CCIT Report #382 May 2002

Noise Cancellation with Static Mixtures of a Nonstationary Signal and Stationary Noise

Sharon Gannot and Arie Yeredor

EE faculty, Technion
Technion City, Haifa 32000, Israel
phone: +972-4-8294756
fax: +972-4-8323041
gannot@siglab.technion.ac.il
http://www-sipl.technion.ac.il/∼gannot/

Dept. of EE-Systems, Tel-Aviv University
P.O.Box 39040, Tel-Aviv 69978, Israel
phone: +972-3-6405314
fax: +972-3-6407095
arie@eng.tau.ac.il
http://www.eng.tau.ac.il/∼arie/

Abstract

We address the problem of cancelling a stationary noise component from its static mixtures with a nonstationary signal of interest. This problem can be treated from two different perspectives, using different processing tools, both based on second-order statistics. The first approach is the Blind Source Separation (BSS) approach, which is aimed at estimating the mixing parameters via approximate joint diagonalization of estimated correlation matrices. Proper exploitation of the nonstationary nature of the desired signal in contrast to the stationarity of the noise signal, allows special parameterization of the joint diagonalization problem in terms of a nonlinear weighted least squares (WLS) problem. The second approach is a denoising approach, which translates into direct estimation of just one of the mixing coefficients via solution of a linear WLS problem, followed by the use of this coefficient to create a noise-only signal to be properly eliminated from the mixture. Under certain assumptions, the BSS approach is asymptotically optimal, yet computationally more intense than the suboptimal denoising approach, since it involves an iterative solution of a nonlinear WLS, whereas the latter only requires a closed-form linear LS solution. We analyze and compare the performance of the two approaches, and provide some simulation results which confirm our analysis.

I. Introduction

In many applications in signal processing and communications a desired signal is contaminated by some unknown, statistically independent, noise signal. Multisensor arrays are often used for the purpose of separating, or denoising, the desired signal. Each sensor receives a linear combination of the desired signal and noise, so that by properly combining the received signals, enhancement of the desired signal is possible.

This problem can be regarded either as a denoising or as a blind source separation (BSS) problem. The difference between these two approaches lies with the treatment of the noise signal: while the former regards the noise merely as a disturbance, the latter regards it as another source signal to be separated from the desired one.

A major practical difference between the two approaches to this problem lies in their computational complexity: while the BSS approach involves approximate joint diagonalization, which amounts to the solution of a nonlinear Weighted Least Squares (WLS) problem, the Denoising approach only requires the solution of a linear WLS problem. It is therefore interesting to compare the performance of the two approaches, in order to gauge the benefit of using the computationally more intense BSS approach.

In order to attain the desired noise cancellation, some special characteristics of the signals and/or the mixing have to be exploited. Traditionally, the BSS approach is only based on statistical independence of the sources. However, in several contexts (e.g., [1], [2], [3]) second-order statistics are sufficient. One such context is the framework of nonstationarity. The key property to be employed in this paper is the assumption that the desired signal is nonstationary, whereas the noise signal is stationary. This assumption holds in several situations of interest, e.g., when the desired signal is a speech signal and some stationary noise signal (such as fan noise) is present.

The mixing that links the source signals to the sensors is usually assumed to be linear and time-invariant (LTI). In its more general form, it consists of different (unknown) LTI systems relating each source signal to each sensor. However, a more degenerate case of an LTI system is a static mixture, in which each sensor receives a memoryless (static) linear combination of the source signals. While the case of static mixtures is not as prevalent in practical situation as the dynamic (convolutive) mixtures case, it has been treated extensively in the context of BSS and Independent Components Analysis (ICA) - see, e.g., [4], [5], [6] for a comprehensive review. In many situations the assumption of a static mixture holds precisely, and in other situations it can be justified as a first-order approximation of a short-memory convolutive system (e.g., in communications applications with narrowband sources or in non-reverberant acoustic situations with closely-spaced directional microphones). The treatment of the static case basically encompasses many of the principles underlying the BSS problem in general, even in the context of convolutive mixtures.

Our purpose in this paper is to present and compare (by analysis and simulations) both the denoising and the separation approaches for the problem of a static mixture of a nonstationary (desired) signal and a stationary (noise) signal.

The problem of BSS in a static mixture of nonstationary signals has recently been treated by Pham and Cardoso in [3], where one proposed method was to apply a special form of joint diagonalization to a set of estimated correlation matrices taken from different segments. It is assumed that the source signals have constant powers within segments, but these powers vary between segments - thus constituting the nonstationarity of the sources. While directly applicable in our problem, this approach cannot exploit the fact that one of the source signals (the noise in our case) is stationary. In the BSS approach we take in this paper, the joint diagonalization problem assumes the form of a WLS problem, in which the parameterization properly exploits the noise stationarity.

Static mixtures in the BSS context were also addressed in [7] by Parra and Spence as a preliminary tool for treatment of the convolutive case. Their model is more general since it also contains uncorrelated additive noise components in each sensor (on top of the signals' mixing). Therefore this model is also over-parameterized for our more concise problem.

In [8] and [9], Rahbar et al. address the case of convolutive mixtures of nonstationary signals, where separation is performed in the frequency domain by applying static source separation to the spectral components at each frequency taken over different segments (and later resolving the scale/permutation ambiguity). Again, exploitation of stationarity of one of the sources is beyond the scope of these contributions (although the extension of the associated diagonalization problems accordingly is possible).

The alternative approach, which regards the separation as a denoising problem was first introduced by Gannot et al. in [10] and analyzed in [11]. It was applied in the convolutive mixture case, and relies on a system-identification method proposed by Shalvi and Weinstein in [12]. This method estimates an LTI system's transfer function by exploiting the nonstationarity of its input signal contrasted with the stationarity of the input/output noise signal. One identification approach in [12] was based on estimated time-domain correlations, while another approach was based on spectral estimates. Only the frequency-domain approach was (approximately) analyzed. However, the degenerate case of a static mixture, which allows exact (small errors) analysis in the time-domain, was not addressed.

This paper is organized as follows. In the next section we provide the problem formulation. In Section III we present the BSS approach and in Section IV we present the denoising approach. While the general approaches in both sections do not make any assumptions on the actual distribution of each source, a small-errors analysis is also provided (for both approaches) for the case of Gaussian, temporally-uncorrelated sources. Based on

these analyses, optimized versions (under the same assumptions) of both approaches are derived. In Section V we present some simulations results comparing the two approaches, as well as showing the agreement with the analytically predicted performance. Some conclusions are drawn in section VI.

II. Problem Formulation

We denote the nonstationary source signal by $s[n]$, and the stationary noise by $v[n]$. The observed signals are $x_1[n]$ and $x_2[n]$:

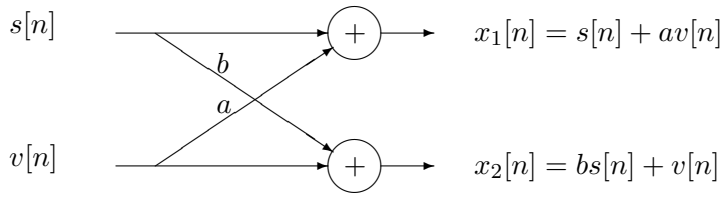

Fig. 1. Static mixing with normalized coefficients

In the blind scenario the scales of neither the source signal nor the noise are known. Therefore, some arbitrary constraints have to be imposed on the mixing coefficients, in order to eliminate the inherent ambiguity involved in the possible commutation of scales between the channel and the signal. We use unity scales in the direct paths, denoting by a and b the two unknown mixing parameters:

$$\begin{aligned} x_1[n] &= s[n] + av[n] \\ x_2[n] &= bs[n] + v[n], \quad n = 1, 2, \ldots N. \end{aligned} \tag{1}$$

The source signal $s[n]$ is assumed to be piece-wise power-stationary, in the following sense: Divide the observation interval into K segments. In each segment, $s[n]$ satisfies

$$\begin{aligned} E[s[n]] &= 0 \\ E[|s[n]|^2] &= \sigma_k^2 \quad N_{k-1} < n \leq N_k \quad k = 1, 2, \ldots K \end{aligned} \tag{2}$$

where

$$\begin{aligned} N_0 &= 0 \\ N_k &= N_{k-1} + L_k \quad k = 1, 2, \ldots K \end{aligned} \tag{3}$$

L_k being the (known) length of the k-th segment (and $N_K = N$). The variances $\sigma_1^2, \sigma_2^2, \ldots \sigma_K^2$ are unknown. Weak ergodicity of $s[n]$ in each segment is assumed.

The noise $v[n]$ is assumed to be zero-mean weakly ergodic WSS, statistically independent of $s[n]$, with unknown variance $\sigma_v^2 = E[|v[n]|^2]$.

It is desired to estimate the source signal $s[n]$ (or a scaled version[1] thereof) from the observations $x_1[n]$ and $x_2[n]$, $n = 1, 2, \ldots N$.

In general, the source signals, as well as the mixing parameters, may be either real-valued or complex-valued. Unfortunately, the real-valued case cannot be regarded as a special case of the complex-valued case, since in the complex-valued case the signals are usually assumed to be circular (see e.g. [13]). A real-valued signal cannot be considered a circular complex-valued signal. While both cases are of interest, the presentation of the real-valued case is considerably more concise. Therefore, in order to capture the essence of the proposed approaches, we shall mainly address the real-valued case, leaving for the appendix the further modifications required to address the complex-valued case.

III. The BSS Approach

In this section we address the denoising problem as a BSS problem, attempting to estimate the mixing parameters explicitly in order to use their estimates for demixing.

Transforming to matrix-vector notation, we define

$$\boldsymbol{M}(a,b) \stackrel{\triangle}{=} \begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix} \tag{4}$$

as the mixing matrix, and $\boldsymbol{x}[n] \stackrel{\triangle}{=} [x_1[n]\ x_2[n]]^T$ as the observation vector.

Since $s[n]$ and $v[n]$ are zero-mean and statistically independent, and are both power-stationary in each segment, the signals $x_1[n]$ and $x_2[n]$ are jointly power-stationary in each segment. Specifically, The zero-lag correlation matrices

$$E\left[\boldsymbol{x}[n]\boldsymbol{x}^T[n]\right] = \boldsymbol{M}(a,b) E\begin{bmatrix} E\left[s^2[n]\right] & 0 \\ 0 & E\left[v^2[n]\right] \end{bmatrix} \boldsymbol{M}^T(a,b) \tag{5}$$

are independent of n within each segment, so that we may define the k-th segment's zero-lag correlation matrix,

$$\boldsymbol{R}_k \stackrel{\triangle}{=} \boldsymbol{M}(a,b) \begin{bmatrix} \sigma_k^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix} \boldsymbol{M}^T(a,b), \quad k = 1, 2, \ldots K. \tag{6}$$

These correlation matrices can be estimated in each segment using straightforward averaging,

$$\hat{\boldsymbol{R}}_k = \frac{1}{L_k} \sum_{n=N_{k-1}+1}^{N_k} \boldsymbol{x}[n]\boldsymbol{x}^T[n] \quad k = 1, 2, \ldots K. \tag{7}$$

[1]Due to the scaling assumption, $s[n]$ can only be recovered up to some (complex) constant scale.

The estimates are unbiased, and moreover, are consistent if the source signal and noise are weakly ergodic within each segment (consistency is per segment, with respect to its length L_k).

A set of K matrices $\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots \boldsymbol{R}_K$ is said to be jointly diagonalized by a matrix $\boldsymbol{M}$ if there exist K diagonal matrices $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots \boldsymbol{\Lambda}_K$ such that $\boldsymbol{R}_k = \boldsymbol{M}\boldsymbol{\Lambda}_k\boldsymbol{M}^T$ for all $k = 1, 2, \ldots K$. Under certain conditions on the $\boldsymbol{\Lambda}_k$-s, the diagonalizing matrix $\boldsymbol{M}$ is unique up to possible scaling and permutation of its columns.

It is evident from (6), that the true correlation matrices $\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots \boldsymbol{R}_K$ are jointly diagonalized by $\boldsymbol{M}(a, b)$. Thus, an estimate of $\boldsymbol{M}(a, b)$ can be obtained by attempting to jointly diagonalize the K estimated correlation matrices $\hat{\boldsymbol{R}}_1, \hat{\boldsymbol{R}}_2, \ldots \hat{\boldsymbol{R}}_K$, which we shall denote the "target matrices". However, if $K > 2$ then it is (almost surely) impossible to attain exact joint diagonalization of these target matrices. We must then resort to approximate joint diagonalization, a concept which has seen extensive use in the field of BSS ([14], [15], [16], [17], [6]) with various selections of sets of "target matrices". Several criteria have been proposed as a measure of the extent of attainable diagonalization, see, e.g., [15], [17], [18], and especially [3] in a context similar to ours.

One possible measure of diagonalization is the straightforward Least-Squares (LS) criterion, which, in our case, assumes the following form:

$$\min_{\hat{a},\hat{b},\hat{\sigma}_v^2,\hat{\sigma}_1^2,\hat{\sigma}_2^2,\ldots\hat{\sigma}_K^2} \sum_{k=1}^{K} \left\{ \left\| \hat{\boldsymbol{R}}_k - \begin{bmatrix} 1 & \hat{a} \\ \hat{b} & 1 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_k^2 & 0 \\ 0 & \hat{\sigma}_v^2 \end{bmatrix} \begin{bmatrix} 1 & \hat{b} \\ \hat{a} & 1 \end{bmatrix} \right\|_F^2 \right\} \tag{8}$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm[2]. Note that the minimization has to be attained with respect to (w.r.t.) the nuisance parameters $\hat{\sigma}_v^2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots \hat{\sigma}_K^2$ (as well as w.r.t. the parameters of interest $\hat{a}, \hat{b}$), since these are additional unknowns.

This formulation differs from the general formulation of standard approximate joint diagonalization problems in two respects: One is the structural constraint on the mixing matrix, which eliminates the scaling and permutation ambiguity by explicitly parameterizing just two degrees of freedom. The other is the constraint on the diagonal matrices, by which the $(2, 2)$ element (namely $\hat{\sigma}_v^2$) must be the same for all k - a direct consequence of the noise's stationarity.

Therefore, with slight manipulations, we prefer to represent this criterion as a standard (nonlinear, possibly weighted) LS problem. First denote, for shorthand, a vector $\hat{\boldsymbol{\theta}} \stackrel{\triangle}{=} [\hat{\sigma}_1^2\ \hat{\sigma}_2^2 \cdots \hat{\sigma}_K^2\ \hat{\sigma}_v^2]^T$ consisting of all nuisance parameters. In addition, define K vectors consisting of the entries of the respective target matrices in vec$\{\cdot\}$ formation:

$$\hat{\boldsymbol{r}}_k \stackrel{\triangle}{=} \text{vec}\{\hat{\boldsymbol{R}}_k\} = [\hat{R}_k^{(1,1)}\ \hat{R}_k^{(2,1)}\ \hat{R}_k^{(1,2)}\ \hat{R}_k^{(2,2)}]^T \quad k = 1, 2, \ldots K. \tag{9}$$

[2]The Frobenius norm of a matrix $\boldsymbol{A}$ is given by $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j A_{i,j}^2 = Trace\{\boldsymbol{A}^T\boldsymbol{A}\}$.

The equivalent vec$\{\cdot\}$ formation of the k-th diagonal form would be

$$\mathrm{vec}\left\{\begin{bmatrix} 1 & \hat{a} \\ \hat{b} & 1 \end{bmatrix}\begin{bmatrix} \hat{\sigma}_k^2 & 0 \\ 0 & \hat{\sigma}_v^2 \end{bmatrix}\begin{bmatrix} 1 & \hat{b} \\ \hat{a} & 1 \end{bmatrix}\right\} = \begin{bmatrix} 1 & \hat{a}^2 \\ \hat{b} & \hat{a} \\ \hat{b} & \hat{a} \\ \hat{b}^2 & 1 \end{bmatrix}\begin{bmatrix} \hat{\sigma}_k^2 \\ \hat{\sigma}_v^2 \end{bmatrix} \quad . \tag{10}$$

Consequently, if we concatenate all $\hat{\boldsymbol{r}}_k$-s into a $4K \times 1$ vector $\hat{\boldsymbol{r}} \stackrel{\triangle}{=} [\hat{\boldsymbol{r}}_1\ \hat{\boldsymbol{r}}_2 \cdots \hat{\boldsymbol{r}}_K]^T$, then the LS criterion (8) can be expressed as

$$\min_{\hat{a},\hat{b},\hat{\boldsymbol{\theta}}}\ [\hat{\boldsymbol{r}} - \tilde{\boldsymbol{G}}(\hat{a},\hat{b})\hat{\boldsymbol{\theta}}]^T[\hat{\boldsymbol{r}} - \tilde{\boldsymbol{G}}(\hat{a},\hat{b})\hat{\boldsymbol{\theta}}], \tag{11}$$

where the $4K \times (K+1)$ matrix $\tilde{\boldsymbol{G}}(\hat{a},\hat{b})$ is given by

$$\tilde{\boldsymbol{G}}(\hat{a},\hat{b}) \stackrel{\triangle}{=} \begin{bmatrix} \tilde{\boldsymbol{b}} & \mathbf{0} & \cdots & \mathbf{0} & \tilde{\boldsymbol{a}} \\ \mathbf{0} & \tilde{\boldsymbol{b}} & & \mathbf{0} & \tilde{\boldsymbol{a}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \tilde{\boldsymbol{b}} & \tilde{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} \otimes \tilde{\boldsymbol{b}} & \mathbf{1} \otimes \tilde{\boldsymbol{a}} \end{bmatrix} \tag{12}$$

with $\tilde{\boldsymbol{b}} = [1\ \hat{b}\ \hat{b}\ \hat{b}^2]^T$, $\tilde{\boldsymbol{a}} = [\hat{a}^2\ \hat{a}\ \hat{a}\ 1]^T$ and $\boldsymbol{I}$, $\mathbf{1}$ and $\mathbf{0}$ as the $K \times K$ identity matrix, a $K \times 1$ all-ones vector and a 4×1 all-zeros vector, respectively. $\otimes$ denotes Kronecker's product.

The concatenation of the K vectors $\hat{\boldsymbol{r}}_k$ would normally comprise the entire "measurements vector" for the LS formulation. However, since $\hat{\boldsymbol{R}}_k$ is symmetric, the second and third elements of each $\hat{\boldsymbol{r}}_k$ are identical, and hence one of them is redundant. To mitigate this redundancy, we define reduced "measurement vectors" $\boldsymbol{y}_k$

$$\boldsymbol{y}_k \stackrel{\triangle}{=} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\hat{\boldsymbol{r}}_k \stackrel{\triangle}{=} \boldsymbol{D}\hat{\boldsymbol{r}}_k \quad k = 1, 2, \ldots K, \tag{13}$$

which we concatenate to form $\boldsymbol{y} = [\boldsymbol{y}_1\ \boldsymbol{y}_2 \cdots \boldsymbol{y}_K]^T$. Adding an arbitrary weight matrix $\boldsymbol{W}$, the WLS criterion becomes

$$\min_{\hat{a},\hat{b},\hat{\boldsymbol{\theta}}}\ [\boldsymbol{y} - \boldsymbol{G}(\hat{a},\hat{b})\hat{\boldsymbol{\theta}}]^T\boldsymbol{W}[\boldsymbol{y} - \boldsymbol{G}(\hat{a},\hat{b})\hat{\boldsymbol{\theta}}], \tag{14}$$

where $\boldsymbol{G}(\hat{a},\hat{b})$ is structured like $\tilde{\boldsymbol{G}}(\hat{a},\hat{b})$,

$$\boldsymbol{G}(\hat{a},\hat{b}) = \left[\boldsymbol{I} \otimes \boldsymbol{b} \quad \mathbf{1} \otimes \boldsymbol{a}\right], \tag{15}$$

this time with

$$\begin{aligned} \boldsymbol{b} &= [\ 1 \quad b \quad b^2\ ]^T \\ \boldsymbol{a} &= [\ a^2 \quad a \quad 1\ ]^T. \end{aligned} \tag{16}$$

Note that this criterion coincides with the criterion in (8) when $\boldsymbol{W} = \mathrm{diag}\{1\ 2\ 1\ 1\ 2\ 1 \cdots 1\ 2\ 1\}$. However, any symmetrical positive definite matrix can be used, and we shall pursue the optimal weight matrix in the sequel.

A. Nonlinear LS Solution

While linear in $\hat{\boldsymbol{\theta}}$, this WLS criterion is nonlinear in $\hat{a}$ and $\hat{b}$. As a minimization approach, we propose to use "alternating coordinates minimization" (ACM) in the following form. Assuming $\hat{a}$ and $\hat{b}$ are fixed, minimization w.r.t. $\hat{\boldsymbol{\theta}}$ is readily attained by the linear WLS solution,

$$\hat{\boldsymbol{\theta}} = \left[\boldsymbol{G}^T(\hat{a},\hat{b})\boldsymbol{W}\boldsymbol{G}(\hat{a},\hat{b})\right]^{-1}\boldsymbol{G}^T(\hat{a},\hat{b})\boldsymbol{W}\boldsymbol{y}. \tag{17}$$

Assuming that $\hat{\boldsymbol{\theta}}$ is fixed, we may take Gauss' method (see e.g. [19]) to solve the nonlinear problem in terms of $\hat{a}$ and $\hat{b}$. Define $\boldsymbol{H}(\hat{a},\hat{b},\hat{\boldsymbol{\theta}})$ to be the following derivative matrix:

$$\boldsymbol{H}(\hat{a},\hat{b},\hat{\boldsymbol{\theta}}) \stackrel{\triangle}{=} \left[\begin{array}{cc} \frac{\partial}{\partial\hat{a}}\{\boldsymbol{G}(\hat{a},\hat{b})\hat{\boldsymbol{\theta}}\} & \frac{\partial}{\partial\hat{b}}\{\boldsymbol{G}(\hat{a},\hat{b})\hat{\boldsymbol{\theta}}\} \end{array}\right] = \left[\begin{array}{cc} \hat{\sigma}_v^2\mathbf{1}\otimes\left[\begin{array}{c} 2\hat{a} \\ 1 \\ 0 \end{array}\right] & \bar{\hat{\boldsymbol{\theta}}}\otimes\left[\begin{array}{c} 0 \\ 1 \\ 2\hat{b} \end{array}\right] \end{array}\right] \tag{18}$$

where $\bar{\hat{\boldsymbol{\theta}}} = [\hat{\sigma}_1^2\ \hat{\sigma}_2^2 \cdots \hat{\sigma}_K^2]^T$ is comprised of the first K elements of $\hat{\boldsymbol{\theta}}$. Gauss' method iteratively updates the estimates $\hat{a}$ and $\hat{b}$ via

$$\begin{aligned} \left[\begin{array}{c} \hat{a}^{[l+1]} \\ \hat{b}^{[l+1]} \end{array}\right] &= \left[\begin{array}{c} \hat{a}^{[l]} \\ \hat{b}^{[l]} \end{array}\right] \\ &+ \left[\boldsymbol{H}^T(\hat{a}^{[l]},\hat{b}^{[l]},\hat{\boldsymbol{\theta}})\boldsymbol{W}\boldsymbol{H}(\hat{a}^{[l]},\hat{b}^{[l]},\hat{\boldsymbol{\theta}})\right]^{-1}\boldsymbol{H}^T(\hat{a}^{[l]},\hat{b}^{[l]},\hat{\boldsymbol{\theta}})\boldsymbol{W}\left[\boldsymbol{y}-\boldsymbol{G}(\hat{a}^{[l]},\hat{b}^{[l]})\hat{\boldsymbol{\theta}}\right] \end{aligned} \tag{19}$$

where $\hat{a}^{[l]}$ and $\hat{b}^{[l]}$ are the l-th iteration values of $\hat{a}$ and $\hat{b}$, respectively.

A "true" ACM algorithm would alternate between minimization of the LS criterion w.r.t. $\hat{\boldsymbol{\theta}}$ assuming $\hat{a}$ and $\hat{b}$ are fixed, and full minimization w.r.t. $\hat{a}$ and $\hat{b}$ assuming $\hat{\boldsymbol{\theta}}$ is fixed. However, these full minimizations may require a large number of inner (Gauss) iterations per each outer (ACM) iteration. In an attempt to speed up the iterative process, it may be desirable to interlace minimizations w.r.t. $\hat{\boldsymbol{\theta}}$ between Gauss iterations. Thus, each Gauss iteration (19) would be preceded with re-estimation of $\hat{\boldsymbol{\theta}}$ using (17).

In a "true" ACM algorithm, the WLS criterion is guaranteed not to increase (usually to decrease) in each iteration. Being bounded below, this property guarantees convergence of the WLS criterion, which, under some reasonable assumptions (see, e.g., [17]), implies convergence of the parameters. Since the criterion is fully minimized w.r.t. either $\hat{\boldsymbol{\theta}}$ or $\hat{a},\hat{b}$ in each iteration, the point of convergence must be a minimum both

w.r.t. $\hat{\boldsymbol{\theta}}$ and w.r.t. $\hat{a}, \hat{b}$. However, it may happen that this point would not be a minimum with respect to $\hat{a}$, $\hat{b}$ and $\hat{\boldsymbol{\theta}}$ simultaneously.

In the "interlaced" ACM algorithm, the WLS criterion is guaranteed not to increase in each application of (17), but not (in general) in each application of a Gauss iteration (19). Nevertheless, under a "small errors assumption", each Gauss iteration solves a linearized WLS problem in the vicinity of a true minimum, thus the nonlinear WLS criterion is decreased as well.

In order to justify such a "small errors assumption", a reasonable initial guess for the parameters has to be used. A possible choice for $\hat{a}^{[0]}$ and $\hat{b}^{[0]}$ can be computed from the (exact) joint diagonalization of any two matrices of the set $\hat{\boldsymbol{R}}_1, \hat{\boldsymbol{R}}_2, \ldots \hat{\boldsymbol{R}}_K$, say $\hat{\boldsymbol{R}}_1$ and $\hat{\boldsymbol{R}}_2$. Since these estimated correlation matrices are symmetric and positive definite, there exist some $\hat{\boldsymbol{M}}$, $\hat{\boldsymbol{\Lambda}}_1$ and $\hat{\boldsymbol{\Lambda}}_2$ that satisfy

$$\hat{\boldsymbol{R}}_1 = \hat{\boldsymbol{M}}\hat{\boldsymbol{\Lambda}}_1\hat{\boldsymbol{M}}^T \qquad\qquad \hat{\boldsymbol{R}}_2 = \hat{\boldsymbol{M}}\hat{\boldsymbol{\Lambda}}_2\hat{\boldsymbol{M}}^T.$$

so that

$$\hat{\boldsymbol{R}}_1\hat{\boldsymbol{R}}_2^{-1} = \hat{\boldsymbol{M}}\left(\hat{\boldsymbol{\Lambda}}_1\hat{\boldsymbol{\Lambda}}_2^{-1}\right)\hat{\boldsymbol{M}}^{-1} \tag{20}$$

meaning that $\hat{\boldsymbol{M}}$ is the eigenvectors matrix of $\hat{\boldsymbol{R}}_1\hat{\boldsymbol{R}}_2^{-1}$ (with eigenvalues given by the diagonal values of $\hat{\boldsymbol{\Lambda}}_1\hat{\boldsymbol{\Lambda}}_2^{-1}$). Thus, initial guesses for $\hat{a}$ and $\hat{b}$ can be obtained from this eigenvectors matrix using proper normalization. The permutation ambiguity can be resolved by ordering the eigenvalues such that the $(2,2)$ element of the eigenvalues matrix be the nearest to unity among the two (reflecting the nominal requirement $\hat{\boldsymbol{\Lambda}}_1^{(2,2)} = \hat{\boldsymbol{\Lambda}}_2^{(2,2)} = \sigma_v^2$).

The minimization algorithm therefore assumes the following form:

Initialization:

Find the eigenvalues λ_1 and λ_2 and corresponding eigenvectors $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ (respectively) of $\hat{\boldsymbol{R}}_1\hat{\boldsymbol{R}}_2^{-1}$, arranged such that λ_2 is the nearest to 1;
Let $\hat{a}^{[0]} = m_{1,2}/m_{2,2}$ and $\hat{b}^{[0]} = m_{2,1}/m_{1,1}$

where $m_{i,j}$ denotes the i-th element of $\boldsymbol{m}_j$, $i,j = 1,2$.

Iterations:

For $l = 0, 1, \ldots$ repeat until convergence:

I. Minimize w.r.t. $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}}^{[l]} = \left[\boldsymbol{G}^T(\hat{a}^{[l]},\hat{b}^{[l]})\boldsymbol{W}\boldsymbol{G}(\hat{a}^{[l]},\hat{b}^{[l]})\right]^{-1}\boldsymbol{G}^T(\hat{a}^{[l]},\hat{b}^{[l]})\boldsymbol{W}\boldsymbol{y}$$

II. Apply one Gauss iteration:

$$\begin{bmatrix}\hat{a}^{[l+1]}\\ \hat{b}^{[l+1]}\end{bmatrix} = \begin{bmatrix}\hat{a}^{[l]}\\ \hat{b}^{[l]}\end{bmatrix} + \left[\boldsymbol{H}^T(\hat{a}^{[l]},\hat{b}^{[l]},\hat{\boldsymbol{\theta}}^{[l]})\boldsymbol{W}\boldsymbol{H}(\hat{a}^{[l]},\hat{b}^{[l]},\hat{\boldsymbol{\theta}}^{[l]})\right]^{-1}\boldsymbol{H}^T(\hat{a}^{[l]},\hat{b}^{[l]},\hat{\boldsymbol{\theta}}^{[l]})\boldsymbol{W}\left[\boldsymbol{y} - \boldsymbol{G}(\hat{a}^{[l]},\hat{b}^{[l]})\hat{\boldsymbol{\theta}}^{[l]}\right]$$

(where the matrices $\boldsymbol{G}(\hat{a},\hat{b})$ and $\boldsymbol{H}(\hat{a},\hat{b},\hat{\boldsymbol{\theta}})$ are given by (15) and (18), respectively).

A reasonable convergence criterion would be to monitor the norm of all the parameters' update in each iteration and compare to a small threshold.

Once a and b are estimated, the demixing matrix can be constructed, and the source (and noise) process(es) estimated:

$$\begin{aligned}\hat{s}[n] &= \frac{x_1[n] - \hat{a}x_2[n]}{1-\hat{a}\hat{b}} \\ \hat{v}[n] &= \frac{x_2[n] - \hat{b}x_1[n]}{1-\hat{a}\hat{b}}\end{aligned} \tag{21}$$

B. Performance Analysis and Optimal Weighting

When some statistical knowledge regarding the source and the noise processes is available, a small-errors performance analysis can be derived, and, moreover, an optimal (or an asymptotically optimal) weight matrix $\boldsymbol{W}$ can be found. A key step in the analysis would be to obtain the covariance matrix of the "measurements" $\boldsymbol{y}$.

To this end, we will now use a statistical model consisting of the following additional assumptions (in addition to the assumptions stated in section II):

- Both the source and the noise are Gaussian processes;
- All non-zero-lag correlations of both processes are zero,

$$E[s[n]s[n-l]] = E[v[n]v[n-l]] = 0 \quad \forall n, \forall l \neq 0.$$

These additional assumptions imply statistical independence between observation signals $\boldsymbol{x}[n]$ belonging to different segments. This statistical independence implies, in turn, zero covariance between the estimates of correlation matrices from two different segments. We therefore need only the covariance between the elements of the estimated $\hat{\boldsymbol{R}}_k$ for each k (segment). By exploiting the Gaussianity and the in-segment whiteness of both signals we obtain

$$\begin{aligned} E[\hat{\boldsymbol{R}}_k^{(i,j)}\hat{\boldsymbol{R}}_k^{(p,q)}] &= \frac{1}{L_k^2}\sum_n\sum_m E\left[x_i[n]x_j[n]x_p[m]x_q[m]\right] \\ &= \frac{1}{L_k^2}\sum_n\sum_m \{E\left[x_i[n]x_j[n]\right]E\left[x_p[m]x_q[m]\right] + E\left[x_i[n]x_p[m]\right]E\left[x_j[n]x_q[m]\right] \\ &\quad + E\left[x_i[n]x_q[m]\right]E\left[x_j[n]x_p[m]\right]\} \\ &= \boldsymbol{R}_k^{(i,j)}\boldsymbol{R}_k^{(p,q)} + \frac{1}{L_k}\left\{\boldsymbol{R}_k^{(i,p)}\boldsymbol{R}_k^{(j,q)} + \boldsymbol{R}_k^{(i,q)}\boldsymbol{R}_k^{(j,p)}\right\} \quad i,j,p,q = 1,2. \end{aligned} \tag{22}$$

Since the first term on the last row equals $E[\hat{\boldsymbol{R}}_k^{(i,j)}]E[\hat{\boldsymbol{R}}_k^{(p,q)}]$, the remaining term equals the desired covariance. Consequently, the entire covariance matrix (per segment k) can be written in matrix form as follows:

$$\boldsymbol{C}_{r,k} \stackrel{\triangle}{=} \mathrm{cov}(\hat{\boldsymbol{r}}_k) = \frac{1}{L_k}(\boldsymbol{R}_k \otimes \boldsymbol{R}_k)\cdot\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}. \tag{23}$$

The covariance matrix of the "measurements" $\boldsymbol{y}_k$ is then given by

$$\boldsymbol{C}_{y,k} = \boldsymbol{D}\boldsymbol{C}_{r,k}\boldsymbol{D}^T \tag{24}$$

where $\boldsymbol{D}$ was defined via (13). Finally, the covariance matrix of the entire "measurements" vector is given by

$$\boldsymbol{C}_y = \mathrm{diag}\{\boldsymbol{C}_{y,1}, \boldsymbol{C}_{y,2}, \ldots \boldsymbol{C}_{y,K}\} \tag{25}$$

where $\mathrm{diag}\{\cdot\}$ is in the matrices-to-matrix block-diagonal sense.

With $\boldsymbol{C}_y$ in hand, we can now proceed to analyze the error in estimating a and b, and the consequent denoising performance. It is well-known that under the "small errors assumption", the nonlinear-WLS estimates are unbiased, and their covariance can be calculated as follows. Define $\boldsymbol{\phi} \stackrel{\triangle}{=} [a\ b\ \hat{\boldsymbol{\theta}}^T]^T$ as the complete vector of unknown parameters, and

$$\boldsymbol{F}(\boldsymbol{\phi}) \stackrel{\triangle}{=} \frac{\partial}{\partial\boldsymbol{\phi}}\{\boldsymbol{G}(a,b)\hat{\boldsymbol{\theta}}\} = [\boldsymbol{H}(a,b,\hat{\boldsymbol{\theta}})\ \ \boldsymbol{G}(a,b)] \tag{26}$$

as the complete derivative matrix. Then

$$\boldsymbol{C}_{\hat{\phi}} \stackrel{\triangle}{=} \mathrm{cov}\{\hat{\boldsymbol{\phi}}\} = \left[\boldsymbol{F}^T(\boldsymbol{\phi})\boldsymbol{W}\boldsymbol{F}(\boldsymbol{\phi})\right]^{-1}\left[\boldsymbol{F}^T(\boldsymbol{\phi})\boldsymbol{W}\boldsymbol{C}_y\boldsymbol{W}\boldsymbol{F}(\boldsymbol{\phi})\right]\left[\boldsymbol{F}^T(\boldsymbol{\phi})\boldsymbol{W}\boldsymbol{F}(\boldsymbol{\phi})\right]^{-1}. \tag{27}$$

The covariance matrix of $\hat{a}$ and $\hat{b}$ is then given by the upper-left 2×2 matrix of $\boldsymbol{C}_{\hat{\phi}}$. Specifically, define as σ_a^2 the $(1,1)$ element of this matrix.

When the estimated demixing matrix is applied to the observed signals, the entire (residual) mixing is given by

$$\frac{1}{1-\hat{a}\hat{b}}\begin{bmatrix} 1 & -\hat{a} \\ -\hat{b} & 1 \end{bmatrix}\begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix} = \frac{1}{1-\hat{a}\hat{b}}\begin{bmatrix} 1-\hat{a}b & a-\hat{a} \\ b-\hat{b} & 1-a\hat{b} \end{bmatrix} \tag{28}$$

such that the denoised signal is given by

$$\hat{s}[n] = \frac{1-\hat{a}b}{1-\hat{a}\hat{b}}s[n] + \frac{a-\hat{a}}{1-\hat{a}\hat{b}}v[n] \stackrel{\triangle}{=} \alpha s[n] + \epsilon v[n]. \tag{29}$$

The residual Interference to Signal ratio (ISR) is usually defined as the expected value of the power of the residual noise coefficient ϵ, normalized by the power of the signal coefficient α. Under the "small error assumption", and assuming further that the true mixing matrix is well-conditioned (the product ab is far from unity), it can be deduced that $\alpha \approx 1$, and

$$\begin{aligned} E[\epsilon] &\approx 0 \\ E[\epsilon^2] &\approx \frac{\sigma_a^2}{(1-ab)^2}\ , \end{aligned} \tag{30}$$

so that $\mathrm{ISR} = E[\epsilon^2/\alpha^2] \approx \sigma_a^2/(1-ab)^2$.

When such a statistical model is in effect, it becomes relatively straightforward to use the optimal weight matrix, which is well-known ([19]) to be given by $\boldsymbol{W}_{opt} = \boldsymbol{C}_y^{-1}$. However, since the true correlation matrices are unknown, the estimated matrices can be used in (23), yielding a sub-optimal weight matrix. Nevertheless, due to the ergodicity of the source and the noise processes, the "estimated" weight is asymptotically optimal ("asymptotically" means here that the number of segments is fixed and their lengths all tend to infinity). The optimality here is in the sense of the resulting mean squared error in estimating a and b, which translates directly into the ISR.

Note, in addition, that when $\boldsymbol{W}_{opt}$ is used, the expression in (27) reduces to $\left[\boldsymbol{F}^T(\boldsymbol{\phi})\boldsymbol{W}_{opt}\boldsymbol{F}(\boldsymbol{\phi})\right]^{-1}$.

IV. Denoising Approach

The BSS approach presented so far is approximately optimal (under several assumptions), but involves an iterative solution of a nonlinear LS problem. We will now derive a different approach, which only involves a

linear LS solution. A comparison between the two methods would be presented in Section V. This solution addresses the noise cancellation problem as a denoising problem, attempting to cancel out noise terms in the first signal, $x_1[n]$. Again the nonstationarity of the desired signal $s[n]$ is exploited in concert with the stationarity of the noise $v[n]$.

A. Algorithm derivation

To get an estimate of the desired signal we first define a noise-only reference signal, $u[n]$,

$$u[n] \stackrel{\triangle}{=} x_2[n] - bx_1[n] = bs[n] + v[n] - b(s[n] + av[n]) = (1 - ab)v[n]. \tag{31}$$

Obviously, $u[n]$ is unavailable since b is unknown. We shall therefore replace b with its estimate $\hat{b}$. The procedure for estimating b will be discussed in the sequel. However, assuming for now that $u[n]$ is available, an estimate of the desired signal $s[n]$ can be obtained by fixing the coefficient h in the following expression:

$$\hat{s}[n] = x_1[n] - hu[n]$$

such that the power of $\hat{s}[n]$ is minimized. This dwells on the fact that $s[n]$ is uncorrelated with $v[n]$ (and hence with $u[n]$). Let the output power be defined by,

$$E[\hat{s}^2[n]] = E[(x_1[n] - hu[n])^2] = E[x_1^2[n] - 2hx_1[n]u[n] + h^2u^2[n]].$$

So,

$$\frac{\partial}{\partial h}E[\hat{s}^2[n]] = 0 \Rightarrow h = \frac{E[x_1[n]u[n]]}{E[u^2[n]]} \stackrel{\triangle}{=} \frac{r_{x_1u}}{r_{uu}}.$$

Since r_{x_1u} and r_{uu} are not directly available, we will express them using the input signals' correlations.

$$\begin{aligned} r_{uu} &= E[u^2[n]] = r_{x_2x_2} - 2br_{x_1x_2} + b^2r_{x_1x_1} \\ r_{x_1u} &= E[x_1[n]u[n]] = r_{x_1x_2} - br_{x_1x_1}. \end{aligned}$$

Using (31) we note that, indeed, if $r_{x_1x_1}$, $r_{x_1x_2}$ and $r_{x_2x_2}$ are known, then

$$\begin{aligned} r_{uu} &= (1 - ab)^2\sigma_v^2 \\ r_{x_1u} &= a(1 - ab)\sigma_v^2 \end{aligned}$$

yielding, $h = \frac{a}{1-ab}$, resulting in $\hat{s}[n] = x_1[n] - \frac{a}{1-ab}v[n] = s[n]$. However, since in practice the cross- and auto-correlations are not known, we should use their estimated values instead,

$$\hat{h} = \frac{\hat{r}_{x_1u}}{\hat{r}_{uu}} = \frac{\hat{r}_{x_1x_2} - \hat{b}\hat{r}_{x_1x_1}}{\hat{r}_{x_2x_2} - 2\hat{b}\hat{r}_{x_1x_2} + \hat{b}^2\hat{r}_{x_1x_1}} \tag{32}$$

where $\hat{r}_{x_1x_1} = \frac{1}{N}\sum_n x_1^2[n]$, $\hat{r}_{x_1x_2} = \frac{1}{N}\sum_n x_1[n]x_2[n]$ and $\hat{r}_{x_2x_2} = \frac{1}{N}\sum_n x_2^2[n]$ are the correlation estimates (at lag zero) taken over the entire observation interval. Zero-lag correlations are sufficient due to the static mixture framework.

When estimates $\hat{h}$ and $\hat{b}$ are used (for h and b respectively), the estimated signal is given by:

$$\begin{aligned}
\hat{s}[n] &= x_1[n] - \hat{h}\hat{u}[n] \\
&= x_1[n] - \hat{h}(x_2[n] - \hat{b}x_1[n]) \\
&= s[n] + av[n] - \hat{h}\left(bs[n] + v[n] - \hat{b}s[n] - \hat{b}av[n]\right) \\
&= \left(1 - \hat{h}(b - \hat{b})\right)s[n] + \left(a - \hat{h}(1 - a\hat{b})\right)v[n] \\
&\triangleq \tilde{\alpha}s[n] + \tilde{\epsilon}v[n].
\end{aligned} \tag{33}$$

The first additive term is (a scaled version of) the desired signal, and the second term is a residual noise term. This expression is similar in structure to (29). However in (29) direct estimates, $\hat{a}, \hat{b}$ of both mixing parameters (a, b respectively) were used, whereas in (34) a is not estimated directly. Instead an external parameter h is introduced and estimated.

We now turn to the estimation of b. To this end, we shall exploit the nonstationarity of $s[n]$ and stationarity of $v[n]$. Rewrite (31) describing $x_2[n]$ as a scaled noisy version of $x_1[n]$,

$$x_2[n] = bx_1[n] + u[n] \tag{34}$$

with $u[n]$ a noise-only term. Given $x_1[n]$ and $x_2[n]$, it is desired to estimate b. If the noise reference signal $u[n]$ were uncorrelated with $x_1[n]$, then a standard system identification estimate, $\hat{b} = \frac{\hat{r}_{x_2x_1}}{\hat{r}_{x_1x_1}}$, could be used to obtain a consistent estimate of b. Unfortunately, by (31), $u[n]$ and $x_1[n]$ are in general correlated, which would cause this estimate to be biased and inconsistent. The bias effect can be mitigated by introducing an extra unknown parameter. To do so, we divide the observations $x_1[n]$, $x_2[n]$ into the segments introduced in

(3). Thus, for the k-th segment we obtain,

$$\begin{aligned}
\hat{r}_{x_2x_1}^{(k)} &\triangleq \frac{1}{L_k}\sum_{n=N_{k-1}+1}^{N_k} x_2[n]x_1[n] \qquad (35)\\
&= \frac{1}{L_k}\sum_{n=N_{k-1}+1}^{N_k} \left(bx_1[n]+u[n]\right)x_1[n]\\
&= \left(\frac{1}{L_k}\sum_{n=N_{k-1}+1}^{N_k} x_1^2[n]\right)b + \frac{1}{L_k}\sum_{n=N_{k-1}+1}^{N_k} u[n]x_1[n]\\
&\triangleq \hat{r}_{x_1x_1}^{(k)}b + \hat{r}_{ux_1}^{(k)}\\
&= \hat{r}_{x_1x_1}^{(k)}b + r_{ux_1} + \epsilon_{ux_1}^{(k)}, \quad k=1,\ldots,K
\end{aligned}$$

where, $\hat{r}_{x2x1}^{(k)}$,$\hat{r}_{ux1}^{(k)}$ and $\hat{r}_{x1x1}^{(k)}$ are (the k-th segment's) consistent correlation estimates (at lag zero), and $\epsilon_{ux_1}^{(k)} \triangleq \hat{r}_{ux_1}^{(k)} - r_{ux_1}$ is the zero mean error in estimating $r_{ux_1} = E\left[u[n]x_1[n]\right]$.

Concatenating (35) for $k = 1, 2, \ldots, K$ we obtain in matrix form:

$$\begin{pmatrix} \hat{r}_{x_2x_1}^{(1)} \\ \hat{r}_{x_2x_1}^{(2)} \\ \vdots \\ \hat{r}_{x_2x_1}^{(K)} \end{pmatrix} = \begin{bmatrix} \hat{r}_{x_1x_1}^{(1)} & 1 \\ \hat{r}_{x_1x_1}^{(2)} & 1 \\ \vdots & \vdots \\ \hat{r}_{x_1x_1}^{(K)} & 1 \end{bmatrix} \begin{pmatrix} b \\ r_{ux_1} \end{pmatrix} + \begin{pmatrix} \epsilon_{ux_1}^{(1)} \\ \epsilon_{ux_1}^{(2)} \\ \vdots \\ \epsilon_{ux_1}^{(K)} \end{pmatrix}.$$

or in short form:

$$\mathbf{z} = \boldsymbol{Q}\boldsymbol{\eta} + \mathbf{e}.$$

Treating (36) as an LS problem in the parameter $\boldsymbol{\eta}$, with $\mathbf{e}$ a zero-mean "noise" vector, the WLS estimate of $\boldsymbol{\eta}$ is given by,

$$\hat{\boldsymbol{\eta}} = (\boldsymbol{Q}^T\boldsymbol{W}\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{W}\mathbf{z} \qquad (36)$$

where $\boldsymbol{W}$ is a possible weight matrix. The desired estimate of b is given by the first term of $\boldsymbol{\eta}$, the second term could be regarded as a nuisance parameter. Choosing an asymptotically optimal weight matrix $\boldsymbol{W}_{opt}$ will be discussed in subsection IV-B.

Summarizing the algorithm,

I. Estimate the correlations $\hat{r}_{x_1x_2}^{(k)}$, $\hat{r}_{x_1x_1}^{(k)}$ for all segments $k = 1, \ldots, K$.
II. Solve the (possibly weighted) LS problem:

$$\begin{pmatrix} \hat{r}_{x_2x_1}^{(1)} \\ \hat{r}_{x_2x_1}^{(2)} \\ \vdots \\ \hat{r}_{x_2x_1}^{(K)} \end{pmatrix} = \begin{bmatrix} \hat{r}_{x_1x_1}^{(1)} & 1 \\ \hat{r}_{x_1x_1}^{(2)} & 1 \\ \vdots & \vdots \\ \hat{r}_{x_1x_1}^{(K)} & 1 \end{bmatrix} \begin{pmatrix} b \\ r_{ux_1} \end{pmatrix} + \begin{pmatrix} \epsilon_{ux_1}^{(1)} \\ \epsilon_{ux_1}^{(2)} \\ \vdots \\ \epsilon_{ux_1}^{(K)} \end{pmatrix}.$$

III. Define the reference noise signal $\hat{u}[n] = x_2[n] - \hat{b}x_1[n]$.
IV. Estimate the correlations $\hat{r}_{x_1x_1}$, $\hat{r}_{x_1x_2}$ and $\hat{r}_{x_2x_2}$.
V. Calculate coefficient $\hat{h} = \dfrac{\hat{r}_{x_1x_2} - \hat{b}\hat{r}_{x_1x_1}}{\hat{r}_{x_2x_2} - 2\hat{b}\hat{r}_{x_1x_2} + \hat{b}^2\hat{r}_{x_1x_1}}$.
VI. Reconstruct the signal $\hat{s}[n] = x_1[n] - \hat{h}\hat{u}[n]$.

B. Performance Analysis and Optimal Weighting

In this section we analyze the expected performance of the suggested denoising algorithm. Using a small error analysis we can write:

$$\begin{aligned} \hat{b} &= b + \epsilon_b \\ \hat{h} &= h + \epsilon_h \end{aligned} \tag{37}$$

where ϵ_b and ϵ_h are zero-mean "small" random variables, such that $|\epsilon_b| \ll |b|$ and $|\epsilon_h| \ll |h|$. Using (34), the residual error is given by $\tilde{\epsilon}v[n]$, where

$$\begin{aligned} \tilde{\epsilon} &= a - \hat{h}(1 - a\hat{b}) \\ &= a - (h + \epsilon_h)(1 - a(b + \epsilon_b)) \\ &= a - h - \epsilon_h + abh + ah\epsilon_b + ab\epsilon_h + a\epsilon_h\epsilon_b. \end{aligned}$$

Neglecting the second order error term $\epsilon_h\epsilon_b$ and using $h = \frac{a}{1-ab}$ we obtain

$$\tilde{\epsilon} = \frac{a^2}{1 - ab}\epsilon_b - (1 - ab)\epsilon_h. \tag{38}$$

The scaling error in (33) is given by

$$1 - \tilde{\alpha} = \hat{h}\left(b - \hat{b}\right) = -\left(h + \epsilon_h\right)\epsilon_b \approx \frac{a}{1 - ab}\epsilon_b \tag{39}$$

where in the last transition we neglected again the second-order error term, $\epsilon_h\epsilon_b$.

Thus, in order to calculate the residual error energy and the scaling distortion, we need to calculate the second-order statistics of ϵ_b and ϵ_h. Since all the error terms in the analysis are due to errors in estimating the

input signals' correlations, we will now define the relations between these segment-wise errors and the error terms of interest.

We now reemploy the additional assumptions of Section III-B namely, that both the signal $s[n]$ and the noise $v[n]$ are Gaussian, temporally uncorrelated. Consequently, the covariance of the k-th segment's estimation error vector $\boldsymbol{\epsilon}^{(k)} \stackrel{\triangle}{=} \left[\epsilon_{x_1x_1}^{(k)} \epsilon_{x_1x_2}^{(k)} \epsilon_{x_2x_2}^{(k)}\right]^T$; $k = 1, 2, \ldots, K$(which equals the covariance of $\boldsymbol{y}_k$ of (13)) is given by $\boldsymbol{C}_{y,k}$ of (24), and the covariance of the augmented vector $\boldsymbol{\epsilon} \stackrel{\triangle}{=} \left[\boldsymbol{\epsilon}^{(1)^T}\ \boldsymbol{\epsilon}^{(2)^T}\ \cdots\ \boldsymbol{\epsilon}^{(K)^T}\right]^T$ is given by $\boldsymbol{C}_y$ of (25).

The error in estimating $\boldsymbol{\eta} = [b\ \ r_{ux1}]^T$ using the LS solution (36) is given by:

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = (\boldsymbol{Q}^T\boldsymbol{W}\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{W}\mathbf{e} \stackrel{\triangle}{=} \left[\begin{array}{c} \mathbf{q}^T \\ \cdots \end{array}\right]\mathbf{e}.$$

Thus, the error term in estimating b is given by this vector's first element, namely:

$$\epsilon_b = \hat{b} - b = \mathbf{q}^T\mathbf{e} = \sum_{k=1}^{K} q_k \epsilon_{ux_1}^{(k)} = \sum_{k=1}^{K} q_k \left(\epsilon_{x_1x_2}^{(k)} - b\epsilon_{x_1x_1}^{(k)}\right) \tag{40}$$

where $q_1, \ldots q_K$ are the elements of $\mathbf{q}$.

Further define,

$$\boldsymbol{A} = \left[\begin{array}{cccccccccc} \frac{L_1}{N} & 0 & 0 & \frac{L_2}{N} & 0 & 0 & \ldots & \frac{L_K}{N} & 0 & 0 \\ 0 & \frac{L_1}{N} & 0 & 0 & \frac{L_2}{N} & 0 & \ldots & 0 & \frac{L_K}{N} & 0 \\ 0 & 0 & \frac{L_1}{N} & 0 & 0 & \frac{L_2}{N} & \ldots & 0 & 0 & \frac{L_K}{N} \\ -bq_1 & q_1 & 0 & -bq_2 & q_2 & 0 & \ldots & -bq_K & q_K & 0 \end{array}\right] \tag{41}$$

and let $\epsilon_{x_1x_1}$, $\epsilon_{x_1x_2}$ and $\epsilon_{x_2x_2}$ denote the errors in estimating the complete (over the entire observation interval) signals' correlations. Then the covariance error of the vector,

$$\boldsymbol{\varepsilon} \stackrel{\triangle}{=} \left[\begin{array}{cccc} \epsilon_{x_1x_1} & \epsilon_{x_1x_2} & \epsilon_{x_2x_2} & \epsilon_b \end{array}\right]^T = \boldsymbol{A}\boldsymbol{\epsilon}$$

is given by $\boldsymbol{C}_{\boldsymbol{\varepsilon}} = \boldsymbol{A}\boldsymbol{C}_y\boldsymbol{A}^T$. Now the error term ϵ_h can be calculated by the following derivation, where for

brevity we replaced $\hat{r}_{x_m x_n}, r_{x_m x_n};\ n,m=1,2$ with $\hat{r}_{mn}, r_{mn}$ (respectively):

$$\begin{aligned}
\hat{h} &= \frac{\hat{r}_{12}-\hat{b}\hat{r}_{11}}{\hat{r}_{22}-2\hat{b}\hat{r}_{12}+\hat{b}^2\hat{r}_{11}} \\
&= \frac{r_{12}+\epsilon_{12}-(b+\epsilon_b)(r_{11}+\epsilon_{11})}{r_{22}+\epsilon_{22}-2(b+\epsilon_b)(r_{12}+\epsilon_{12})+(b+\epsilon_b)^2(r_{11}+\epsilon_{11})} \\
&\approx \frac{r_{12}-br_{11}+\boldsymbol{\beta}_1^T\boldsymbol{\varepsilon}}{r_{22}-2br_{12}+b^2r_{11}+\boldsymbol{\beta}_2^T\boldsymbol{\varepsilon}} \\
&\approx \frac{r_{12}-br_{11}+\boldsymbol{\beta}_1^T\boldsymbol{\varepsilon}}{r_{22}-2br_{12}+b^2r_{11}}\left(1-\frac{\boldsymbol{\beta}_2^T\boldsymbol{\varepsilon}}{r_{22}-2br_{12}+b^2r_{11}}\right) \\
&= \left(h+\frac{\boldsymbol{\beta}_1^T\boldsymbol{\epsilon}}{r_{22}-2br_{12}+b^2r_{11}}\right)\left(1-\frac{\boldsymbol{\beta}_2^T\boldsymbol{\epsilon}}{r_{22}-2br_{12}+b^2r_{11}}\right) \\
&\approx h+\frac{(\boldsymbol{\beta}_1-h\boldsymbol{\beta}_2)^T\boldsymbol{\epsilon}}{r_{22}-2br_{12}+b^2r_{11}},
\end{aligned}$$

where $\boldsymbol{\beta}_1^T=[-b\ \ 1\ \ 0\ \ -r_{11}]$ and $\boldsymbol{\beta}_2^T=\left[b^2\ \ -2b\ \ 1\ \ 2(-r_{12}+br_{11})\right]$, neglecting second and higher order terms in all approximations. Consequently, we identify

$$\epsilon_h \approx \boldsymbol{\beta}^T\boldsymbol{\varepsilon}. \tag{42}$$

with

$$\boldsymbol{\beta}=\frac{(\boldsymbol{\beta}_1-h\boldsymbol{\beta}_2)^T}{r_{22}-2br_{12}+b^2r_{11}}.$$

Using (38),

$$\tilde{\epsilon}=ah\epsilon_b-(1-ab)\epsilon_h=[0\ \ 0\ \ 0\ \ (ah)]\,\boldsymbol{\epsilon}-(1-ab)\boldsymbol{\beta}^T\boldsymbol{\epsilon}\stackrel{\triangle}{=}\boldsymbol{\gamma}^T\boldsymbol{\epsilon}.$$

Then the ISR is defined by

$$\text{ISR}=E\{\tilde{\epsilon}^2\}=\boldsymbol{\gamma}^T\boldsymbol{C}_{\boldsymbol{\epsilon}}\boldsymbol{\gamma}.$$

As we did in the BSS context, we may, under the same statistical assumption, employ an asymptotically optimal weight matrix in the WLS problem (36). Using the identity $\epsilon_{ux_1}^{(k)}=\epsilon_{x_1x_2}^{(k)}-b\epsilon_{x_1x_1}^{(k)}$, we can obtain the optimal weight matrix

$$\boldsymbol{W}=\left(\text{diag}\left\{\text{Var}\left(\epsilon_{ux_1}^{(1)}\right),\text{Var}\left(\epsilon_{ux_1}^{(2)}\right),\ldots,\text{Var}\left(\epsilon_{ux_1}^{(K)}\right)\right\}\right)^{-1} \tag{43}$$

where

$$\text{Var}\left(\epsilon_{ux_1}^{(k)}\right)=\boldsymbol{\delta}^T\boldsymbol{C}_{y,k}\boldsymbol{\delta} \tag{44}$$

with $\boldsymbol{\delta}^T = [-b \;\; 1 \;\; 0]$. Since the true correlation terms are unknown, the estimated terms can be used instead. Note that this also requires an estimate $\hat{b}$ of b. Thus, in order to use the optimal weighting matrix we may first estimate b using (36) with $\boldsymbol{W} = \boldsymbol{I}$ (the identity matrix) and then use (43) to obtain the (asymptotically) optimal $\boldsymbol{W}$. Note that, as in the BSS approach, this procedure requires reasonably "good" estimates in order for the estimated $\boldsymbol{W}$ to be close to the true optimal weight. Recall further, that this weight matrix is only optimal under the assumption that the source and noise signals are Gaussian, temporally uncorrelated. When this is not the case, the algorithm can still be applied using either $\boldsymbol{W} = \boldsymbol{I}$ or any other properly calculated weight matrix.

V. Performance Evaluation and Comparison

In this section we compare the performance of the two approaches, both analytically and empirically. The setup used is as follows. All signals involved are temporally uncorrelated zero-mean Gaussian. We use 7 equal-length segments ($L_1 = L_2 = \cdots = L_7 \stackrel{\triangle}{=} L$) with signal powers $\sigma_1^2, \sigma_2^2, ...\sigma_7^2 = 0.1, 10, 2, 8, 4, 2, 0.3$ (respectively), and with unity noise power $\sigma_v^2 = 1$. The true mixing matrix is $\boldsymbol{M} = \left[\begin{smallmatrix} 1 & 0.6 \\ 1.4 & 1 \end{smallmatrix}\right]$.

In Fig. 2 we present analytical and empirical results for three algorithms: The optimally weighted BSS algorithm, the unweighted denoising algorithm and the optimally weighted denoising algorithm. All results are displayed in terms of ISR vs. the entire observation length $N = 7L$. The empirical (simulations) results represent averages over 250 trials each. All algorithms were applied to the same data.

The empirical results are seen to coincide (asymptotically) with the theoretically predicted values. As expected, the computationally more intensive BSS approach outperforms the denoising approach in both its weighted and unweighted versions. However, this advantage is more pronounced at the longer observation lengths. At the shorter lengths the BSS weighting departs from its optimal value and hence the advantage in performance decreases. As for the Denoising approach, its weighted version attains an improvement over the unweighted version.

VI. Conclusion

We presented and compared two approaches for the noise cancellation problem in static mixtures of a nonstationary desired signal and stationary noise. Both approaches are based on second-order statistics. However, the BSS approach requires the solution of a nonlinear WLS problem, whereas the Denoising approach only requires the solution of a linear WLS problem. Consequently, the performance obtained by the BSS approach is superior to that obtained by the Denoising approach.

Both approaches can be extended and applied in the convolutive case, possibly expressing similar tradeoffs between computational complexity and performance.

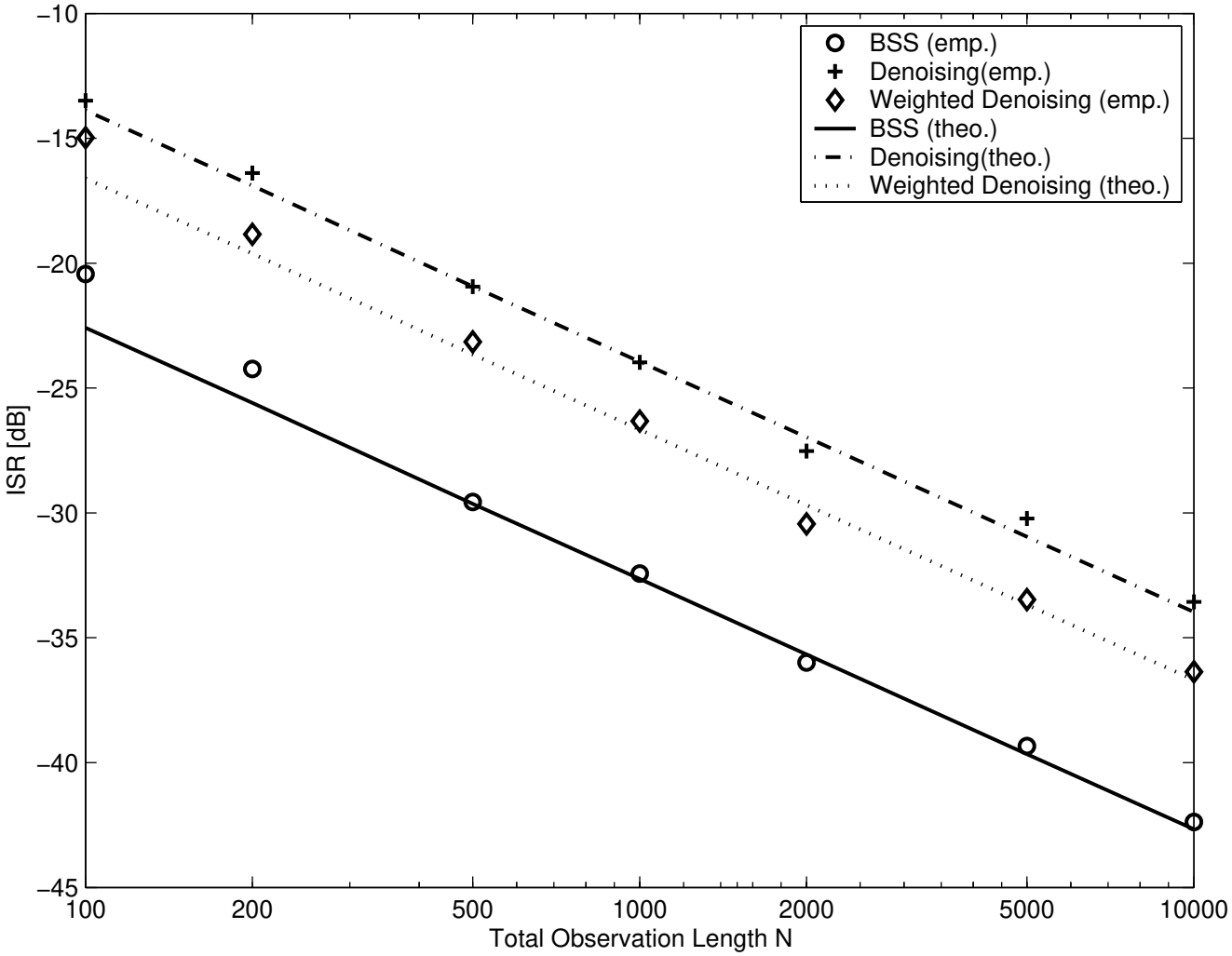

Fig. 2. Empirical and theoretical results for the BSS, Denoising and Weighted Denoising approaches in terms of ISR $[dB]$ vs. the entire observation length N.

Appendix

I. Modifications for the complex-valued case

For the complex-valued case we assume that both the source signal and the noise are complex-valued circular random processes. The circularity property [13], often assumed in the context of complex random processes, implies that $E[s[n]s[m]] = 0$ and $E[v[n]v[m]] = 0\ \forall n, m$. In other words, we have

$$E[v_R[n]v_R[m]] = E[v_I[n]v_I[m]] \tag{45}$$

$$E[v_R[n]v_I[m]] = -E[v_I[n]v_R[m]] \quad \forall n, m \tag{46}$$

where $v_R[n]$ and $v_I[n]$ denote the real and imaginary parts (respectively) of $v[n]$. A similar property holds for $s[n]$ in each segment. Note that this implies that the real and imaginary parts at each time instant n are uncorrelated.

In addition, we assume a properly normalized complex mixing matrix $\boldsymbol{M}(a, b)$ as in (4), with $a = a_R + j \cdot a_I$

and $b = b_R + j \cdot b_I$, so these are now four real-valued parameters of interest, a_R, a_I, b_R and b_I. The other $K+1$ nuisance parameters remain unchanged (since they represent real-valued positive variances).

The modifications to the BSS approach are as follows:

The segmental correlation matrices are now estimated using

$$\hat{\boldsymbol{R}}_k = \frac{1}{L_k} \sum_{n=N_{k-1}+1}^{N_k} \boldsymbol{x}[n]\boldsymbol{x}^H[n] \quad k = 1, 2, \ldots K, \tag{47}$$

where the superscript ^{H} denotes the conjugate-transpose. With $\hat{\boldsymbol{r}}_k = \mathrm{vec}\{\hat{\boldsymbol{R}}_k\}$ and $\boldsymbol{y} = \boldsymbol{D}\hat{\boldsymbol{r}}_k$ defined as in (9) and (13) (resp.), the matrix $\boldsymbol{G}(\hat{a}, \hat{b})$ is still defined as in (15), but now $\boldsymbol{b} = [\ 1 \quad b^* \quad |b|^2\ ]^T$ and $\boldsymbol{a} = [\ |a|^2 \quad a \quad 1\ ]^T$. The matrix $\boldsymbol{H}(\hat{a}, \hat{b}, \hat{\boldsymbol{\theta}})$ of (18) is now defined as

$$\boldsymbol{H}(\hat{a}, \hat{b}, \hat{\boldsymbol{\theta}}) \stackrel{\triangle}{=} \left[\hat{\sigma}_v^2 \mathbf{1} \otimes \begin{bmatrix} 2\hat{a}_R \\ 1 \\ 0 \end{bmatrix} \quad \hat{\sigma}_v^2 \mathbf{1} \otimes \begin{bmatrix} 2\hat{a}_I \\ j \\ 0 \end{bmatrix} \quad \bar{\hat{\boldsymbol{\theta}}} \otimes \begin{bmatrix} 0 \\ 1 \\ 2\hat{b}_R \end{bmatrix} \quad \bar{\hat{\boldsymbol{\theta}}} \otimes \begin{bmatrix} 0 \\ -j \\ 2\hat{b}_I \end{bmatrix} \right]. \tag{48}$$

Therefore, the minimization w.r.t. $\hat{\boldsymbol{\theta}}$ still takes the form of (17), with the ^{T} superscript replaced by ^{H}. However, the Gauss iterations take the augmented form,

$$\begin{bmatrix} \hat{a}_R^{[l+1]} \\ \hat{a}_I^{[l+1]} \\ \hat{b}_R^{[l+1]} \\ \hat{b}_I^{[l+1]} \end{bmatrix} = \begin{bmatrix} \hat{a}_R^{[l]} \\ \hat{a}_I^{[l]} \\ \hat{b}_R^{[l]} \\ \hat{b}_I^{[l]} \end{bmatrix} + \mathrm{Re}\left\{ \boldsymbol{H}^H(\hat{a}^{[l]}, \hat{b}^{[l]}, \hat{\boldsymbol{\theta}}) \boldsymbol{W} \boldsymbol{H}(\hat{a}^{[l]}, \hat{b}^{[l]}, \hat{\boldsymbol{\theta}}) \right\}^{-1} \mathrm{Re}\left\{ \boldsymbol{H}^H(\hat{a}^{[l]}, \hat{b}^{[l]}, \hat{\boldsymbol{\theta}}) \boldsymbol{W} \left[\boldsymbol{y} - \boldsymbol{G}(\hat{a}^{[l]}, \hat{b}^{[l]}) \hat{\boldsymbol{\theta}} \right] \right\}, \tag{49}$$

where $\mathrm{Re}\{\cdot\}$ denotes the real part of the enclosed expression. This is the special form of a linear WLS solution obtained when using complex-valued measurements and model matrix, while constraining the estimated parameters to be real-valued.

As for calculating the optimal weight matrix $\boldsymbol{W}_{opt}$, the only modification is to $\boldsymbol{C}_{r,k}$, which is now given (still under the assumption of complex circular Gaussian, temporally uncorrelated source signal and noise) by

$$\boldsymbol{C}_{r,k} = \frac{1}{L_k} \boldsymbol{R}_k^* \otimes \boldsymbol{R}_k. \tag{50}$$

The matrices $\boldsymbol{C}_{y,k}$, $\boldsymbol{C}_y$ and $\boldsymbol{W}_{opt} = \boldsymbol{C}_y^{-1}$ are automatically updated accordingly.

The modifications to the denoising approach are more simple:

Naturally, all correlations should be estimated using conjugation, as indicated above in (47). The linear LS problem (36) still holds, so the estimation of the complex value of b is still given by (36), but with the ^{T}

superscript replaced by ^{H}. All other procedures, including calculation of the optimal weight in (43), (44) remain unchanged, provided that (50) is used for $\boldsymbol{C}_{r,k}$ in (44).

References

[1] Adel Belouchrani, Karim Abed-Meraim, Jean-François Cardoso, and Eric Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[2] Arie Yeredor, "Blind separation of Gaussian sources via second-order statistics with asymptotically optimal weighting," *IEEE Signal Processing Letters*, vol. 7, no. 7, pp. 197–200, 2000.

[3] Dinh-Tuan Pham and Jean-François Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[4] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[5] Jean-François Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[6] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent Component Analysis*, John Wiley and Sons, Inc., 2001.

[7] Lucas Parra and Clay Spence, "Convolutive blind source separation of nonstationary sources," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 320–327, 2000.

[8] Kamran Rahbar and James P. Reilly, "Blind source separation algorithm for MIMO convolutive mixtures," *Proceedings ICA 2001*, pp. 242–247, 2001, San Diego.

[9] Kamran Rahbar, James P. Reilly, and Jonathan H. Manton, "A frequency-domain approach to blind dentification of MIMO FIR systems driven by quasi-stationary signals," *Proceedings ICASSP 2002*, 2002, Orlando.

[10] Sharon Gannot, David Burshtein, and Ehud Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[11] Sharon Gannot, David Burshtein, and Ehud Weinstein, "Theoretical analysis of the general transfer function GSC," in *The 2001 International Workshop on Acoustic Echo and Noise Control (IWAENC01)*, Darmstadt, Germany, Sep. 2001.

[12] Ofir Shalvi and Ehud Weinstein, "System identification using nonstationary signals," *IEEE trans. Signal Processing*, vol. 44, no. 8, pp. 2055–2063, 1996.

[13] Bernard Picinbono, "On circularity," *IEEE trans. Signal Processing*, vol. 42, no. 12, pp. 3473–3482, 1994.

[14] Jean-François Cardoso and Antoine Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 1, pp. 161–164, 1996.

[15] Arie Yeredor, "Approximate joint diagonalization using non-orthogonal matrices," *Proceedings ICA2000*, pp. 33–38, 2000.

[16] Dinh-Tuan Pham, "Joint approximate diagonalization of positive-definite Hermitian matrices," 1999, Technical Report LMC/IMAG, `http://www-lmc.imag.fr/lmc-sms/Dinh-Tuan.Pham/jadiag/jadiag.ps.gz`.

[17] Arie Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," 2002, to appear in *IEEE Trans. on Signal Processing*.

[18] Mati Wax and Jacob Sheinvald, "A least-squares approach to joint diagonalization," *IEEE Signal Processing Letters*, vol. 4, no. 2, pp. 52–53, 1997.

[19] H.W. Sorenson, *Parameter Estimation*, Marcel-Dekker, 1980.