac{\partial J}{\partial \mathbf{A}} - \mathbf{A}\,\mathrm{diag}\left[\mathbf{A}^T \frac{\partial J}{\partial \mathbf{A}}\right]\,. \tag{12}$$

C. Estimation of source signals

In the case of a square and invertible mixing matrix $\hat{\mathbf{A}}$, the signal estimates are trivially computed to be $\hat{\mathbf{s}} = \hat{\mathbf{A}}^{-1}\mathbf{x}$. In the non-square case for $d_s < d_x$ we can compute the LS estimate

$$\hat{\mathbf{s}}_{\mathrm{LS}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \|\mathbf{x}(t) - \hat{\mathbf{A}}\mathbf{s}(t)\| = \left(\hat{\mathbf{A}}^T\hat{\mathbf{A}}\right)^{-1}\hat{\mathbf{A}}^T\mathbf{x}(t)\,. \tag{13}$$

If we assume the additive noise to be Gaussian, but not necessarily white or stationary, we can compute the maximum likelihood (ML) estimate

$$\begin{aligned}\hat{\mathbf{s}}_{\mathrm{ML}}(t) &= \underset{\mathbf{s}(t)}{\arg\max}\, p[\mathbf{x}(t)|\mathbf{s}(t); \hat{\mathbf{A}}, \hat{\Lambda}_n(t)] \\ &= \left[\hat{\mathbf{A}}^T \hat{\Lambda}_n(t)^{-1} \hat{\mathbf{A}}\right]^{-1} \hat{\mathbf{A}}^T \hat{\Lambda}_n(t)^{-1}\mathbf{x}(t)\,.\end{aligned} \tag{14}$$

where $p()$ is the Gaussian probability density given by the noise density. If we further assume the signal to be Gaussian, again not necessarily white or stationary, we can compute the maximum a posteriori probability (MAP) estimate. For Gaussian densities the MAP estimate is equal to the conditional expectation $E[\mathbf{s}(t)|\mathbf{x}(t); \hat{\mathbf{A}}, \hat{\Lambda}_n(t), \hat{\Lambda}_s(t)]$

$$\begin{aligned}\hat{\mathbf{s}}_{\mathrm{MAP}}(t) &= \underset{\mathbf{s}(t)}{\arg\max}\, p[\mathbf{s}(t)|\mathbf{x}(t); \hat{\mathbf{A}}, \hat{\Lambda}_n(t), \hat{\Lambda}_s(t)] \\ &= E[\mathbf{s}(t)|\mathbf{x}(t); \hat{\mathbf{A}}, \hat{\Lambda}_n(t), \hat{\Lambda}_s(t)] \\ &= \left[\hat{\mathbf{A}}^T \hat{\Lambda}_n(t)^{-1} \hat{\mathbf{A}} + \hat{\Lambda}_s(t)^{-1}\right]^{-1} \hat{\mathbf{A}}^T \hat{\Lambda}_n(t)^{-1}\mathbf{x}(t)\,.\end{aligned} \tag{15}$$

Note however that the resulting estimates may not be uncorrelated. Assuming that the model is correct and that we found the correct estimate $\hat{\mathbf{A}} \approx \mathbf{A}$,

$$\left\langle \hat{\mathbf{s}}_{\mathrm{LS}} \hat{\mathbf{s}}_{\mathrm{LS}}^T \right\rangle \approx \left\langle \mathbf{s}\mathbf{s}^T \right\rangle + \left(\hat{\mathbf{A}}^T\hat{\mathbf{A}}\right)^{-1}\hat{\mathbf{A}}^T\Lambda_n\hat{\mathbf{A}}\left(\hat{\mathbf{A}}^T\hat{\mathbf{A}}\right)^{-1}\,. \tag{16}$$

[1] We will abbreviate the notation in the reminder of the paper by writing $\Lambda_s(k)$ for $\Lambda_s(t_k)$ and dropping the argument, i.e. Λ_s, when we refer to all $\Lambda_s(t_1), ..., \Lambda_s(t_K)$. We use this notation also for $\Lambda_n(t)$ and $R_x(t)$.

[2] One can see this by re-writing the inequality as $K(d_x^2 - 3d_x) + 2(d_x - d_x^2) \geq 0$. The second term is never positive, and the first is only positive if $d_x \geq 4$.

[3] The diagonalization operator here zeros the off-diagonal elements, i.e. $\mathrm{diag}(M)_{ij} = \begin{cases} M_{ij}, & i = j \\ 0, & i \neq j \end{cases}$

Since the second term may not be diagonal, the resulting estimates can be correlated. However, this is not a problem since the correlation is entirely due to correlated noise and the signal portion of the estimates remains uncorrelated.

D. Backward model

Instead of estimating a forward model and then from that further estimating the source signal, one may try to directly estimate a backward model in the form of (2), in order to model separated sources $\hat{\mathbf{s}}(t)$, which we define as

$$\hat{\mathbf{s}}(t) \equiv \mathbf{W}\mathbf{A}\mathbf{s}(t)\,. \tag{17}$$

We are looking therefore for a $\mathbf{W}$ that inverts $\mathbf{A}$. This will be especially relevant for the discussion of the convolutive case in the next section. In analogy with the previous discussions and using definition (17) and assuming (3) we have

$$\left\langle \hat{\mathbf{s}}(t)\hat{\mathbf{s}}(t)^T \right\rangle = \mathbf{W}\left[R_x(t) - \Lambda_n(t)\right]\mathbf{W}^T\,. \tag{18}$$

We will search for $\mathbf{W}$ such that $\left\langle \hat{\mathbf{s}}(t)\hat{\mathbf{s}}(t)^T \right\rangle$ diagonalizes simultaneously for K different times[4]. The LS estimate is then

$$\hat{\mathbf{W}}, \hat{\Lambda}_s, \hat{\Lambda}_n = \underset{\mathbf{W},\Lambda_s,\Lambda_n,W_{ii}=1}{\arg\min} \sum_{k=1}^{K} \|E(k)\|^2\,, \tag{19}$$
$$\text{where}\quad E(k) = \mathbf{W}\big(\hat{R}_x(k) - \Lambda_n(k)\big)\mathbf{W}^T - \Lambda_s(k)\,.$$

In analogy to the discussion in section II-A we can find the solutions with an iterative gradient algorithm.

III. Convolutive Mixture

In the previous section we described how one can treat the case of instantaneous mixtures by decorrelating the covariance matrices at several times. This approach requires non-stationary sources. The problem can also be treated by decorrelating the cross-correlation at different lags. This requires that the signals be non-white rather than non-stationary. This is the approach traditionally taken in the literature [2], [28], [4], [6].

Recall now equation (1) of the convolutive case:

$$\mathbf{x}(t) = \mathbf{A} * \mathbf{s}(t) + \mathbf{n}(t)\,. \tag{20}$$

As suggested for other source separation algorithms, our approach to the convolutive case is to transform the problem into the frequency domain and to solve simultaneously a separation problem for every frequency [33], [34], [22], [27], [31]. The solution for each frequency would seem to have an arbitrary permutation. The main issues to be addressed here are how to obtain equations equivalent to (4) or (18) in the frequency domain, and how to choose the arbitrary permutations for all individual problems consistently. We will take up these issues in the following sections.

[4]Similar considerations to those given for (16) show that decorrelating $\mathbf{u}(t) = \mathbf{W}\mathbf{x}(t)$, as defined in (2), rather than $\hat{\mathbf{s}}(t)$ may not lead to the correct solution in the presence of sensor noise.

A. Cross-correlations, circular and linear convolution

First consider the cross-correlations $R_x(t, t+\tau) = \left\langle \mathbf{x}(t)\mathbf{x}(t+\tau)^T \right\rangle$. For stationary signals the absolute time does not matter and the correlations depend on the relative time, i.e. $R_x(t, t+\tau) = R_x(\tau)$. Denote with $R_x(z)$ the z-transform of $R_x(\tau)$. We can then write

$$R_x(z) = \mathbf{A}(z)\Lambda_s(z)\mathbf{A}^H(z) + \Lambda_n(z) \tag{21}$$

where $\mathbf{A}(z)$ represents the matrix of z-transforms of the FIR filters $\mathbf{A}(\tau)$, and $\Lambda_s(z)$, and $\Lambda_n(z)$ are the z-transform of the auto-correlation of the sources and noise. Again they are diagonal due to the independence assumptions.

For practical purposes we have to restrict ourself to a limited number of sampling points of z. Naturally we will take T equidistant samples on the unit circle such that we can use the discrete Fourier transform (DFT). For periodic signals the DFT allows us to express circular convolutions as products such as in (21). However, in (1) and (2) we assumed linear convolutions. A linear convolution can be approximated by a circular convolution if the frame size T of the DFT is much larger than the channel length P. We can then write approximately

$$\mathbf{x}(\omega, t) \approx \mathbf{A}(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t), \text{ for } P \ll T \tag{22}$$

where $\mathbf{x}(\omega, t)$ represents the DFT of the frame of size T starting at t, $[\mathbf{x}(t), ..., \mathbf{x}(t+T)]$, given by $\mathbf{x}(\omega, t) = \sum_{\tau=0}^{T-1} e^{-\mathrm{i}2\pi\omega\tau/T}\mathbf{x}(t+\tau)$, and corresponding expressions apply for $\mathbf{s}(\omega, t)$ and $\mathbf{A}(\omega)$.

For non-stationary signals, the cross-correlation will be time dependent. Estimating the cross-power-spectrum at the desired resolution of $1/T$ is difficult if the stationarity time of the signal is on the order of magnitude of T or smaller. We are content, however, with any cross-power-spectrum average that diagonalizes for the source signals. One such sample average is

$$\bar{R}_x(\omega, t) = \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{x}(\omega, t+nT)\mathbf{x}^H(\omega, t+nT)\,. \tag{23}$$

We can then write for such averages

$$\bar{R}_x(\omega, t) = \mathbf{A}(\omega)\Lambda_s(\omega, t)\mathbf{A}^H(\omega) + \Lambda_n(\omega, t)\,. \tag{24}$$

If N is sufficiently large we can model $\Lambda_s(\omega, t)$ and $\Lambda_n(\omega, t)$ as diagonal, again due to the independence assumption. For equations (24) to be linearly independent for different times t it will be necessary that $\Lambda_s(\omega, t)$ changes over time for a given frequency, i.e. the signals are non-stationary.

B. Backward model

Given a forward model $\mathbf{A}$, it is not guaranteed that we can find a stable inverse. In the two-dimensional square case the inverse channel is easily determined from the forward model [6], [19]. However, it is not apparent how to compute a stable inversion for arbitrary dimensions. In this present work we prefer to directly estimate a stable multi-path backward FIR model such as (2). In analogy to

the discussion above and in Section II-D, we wish to find model sources with cross-power-spectra satisfying[5]

$$\Lambda_s(\omega,t) = \mathbf{W}(\omega)\left[\bar{R}_x(\omega,t) - \Lambda_n(\omega,t)\right]\mathbf{W}^H(\omega)\,. \quad (25)$$

In order to obtain independent conditions for every time, we choose the times such that we have non-overlapping averaging times for $\bar{R}_x(\omega,t_k)$, i.e. $t_k = kTN$. But if the signals vary sufficiently fast, overlapping averaging times could have been chosen. A multipath model $\mathbf{W}$ that satisfies these equations for K times simultaneously can be found, again with a LS estimate[6]

$$E(\omega,k) = \mathbf{W}(\omega)\left[\bar{R}_x(\omega,k) - \Lambda_n(\omega,k)\right]\mathbf{W}^H(\omega) - \Lambda_s(\omega,k)\,,$$

$$\hat{\mathbf{W}},\hat{\Lambda}_s,\hat{\Lambda}_n = \underset{\substack{\mathbf{W},\Lambda_s,\Lambda_n,\\ \mathbf{W}(\tau)=0,\tau>Q\ll T,\\ W_{ii}(\omega)=1}}{\arg\min} \sum_{\omega=1}^{T}\sum_{k=1}^{K} \|E(\omega,k)\|^2\,. \quad (26)$$

Note the additional time domain constraint on the filter size Q relative to the frame size T. This condition can be satisfied by choosing short filters or alternatively larger frame sizes T. Up to that constraint it would seem the various frequencies $\omega = 1,...,T$ represent independent problems. However the solutions $\mathbf{W}(\omega)$ are restricted to those filters that have zero time response for $\tau > Q \ll T$. Effectively we are parameterizing Td_sd_x filter coefficients in $\mathbf{W}(\omega)$ with Qd_sd_x parameters $\mathbf{W}(\tau)$. The LS solutions can again be found with a gradient descent algorithm. We will first compute the gradients with respect to the complex valued filter coefficients $\mathbf{W}(\omega)$ and discuss their projections onto the subspace of permissible solutions in the following section.

For any real valued function $f(\mathbf{z})$ of a complex valued variable $\mathbf{z}$ the gradients with respect to the real and imaginary part are obtained by taking derivatives formally with respect to the conjugate quantities $\mathbf{z}^*$, ignoring the non-conjugate occurrences of $\mathbf{z}$ [35], [36], i.e.,

$$\frac{\partial f(z)}{\partial \Re(z)} + i\frac{\partial f(z)}{\partial \Im(z)} = 2\frac{\partial f(z)}{\partial z^*}\,. \quad (27)$$

Therefore the gradients of the LS cost in (26) are

$$\frac{\partial J}{\partial \mathbf{W}^*(\omega)} = 2\sum_{k=1}^{K} E(\omega,k)\mathbf{W}(\omega)\left[\bar{R}_x(\omega,k) - \Lambda_n(\omega,k)\right] \quad (28)$$

$$\frac{\partial J}{\partial \Lambda_s^*(\omega,k)} = -\operatorname{diag}\left[E(\omega,k)\right] \quad (29)$$

$$\frac{\partial J}{\partial \Lambda_n^*(k)} = -\operatorname{diag}\left[\mathbf{W}^H(\omega)E(\omega,k)\mathbf{W}(\omega)\right]\,. \quad (30)$$

[5] $\mathbf{W}(\omega)$ represents the DFT with frame size T in the time domain $\mathbf{W}(\tau)$. In the following, time and frequency domain are identified by their argument τ or ω.

[6] Again, we abbreviate the notation by writing $\Lambda_s(\omega,k)$ for $\Lambda_s(\omega,t_k)$ and dropping the argument, i.e. Λ_s, when we refer to all $\Lambda_s(\omega,t_k)$ for all ω and k. We use this notation also for $\Lambda_n(\omega,t)$ and $R_x(\omega,t)$.

Again, we can find the minimum with respect to $\mathbf{W}(\omega)$, and $\Lambda_n(\omega,k)$ with a constrained gradient descent algorithm using the gradients (28), and (30). The optimal $\Lambda_s(\omega,k)$ for given $\mathbf{W}(\omega)$ and $\Lambda_n(\omega,k)$ at every gradient step can be computed explicitly by setting the gradient in (29) to zero, which yields $\hat{\Lambda}_s(k) = \operatorname{diag}\left[\mathbf{W}(\omega)\hat{R}_x(\omega,k)\mathbf{W}^T(\omega) - \Lambda_n(\omega,k)\right]$.

For the simulations in Section IV we have found that convergence of the gradient algorithm can be improved substantially if a different adaptation constant is used for every frequency. Note that the gradient terms scale with the square of the signal powers R_x. The signal powers vary considerably across frequency. As a result, the gradient terms for different frequencies have very different magnitudes. Normalizing by the powers will therefore scale the gradient to give comparable update steps for different frequencies. This can be achieve easily by defining a weighted cost, $J = \sum_{\omega=1}^{T}\sum_{k=1}^{K} m(\omega)\,\|E(\omega,k)\|^2$. We find good results by choosing a straightforward power normalization, $m(\omega) = \left(\sum_{k=1}^{K}\|R_x(\omega,k)\|^2\right)^{-1}$.

C. Permutations and constraints

Note that arbitrary permutations of the coordinates for each frequency ω will lead to the same error $E(\omega,k)$. Therefore the total cost will not change if we choose a different permutation of the solutions for each frequency ω. This seems to be a serious problem since only consistent permutations for all frequencies will correctly reconstruct the sources.

Arbitrary permutations, however, will not satisfy the condition on the length of the filter, $\mathbf{W}(\tau) = 0$ for $\tau > Q \ll T$. Effectively, requiring zero coefficients for elements with $\tau > Q$ will restrict the solutions to be continuous or "smooth" in the frequency domain, e.g., if $Q/T = 8$ the resulting DFT corresponds to a convolved version of the coefficients with a sinc function 8 times wider than the sampling rate.

The constraint on the filter size Q versus the frequency resolution $1/T$ links the otherwise independent frequencies, and solves the frequency permutation problem — a crucial point that may have not been realized in previous literature. In addition, it is a necessary condition for equations (25) to hold to a good approximation. Note also that it does not limit the actual filter size, as in principle one can choose an appropriately large frame size T for any given Q.

We can enforce the filter size constraint by properly projecting the unconstrained gradients (28)-(30) to the subspace of permissible solutions. The projection operator that zeros the appropriate delays for every channel $W_{ij} = [W_{ij}(0),...,W_{ij}(\omega),...,W_{ij}(T)]^T$ is

$$P^{(2)} = FZF^{-1} \quad (31)$$

where the DFT is given by $F_{ij} = 1/\sqrt{T}e^{-\mathrm{i}2\pi ij}$, and Z is diagonal with $Z_{ii} = 1$ for $i < Q$ and $Z_{ii} = 0$ for $i \geq Q$.

The projection operator that enforces unit gains on diagonal filters $W_{ii}(\omega) = 1$ is simply applied by setting the

diagonal terms of the gradients to zero. These projections are orthogonal and can be applied independently of each other. This stands in contrast to the normalization constraint $P^{(1)}$ outlined in Section II-B. That projection operator is not orthogonal to $P^{(2)}$ and care has to be taken to apply a proper projection that maps the gradient to the joint subspace of $P^{(1)}$ and $P^{(2)}$. A simple, though admittedly inefficient, solution is to apply $P^{(1)}$ and $P^{(2)}$ successively and repeatedly to the gradients until convergence. In our simulations, 3-5 iterations were sufficient. The resulting constrained gradient can be used in a gradient update of the filter parameters.

The computational cost of the algorithm is dominated by the costs of estimating $\bar{R}_x(\omega, t)$ in (23), the gradient computation in (28), and the projection (31). Before the gradient descent starts, one needs to evaluate (23) K times, resulting in a computational cost of $O[KNd_xT(\log T + d_x)]$. Thus every gradient step requires a computation of $O[KTd_xd_s(2d_s + d_x)]$ in (28) and $O(d_xd_sKT\log T)$ in (31).

IV. Experimental results

We have done experiments with algorithm (28)-(31)[7] in various realistic environments and obtained very different results depending on the number of sources and microphones, stationarity of the signal, reverberation of the room (size and wall reflectance), type of microphones, etc. Obviously it is hard to evaluate the entire space of possible setups. In this section we report results on experiments with recordings in realistic environments such as offices and conference rooms. We show improvement in the signal to interference ratio and demonstrate improvement over a distant talking microphone for a large vocabulary speech recognizer. For a more systematic study of the various parameters we also performed experiments with simulated room responses.

The following results report the improvements in terms of the signal to interference ratio (SIR), which we define for a signal $\mathbf{s}(t)$ in a multi-path channel $H(\omega)$ to be the total signal powers of the direct channel versus the signal power stemming from cross channels,

$$SIR[H,s] = \frac{\sum_\omega \sum_i |H_{ii}(\omega)|^2 \left\langle |s_i(\omega)|^2 \right\rangle}{\sum_\omega \sum_{i\neq j} \sum_j |H_{ij}(\omega)|^2 \left\langle |s_j(\omega)|^2 \right\rangle}. \qquad (32)$$

In the case of known channels and source signals we can compute the expressions directly by using a sample average over the available signal and multiplying the powers with the given direct and cross channel responses. In the case of unknown channel response and underlying signals we can estimate the direct powers (numerator) and cross-powers (denominator) by using alternating signals. We estimate the contributions of source j while source j is 'on' and all other sources are 'off'. During periods of silence, i.e. all sources are 'off' we can estimate background noise powers in all channels to subtract from the signal powers. This information is only used for reporting SIR improvements and is obviously not used in the adaptation phase.

[7]We used the weighted gradients as described at the end of section III-B.

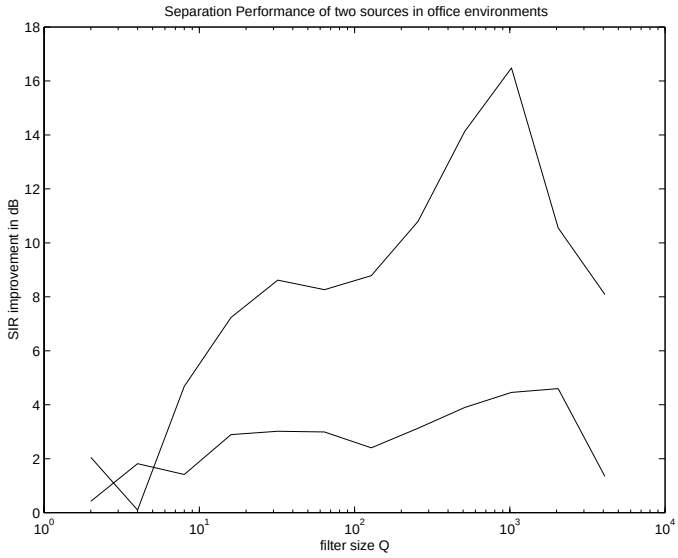

Fig. 1. Separation performance for two speakers recorded with two microphones in two different office environments as a function of separation filter size Q. Upper curve: uni-directional microphones in a $3\,m \times 3.6\,m \times 2.3\,m$ room, 30 s recordings at 8 KHz, 15 s alternating and 15 s simultaneous speech. Lower curve: 10 s simultaneous speech recorded at 16 KHz in a $4.2\,m \times 5.5\,m \times 3.1\,m$ room with omni-directional microphones.

The results obtained for real recordings vary widely. Figure 1 shows the results for varying filter sizes on the separation of two competing speakers whom we recorded with two microphones. In all experiments in this section we used $T/Q = 8$. The improvement in SIR can be as high as 15 dB for recordings obtained in an office room using uni-directional (cardioid) microphones (upper curve). Separating two speakers from the recordings in a second room with omni-directional microphones seems more challenging (lower curve)[8]. As expected, the performance initially increases with increasing filter size, as the inverse of the room can be modeled more accurately. However, larger filters may require more training data, and so the performance eventually decreases given the constant amount of data.

We observed in further experiments that separation works better in large conference rooms than in small office rooms with stronger reflecting walls, most likely due to the increased reverberation. To verify these result more systematically we used simulated room responses, according to [38][9], with 1000-2000 filter taps at 8 KHz. Figure 2 shows the results obtained for varying room sizes. As one can see, with increasing room size the separation perfor-

[8]The data for this second example was provided by the authors of [37].

[9]We thank Joseph G. Desloge and Michael P. O'Connell of the Sensory Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology for the software implementation in MATLAB

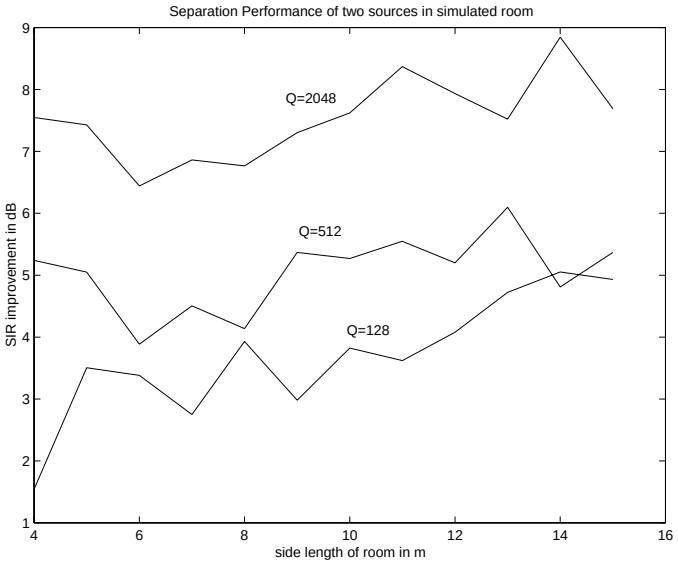

Fig. 2. SIR improvement in simulated rooms of varying size $a \times a \times a/2$, with side length a. Rooms with a typical office room reflectance characteristic were used (gypsum walls, ceiling with acoustic tiles, carpet on concrete floor). The two microphones where placed two meters apart at an anechoic wall, making their response characteristic effectively directional. The speaker is in front of one microphone at a distance of $1\,m$. The interfering music source is at $4\,m$ in front of the second microphone. Results are shown for different separation filter sizes Q. We used 15 s signals at 8 KHz.

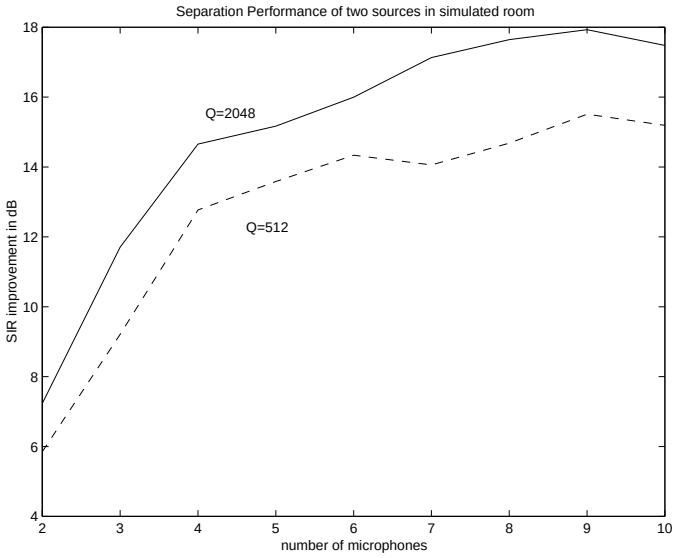

Fig. 3. Same simulation as in previous figure with increasing number of microphones in a $2\,m$ array.

mance improves.

The SIRs in Figs. 1 and 2 do not change smoothly, which may be explained by the fact that the algorithm is not optimizing the SIR directly but instead multiple decorrelations. Also, the gradient algorithm may be reaching different local minima of the diagonalization criterion.

Another interesting question is how the performance improves if we use additional microphones, given a constant number of sources. Again, for a systematic evaluation we used the same simulated room and microphone setup as in the previous experiment. The results in figure 3 show a clear improvement in separating two sources as the number of microphones increases.

From previous experiments we know, however, that for an increasing number of sources and microphones the performance degrades [29]. This is expected as the amount of data increases linearly with the number of microphones, while the number of parameters in our current parameterization increases with the square of the number of channels.

We want to show now the utility of the algorithm for automatic speech recognition, at least in a situation in which we obtain reasonably good SIR improvement. We used a commercial large vocabulary recognizer (IBM's ViaVoice) that was adapted to the speaker and the same distant-talking microphone used in the separation experiments. The recognition performance was estimated on a short text (760 words Wall Street Journal article), and resulted in a word error rate (WER) of 11.9% in a quiet office environment. This number represents a lower bound, as the separation algorithm can at best remove the undesired interference, and is not designed to undo the room acoustics of the direct channel. The results shown in Fig. 4 correspond to SIR improvements between 0 to 10 dB. In almost all instances, the recognition performance was improved. Actual deterioration is also possible as there is no guarantee that the resulting separated signals will not be distorted. Distortion is minimized however by constraining the direct channels to be unit filters.

Unfortunately the most realistic case in a practical application may be that we have only two microphones available but a multitude of sources. We have observed that in the presence of a strong main speech signal, the algorithm will give a good estimate of the background of multiple sources in one channel. It will do little, however, to remove the background from the one main source. This is not surprising as we know from the beam-forming literature that with two microphones, one can at best zero one orientation. However, this background estimate may be useful for further single-channel enhancement.

Another problem to be addressed in practice is that the channel is typically non-stationary as well. A slight change in the location or orientation of a source may cause drastic changes in the response characteristic of a room. A crucial question therefore is the amount of signal required for any algorithm to produce a reasonable separation. Earlier results [29] and current work on an on-line version of this algorithm suggest that 1-2s of signal may be sufficient, provided the channel is easy to invert in a stationary situation.

V. Conclusion

A large body of work has accumulated in the last two decades on the problem of blind source separation. We have concentrated on the rather general case of recovering

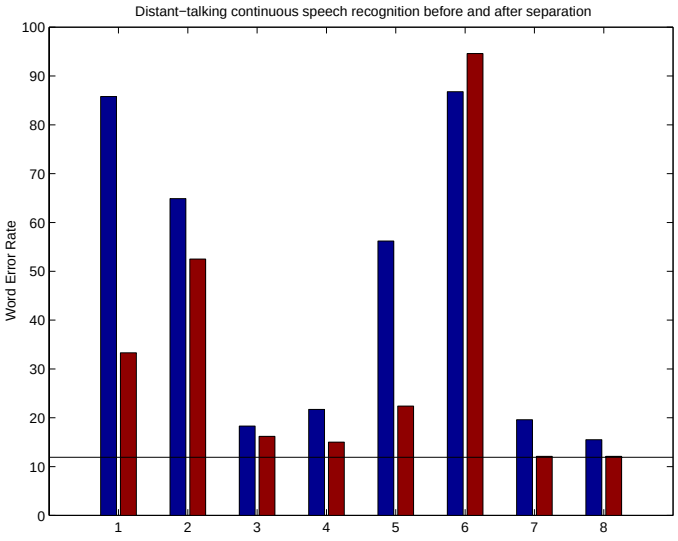

Fig. 4. Word error rate of a large vocabulary, continuous speech recognizer before and after source separation with the current algorithm (left and right bars respectively). The experiments were performed in a small conference room with two cardioid microphones at a sampling rate of 8 KHz. The source of interference was either a second speaker or music from a single loud-speaker. Parallel setup: Microphones placed side by side, one pointing mainly towards the speaker, located at a 150 cm distance, and the second microphone pointing approximately at the interfering source at a 250 cm distance. Opposite setup: Speaker and jammer placed on opposite sides of the two microphones, which also pointed in opposite directions. Speaker and jammer again at 150 cm and 250 cm distance respectively. 1: music, parallel 150 cm; 2: music, parallel 20 cm; 3: music, opposite 45°; 4: music, opposite 180°; 5: speech, parallel 150 cm; 6: speech, parallel 20 cm; 7: speech, opposite 45°; 8: speech, opposite 180°.

convolutive mixtures of wideband signals with at least as many sensors as sources. The main contributions of this work are the explicit use of non-stationarity of the source signals and an efficient solution to the permutation problem of the frequency domain algorithm. Careful considerations of how to measure second-order statistics in the frequency domain allow us to obtain a constrained LS cost that is optimal at the desired solutions. The constraint on the filter size solves the permutation problem of wideband signals. The current experimental results suggest that under proper conditions for two channels we can achieve a crosstalk reduction of up to 14 dB in an office environment. We have demonstrated the algorithm's utility for automatic speech recognition in presence of a single source of interference by using two microphones.

Acknowledgments

The authors would like to thank foremost Bert De Vries from the Adaptive Signal Processing group at Sarnoff for educating the authors on many signal processing issues. Thanks are also due to Jose Principe from the University of Florida for useful suggestions. We also thank Joseph G. Desloge and Michael P. O'Connell of the Sensory Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology for the software implementation in MATLAB that simulates real room acoustics. We thank Daniel Schobben from the EE department of the Eindhoven University of Technology who provided signals used in one of the experiments reported in Fig. 1. Finally we are extremely grateful to the one anonymous reviewer who did a remarkable thorough review of the paper.

References

[1] D. Bradwood, "Cross-coupled cancellation systems for improving cross-polarisation discrimination", in *Proc. IEEE Int. Conf. Antennas and Propagation*, Nov. 1978, vol. I, pp. 41–45.

[2] Y. Bar-Ness, J. Carlin, and M. Steinberger, "Bootstrapping adaptive cross-pol canceller for satellite communications", in *Proc. IEEE Int. Conf. Comunications*, 1982, pp. 4F.5.1–4F.5.5.

[3] S. Van Gerven and D. Van Compernolle, "Signal separation in a symmetric adaptive noise canceler by output decorrelation", in *Proc. ICASSP 92*, 1992, vol. IV, pp. 221–224.

[4] S. Van Gerven and D. Van Compernolle, "Signal Separation by Symetric Adaptive Decorrelation: Stability, Convergence, and Uniqueness", *IEEE Trans. Signal Processing*, vol. 43, no. 7, pp. 1602–1612, July 1995.

[5] R.L.L Tong and Y.H.V.C. Soon, "Interdeterminacy and Identifiability of blind identification", *IEEE Trans. Circuits Syst.*, vol. 38, no. 5, pp. 499–509, May 1991.

[6] E. Weinstein, M. Feder, and A.V. Oppenheim, "Multi-Channel Signal Separation by Decorrelation", *IEEE Trans. Spreech Audio Processing*, vol. 1, no. 4, pp. 405–413, Apr. 1993.

[7] P. Comon, "Blind separation of sources: Problem statement", *Signal Processing*, vol. 24, no. 1, pp. 11–20, 1991.

[8] P. Comon, "Independent Comonent Analysis, a new concept?", *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[9] G. Giannakis, Y. Inouye, and J.M. Mendel, "Cumulant based identification of multichannel MA models", *IEEE Trans. Automatic Contr.*, vol. 34, no. 7, pp. 783–787, July 1989.

[10] S. Shamsunder and G.B. Giannakis, "Modeling of non-Gaussian array data using cumulants: DOA estimation of more sources with less sensors", *Signal Processing*, vol. 30, no. 3, pp. 279–297, 1993.

[11] J.-F. Cardoso, "Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem", in *Proc. ICASSP 89*, 1989, pp. 2109–2112.

[12] C. Jutten and J. Herault, "Blind separation of sources part I: An adaptive algorithm based on neuromimetic architecture", *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[13] D.T. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach", in *Proc. EUSIPCO*, 1992, pp. 771–774.

[14] A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution", *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[15] B. Pearlmutter and L. Parra, "Maximum Likelihood Source Separation: a context-sensitive generalization of ICA", in *Proc. NIPS 9*, 1997, pp. 613–619.

[16] J.F. Cardoso, "Infomax and maximum likelihood for blind source separation", *Signal Processing Lett.*, vol. 4, no. 4, pp. 112–114, Apr. 1997.

[17] S. Amari, A. Cichocki, and Yang A.A., "A new learning algorithm for blind signal separation", in *Proc. NIPS 95*, 1996, pp. 752–763.

[18] D. Yellin and E. Weinstein, "Multichannel Signal Separation: Methods and Analysis", *IEEE Trans. Signal Processing*, vol. 44, no. 1, pp. 106–118, 1996.

[19] H.-L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures", *Signal Processing*, vol. 45, no. 2, pp. 209–229, 1995.

[20] S. Shamsunder and G. Giannakis, "Multichannel Blind Signal Separation and Reconstruction", *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 515–528, Nov. 1997.

[21] R. Lambert and A. Bell, "Blind Separation of Multiple Speakers in a Multipath Environment", in *Proc. ICASSP 97*, 1997, pp. 423–426.

[22] T. Lee, A. Bell, and R. Lambert, "Blind separation of delayed and convolved sources", in *Proc. NIPS 96*, 1997.

[23] S. Amari, S.C. Douglas, A. Cichocki, and A.A. Yang, "Multichannel blind deconvolution using the natural gradient", in *Proc. 1st IEEE Workshop on Signal Processing App. Wireless Comm.*, 1997, pp. 101–104.

[24] L. Parra, "Temporal Models in Blind Source Separation", in *Adaptive Processing of Sequences and Data Structures*, L. Giles and M. Gori, Eds., Lecture Notes in Computer Science, pp. 229–247. Springer, 1998.

[25] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals", *Neural Networks*, vol. 8, no. 3, pp. 411–420, 1995.

[26] A. Soulourmiac, "Blind source detection and separation using second order non-stationarity", in *Proc. ICASSP 95*, 1995, vol. 3, pp. 1912–1915.

[27] F. Ehlers and H.G. Schuster, "Blind Separation for Convolutive Mixtures and an Application in Automatic Speech Recognition in a Noisy Environment", *IEEE Trans. Signal Processing*, vol. 45, no. 10, pp. 2608–2612, Oct. 1997.

[28] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations", *Phys. Rev. Lett.*, vol. 72, no. 23, pp. 3634–3637, 1994.

[29] L. Parra, C. Spence, and B. De Vries, "Convolutive Blind Source Separation based on Multiple Decorrelation", in *IEEE Workshop on Neural Networks and Signal Processing*, Cambridge, UK, September 1998, also presented at "Machines that Learn" Workshop, Snowbird, April 1998.

[30] H. Wee and J. Principe, "A criterion for BSS based on simultaneous diagonalization of time correlation matrices", in *Proc. IEEE Workshop NNSP'97*, 1997, pp. 496–508.

[31] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals", *IEEE Trans. Signal Processing*, submitted.

[32] M. Kawamoto, "A method of blind separation for convolved non-stationary signals", *Neurocomputing*, vol. 22, no. 1-3, pp. 157–171, 1998.

[33] C. Jutten, *Calcul neuromimetique et traitment du signal: Analyse en composantes independantes*, PhD thesis, UJF-INP Grenoble, 1987.

[34] V. Capdevielle, C. Serviere, and J.L. Lacoume, "Blind separation of wide-band sources in the frequency domain", in *Proc. ICASSP 95*, 1995, pp. 2080–2083.

[35] D Brandwood, "A complex gradient operator and its application in adaptive array theory", *IEE Proc.*, vol. 130, no. 1, pp. 11–16, Feb. 1983.

[36] K. Jänich, *Einführung in die Funktionentheorie*, Springer-Verlag, 1977, ch. 2.

[37] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of Blind Signal Separation Methods", in *Proceedings Int. Workshop Independent Component Analysis and Blind Signal Separation*, J.F. Cardoso, Ch. Jutten, and Ph. Loubaton, Eds., Aussois, France, January 11-15 1999, pp. 261–266.

[38] P.M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room", *J. Acoust. Soc. Am.*, vol. 80, no. 5, pp. 1527–1529, Nov. 1986.