A NATURAL GRADIENT CONVOLUTIVE BLIND SOURCE SEPARATION ALGORITHM FOR SPEECH MIXTURES

Xiaoan Sun and Scott C. Douglas

Department of Electrical Engineering
Southern Methodist University
Dallas, Texas 75275 USA

ABSTRACT

In this paper, a novel algorithm for separating mixtures of multiple speech signals measured by multiple microphones in a room environment is proposed. The algorithm is a modification of an existing approach for density-based multichannel blind deconvolution using natural gradient adaptation. It employs linear predictors within the coefficient updates and produces separated speech signals whose autocorrelation properties can be arbitrarily specified. Stationary point analyses of the proposed method illustrate that, unlike multichannel blind deconvolution methods, the proposed algorithm maintains the spectral content of the original speech signals in the extracted outputs. Performance comparisons of the proposed method with existing techniques show its desirable properties in separating real-world speech mixtures.

1. INTRODUCTION

In blind source separation (BSS), multiple independent source signals are extracted from their linear mixtures with little to no knowledge of the sources and the mixing system. Numerous applications of BSS have been found in a wide variety of fields, such as antenna arrays for wireless communications, biomedical signal processing, seismic signal processing, passive sonar, and speech processing.

Many algorithms have been proposed for the instantaneous BSS problem in which the mixtures are spatial linear combinations of independent sources [1]. A more challenging problem involves mixtures that are spatio-temporal in nature [2]-[12]. Algorithms for such situations can be classified into two classes. One class of algorithms, termed *multichannel blind deconvolution* methods, attempt to make the system's outputs both spatially and temporally independent. The other class of algorithms, termed *convolutive BSS* methods, attempt to perform separation without specifically deconvolving the system's outputs. The former class of algorithms are only truly appropriate for situations in which the source signals are spatially and temporally independent. Such is not the case for most acoustic source signals.

This paper focuses on the blind separation of speech signal mixtures as measured in a room environment, or the so-called "cocktail party problem." In this case, we assume that there are m talkers and n randomly-located microphones in the environment, where $n \geq m$. Since speech is temporally-correlated, convolutive BSS algorithms are most appropriate for this task. Despite this fact, many researchers have applied multichannel blind deconvolution algorithms to speech separation tasks with moderate success [4, 6, 7]. Such algorithms impose undesirable constraints on the extracted output signals, however, causing their spectra to be nearly equalized. Previously, three solutions to this problem have been proposed:

1. Apply pre- and post-filters to the inputs and outputs, respectively, of the separation system [8].
2. Employ a frequency-domain separation method that applies independent separation systems to each input signal frequency bin [9].
3. Employ a separation method that imposes no constraints on the individual temporal structures of the extracted output signals [12].

The first solution relies on the similarities of the correlation properties of the source signals to work properly. The second solution creates the problem of permuted solutions in the frequency domain that require additional temporal constraints to solve. The third solution can result in slightly-increased reverberation in the extracted signals due to the unconstrained nature of the separation system.

In this paper, we propose a new convolutive BSS algorithm for separating speech signal mixtures. The proposed algorithm modifies the time-domain natural gradient multichannel blind deconvolution method in [5, 7] by incorporating linear prediction filters on each of the extracted output signals. These filters allow the output correlation properties of the extracted signals to match those of the individual speech signals while maintaining the good convergence properties of the natural gradient updates. Numerical evaluations of the method show that it provides excellent performance in real-world speech signal separation for a variety of publicly-available data sets.

2. CONVOLUTIVE BLIND SOURCE SEPARATION

2.1. Problem Formulation

In the convolutive blind source separation (BSS) task, m source signals $\{s_j(k)\}$, $1 \leq j \leq m$, pass through an unknown m-input, n-output linear time-invariant mixing system to yield the n mixed signals $\{x_i(k)\}$. Defining the vectors $\mathbf{s}(k) = [s_1(k) \cdots s_m(k)]^T$ and $\mathbf{x}(k) = [x_1(k) \cdots x_n(k)]^T$,

[0]This material is based in part upon work supported by the Texas Advanced Technology Program under Grant No. 003613-0031-1999.

we can represent the mixing process as

$$\mathbf{x}(k) \quad = \quad \sum_{l=0}^{\infty} \mathbf{A}_l \mathbf{s}(k-l), \tag{1}$$

where $\{\mathbf{A}_l\}$ is a sequence of $(n \times m)$ matrices which is the impulse response of the acoustical environment. The matrix sequence $\{\mathbf{A}_l\}$ is not arbitrary and must satisfy the following weak condition that is typically satisfied in practice:

> The discrete-time Fourier transform $\mathcal{A}(e^{j\omega})$ of the matrix sequence $\{\mathbf{A}_l\}$ must be of rank m for all $|\omega| \leq \pi$.

For this paper, ambient sensor noise is assumed to be negligible.

The goal of the convolutive BSS task is to calculate a demixing system with a causal matrix impulse response $\{\mathbf{B}_l\}$, $0 \leq l \leq \infty$ such that the outputs of this system given in vector form by

$$\mathbf{y}(k) \quad = \quad \sum_{l=0}^{\infty} \mathbf{B}_l \mathbf{x}(k-l) \tag{2}$$

with $\mathbf{y}(k) = [y_1(k) \cdots y_m(k)]^T$ contain estimates of the m source signal sequences in $\{\mathbf{s}(k)\}$ without crosstalk. Without any *a priori* information about the temporal characteristics of the source signals, such solutions have the form

$$y_i(k) \quad = \quad \sum_{l=0}^{\infty} \delta_{ij} \varepsilon_{jjl} s_j(k-l) \tag{3}$$

where δ_{ij} is the Kronnecker impulse function, ε_{jjl} is the impulse response of an unknown filter, and the assignment $j \rightarrow i$ occurs without replacement. If such a solution is attained, each output of the system in $\mathbf{y}(k)$ contains a filtered version of a unique source signal in $\mathbf{s}(k)$.

The ability to solve the convolutive BSS task depends entirely on the characteristics of the source signals $\{s_i(k)\}$. In this paper, we shall employ a strong statistical assumption that is often reasonable in speech separation tasks:

> Each speech signal $s_i(k)$ is statistically independent of every other speech signal $s_j(l)$ for $i \neq j$ and for all k and l.

Algorithms that use this independence assumption to perform separation have shown a reasonable level of success even in situations involving potentially-correlated source signals. In addition, we shall assume that sampled versions of the speech signals have a non-Gaussian amplitude density that approximately follows a Laplacian distribution model given by

$$p_{s_i}(s) \quad = \quad \frac{1}{\sqrt{2}\sigma_i} \exp\left(-\frac{\sqrt{2}|s|}{\sigma_i}\right), \tag{4}$$

where σ_i^2 is the variance of the ith speech signal. This assumption is well-motivated by statistical tests [13], and its accuracy is less critical to the success of the separation methods as a whole.

In the sequel, we shall assume well-behaved mixing conditions that allow for approximate truncated implementations of the causal demixing system, whereby (2) is replaced by

$$\mathbf{y}(k) \quad = \quad \sum_{l=0}^{L} \mathbf{B}_l \mathbf{x}(k-l) \tag{5}$$

and L is a finite integer. Approximating linear systems by truncated models is an oft-used procedure in signal processing tasks, as it allows for practical adaptation procedures for the matrices $\{\mathbf{B}_l\}$ to be developed. Such approximations shall not be discussed further in any great detail.

2.2. Relationship to Multichannel Blind Deconvolution

Convolutive BSS is closely related to multichannel blind deconvolution, a task that we now describe. The mixing and separation systems for multichannel blind deconvolution are identical to those in (1) and (2), respectively. Only the assumptions regarding the source signals are different and are as follows:

> Each signal $s_i(k)$ is statistically-independent of every other signal $s_j(l)$ for both $i \neq j$ and all k and l as well as for $i = j$ and all $k \neq l$.

In other words, the source signals $\{s_i(k)\}$ are independent of each other, and each signal is a sequence of independent samples as well. With such an assumption, it is possible to obtain a stronger separation result than (3), as given by

$$y_i(k) \quad = \quad \delta_{ij} \varepsilon_{jj\Delta_j} s_j(k-\Delta_j) \tag{6}$$

where Δ_j is an integer delay value. If such a solution is attained, each output of the system in $\mathbf{y}(k)$ contains a scaled, delayed version of a unique source signal in $\mathbf{s}(k)$. This solution corresponds to a removal of both spatial and temporal crosstalk in the extracted source signals.

The assumption of temporal independence is reasonable in certain signal processing tasks that yield the mixing model in (1). One such situation is in multiuser wireless communications, where each $s_i(k)$ corresponds to a modulated bit sequence [14]. Temporal independence is clearly not appropriate, however, for convolutive blind separation of speech. Hence, algorithms for multichannel blind deconvolution are not entirely appropriate for separating speech signals, as they typically impose unnatural contraints on the temporal structures of the extracted signals in $\mathbf{y}(k)$. Despite this fact, many researchers have used multichannel blind deconvolution procedures in convolutive BSS tasks with some degree of success [6, 7, 8]. In the next section, we modify such procedure to yield a new algorithm that is better-suited to, and provide better performance in, convolutive BSS tasks.

3. ALGORITHM DERIVATION

3.1. The Natural Gradient Algorithm for Multichannel Blind Deconvolution

The approaches developed in this paper are based on a recently-developed algorithm for multichannel blind deconvolution tasks. This algorithm employs a well-studied and statistically-plausible criterion. It is computationally-simple and has shown success in a number of deconvolution tasks [5, 7]. For these reasons, the algorithm is a good starting point from which to develop speech separation approaches.

This algorithm employs the following density-matching criterion to perform separation:

$$\mathcal{J}(\{\mathbf{B}_l\}) = \\ -\oint \log|\det \mathcal{B}(z)| z^{-1} dz - \sum_{i=1}^{m} E\{\log \widehat{p}_{s_i}(y_i(k))\} \tag{7}$$

where $\mathcal{B}(z)$ is the z-transform of the impulse response sequence $\{\mathbf{B}_l\}$, $\widehat{p}_{s_i}(y_i)$ is a density model for the ith extracted source, and $E\{\cdot\}$ denotes statistical expectation. While many interpretations can be given to this criterion, Pham has shown that this cost function is, up to an additive constant, proportional to the relative entropy of the extracted output sources assuming the amplitude density models $\widehat{p}_{s_i}(y_i)$ [15]. When each $\widehat{p}_{s_i}(y_i)$ matches the actual amplitude density of each source, minimizing this criterion corresponds to minimizing the mutual information of the extracted source signals in $\mathbf{y}(k)$, which in the case of speech mixtures corresponds to separated speech signals.

Having chosen a multichannel deconvolution criterion for minimization, we now describe a procedure for adapting the matrices $\mathbf{B}_l$ to minimize it. While a standard gradient approach could be used, gradient methods are known for their slow convergence properties. In addition, it can be shown [5] that the gradient of the criterion in (7) requires calculating the inverse of the impulse response of the multichannel separation system $\mathcal{B}(z)$, a daunting task. Fortunately, a modification of standard gradient-based procedures has been developed that overcomes these two difficulties in the multichannel blind deconvolution task. This modification, termed the *natural gradient* by Amari [16], modifies the standard gradient update by a linear transformation whose elements are determined by the Riemannian metric tensor for the assumed parameter space. Details regarding this modification can be found in [5] and [17]. The resulting coefficient updates for this natural gradient multichannel blind deconvolution procedure are: for $0 \leq l \leq L$,

$$\mathbf{B}_l(k+1) = \mathbf{B}_l(k) + \mathbf{M}(k)\left[\mathbf{B}_l(k) - \mathbf{f}(\mathbf{y}(k-L))\mathbf{u}^T(k-l)\right] \quad (8)$$

$$\mathbf{y}(k) = \sum_{l=0}^{L} \mathbf{B}_l(k)\mathbf{x}(k-l) \quad (9)$$

$$\mathbf{u}(k) = \sum_{q=0}^{L} \mathbf{B}_{L-q}^T(k)\mathbf{y}(k-q) \quad (10)$$

where $\mathbf{M}(k)$ is a diagonal matrix of positive step sizes $\mu_i(k)$, $1 \leq i \leq m$ and $\mathbf{f}(\mathbf{y}) = [f_1(y) \cdots f_m(y)]^T$ is a vector nonlinearity function. For the cost function in (7), the corresponding forms for each $f_i(y)$ are

$$f_i(y) = -\frac{\partial \log p_{s_i}(y)}{\partial y}. \quad (11)$$

For example, the Laplacian density model in (4) yields the identical nonlinearities $f_i(y) = \mathrm{sgn}(y)$ when source scaling is ignored.

3.2. Temporal Constraints in Convolutive BSS

Multichannel blind deconvolution is a special case of convolutive BSS in which particular temporal constraints are placed on the extracted sources. For this reason, many researchers have employed multichannel blind deconvolution algorithms in convolutive BSS as a first attempt [6, 7, 8]. As we shall show, however, such methods are inappropriate for the speech separation and require some modifications to obtain best performance.

Both multichannel blind deconvolution and convolutive BSS assume source signals that are independent of each other; only the assumptions on the temporal structures of the sources differ. In the former task, the sources are additionally assumed to be independent from sample to sample, such that for any two samples $s_1 = s_i(k)$ and $s_2 = s_i(k-l)$, the joint probability density function (pdf) of s_1 and s_2 is

$$p_{s_1 s_2}(s_1, s_2) = p_{s_1}(s_1)p_{s_2}(s_2), \quad (12)$$

where $p_{s_i}(s_i)$ is the marginal p.d.f. of s_i for $i \in \{1, 2\}$. Multichannel blind deconvolution algorithms attempt to enforce this temporal independence on the extracted output signals $\{y_i(k)\}$. The conditions being enforced can generally be found by analyzing the stationary points of the corresponding algorithm's update equations. A stationary point of an adaptive algorithm is a coefficient solution such that, on average, coefficient values do not change from one update to the next for statistically-stationary signals. A necessary condition for a stationary point to exist in either multichannel blind deconvolution or convolutive BSS algorithms is

$$E\{\mathbf{B}_l(k+1)\} = E\{\mathbf{B}_l(k)\} \quad (13)$$

for all l. Sufficient conditions for stationary points require the analysis of the second-order properties of the algorithm updates; see [11] for examples.

In the case of the natural gradient multichannel blind deconvolution algorithm in (8)–(10), one can determine the conditions on the output signal sequence $\mathbf{y}(k)$ such that (13) holds assuming stationary source signal statistics. The resulting conditions are

$$\mathbf{I}\delta_l - E\{\mathbf{f}(\mathbf{y}(k-L))\mathbf{y}^T(k-L-l)\} = \mathbf{0} \quad (14)$$

which, in scalar form, become

$$E\{f_i(y_i(k))y_j(k-l)\} = \delta_{ij}\delta_l \quad (15)$$

for all $1 \leq \{i, j\} \leq m$ and $-\infty < l < \infty$. Eqn. (15) implies both spatial and temporal statistical independence of the extracted output signals when such signals are non-Gaussian and have zero mean and symmetric p.d.f.'s, so long as $f_i(y)$ is a nonlinear function [12]. If $f_i(y)$ is linear, then (15) implies that the extracted signals are uncorrelated, which is not sufficient to guarantee independence.

In convolutive BSS, one cannot assume that the signals are temporally-independent. Speech signals maintain a correlated temporal structure due both to the acoustic properties of the vocal production system and the quasi-periodic nature of voiced speech sounds. Hence, the temporal conditions imposed by multichannel blind deconvolution algorithms are undesirable and unnatural. Experiments with the natural gradient multichannel blind deconvolution algorithm in (8)–(10) for speech separation indicate that the main artifact imposed by the temporal constraints is an approximate "whitening" or spectral flattening of the extracted speech signals. While such speech is still understandable, it loses much of its realism and cannot be expected to be listenable for long periods. For these reasons, we now explore modifications to that attempt to preserve the temporal characteristics of the source signals in the extracted outputs.

3.3. Convolutive BSS Using Linear Prediction Constraints

Now we propose a novel modification of the natural gradient multichannel blind deconvolution algorithm in (8)–(10) to control the correlation properties of the extracted signals. This modified algorithm employs the following assumption about the speech signals being extracted.

Each speech signal $s_i(k)$ can be approximately modelled as the output of an autoregressive (AR) system driven by a temporally-independent input signal, such that

$$s_i(k) = -\sum_{j=1}^{M} d_{ij} s_i(k-j) + p_i(k), \quad (16)$$

where d_{ij}, $1 \leq j \leq M$ are the coefficients of the ith AR system and $p_i(k)$ is the temporally-independent input sequence driving this system.

This assumption is well-justified by the statistical characteristics of human speech. Speech can be modelled as a temporally-correlated quasi-periodic signal. The autoregressive model in (16) is often used to represent the correlation properties of speech in a number of tasks and speech coding in particular. Thus, processing this speech by an all-zero linear predictor removes a significant portion of the redundancy in speech, one obtains a nearly-independent sequence of random samples [18].

With this assumption, we propose to modify the natural gradient multichannel blind deconvolution algorithm as shown in Figure 1. The separation system is now described by two subsystems, denoted as $\mathcal{W}(z)$ and $\overline{\mathcal{D}}(z)$ in the figure. The first subsystem is a multiple-input, multiple-output system with an FIR matrix impulse response $\mathbf{W}_l(k)$, $0 \leq l \leq N$, whereas the second system consists of m different FIR filters with coefficients $\overline{d}_{il} = d_{jl}$, $0 \leq l \leq M$ with $d_{j0} = 1$, where the mapping $j \rightarrow i$ corresponds to the permutation relationship between the ith source signal $p_i(k)$ and the jth output signal $y_j(k)$ at convergence. Letting $\overline{\mathbf{D}}_l$ be a diagonal matrix whose diagonal entries are the order-permuted filter taps d_{jl}, $1 \leq j \leq m$, we can write the impulse response of the entire separation system as

$$\mathbf{B}_l(k) = \overline{\mathbf{D}}_l * \mathbf{W}_l(k) \quad (17)$$

$$= \sum_{j=0}^{M} \overline{\mathbf{D}}_j \mathbf{W}_{l-j}(k). \quad (18)$$

To develop the algorithm for adjusting each $\mathbf{W}_l(k)$, we employ the multichannel blind deconvolution approach in (8)–(10) to adjust the *combined* system impulse responseas

$$\overline{\mathbf{D}}_l * \mathbf{W}_l(k+1) = \overline{\mathbf{D}}_l * \mathbf{W}_l(k) + \mathbf{M}(k)\big[\overline{\mathbf{D}}_l * \mathbf{W}_l(k) - \mathbf{f}(\mathbf{y}(k-L)) \mathbf{y}^T(k-L-l) * \overline{\mathbf{D}}_l * \mathbf{W}_l(k)\big] \quad (19)$$

$$\mathbf{y}(k) = \overline{\mathbf{D}}_l * \mathbf{W}_l(k) * \mathbf{x}(k-l). \quad (20)$$

Translating this update to the coefficients $\mathbf{W}_l(k)$ of the separation system requires defining the impulse response of the inverse of $\overline{\mathcal{D}}(z)$ as $\overline{\mathbf{D}}_{inv,l}$. Clearly, $\overline{\mathbf{D}}_{inv,l}$ exists if each sequence $\overline{d}_{il}$ corresponds to an FIR linear predictor, as such systems are always minimum phase [18]. Applying this inverse system to both sides of (19) yields

$$\mathbf{W}_l(k+1) = \mathbf{W}_l(k) + \mathbf{M}(k)\left[\mathbf{W}_l(k) - \overline{\mathbf{D}}_{inv,l} * \mathbf{f}(\mathbf{y}(k-L)) \mathbf{y}^T(k-L-l) * \overline{\mathbf{D}}_l * \mathbf{W}_l(k)\right]. \quad (21)$$

Finally, the following approximation proves useful:

$$\overline{\mathbf{D}}_{inv,l} * \mathbf{f}(\mathbf{y}(k-L)) \mathbf{y}^T(k-L-l)$$

$$= \sum_{i=0}^{\infty} \overline{\mathbf{D}}_{inv,i} \mathbf{f}(\mathbf{y}(k-L)) \mathbf{y}^T(k-L-l+i) \quad (22)$$

$$\approx \sum_{i=0}^{\infty} \overline{\mathbf{D}}_{inv,i} \mathbf{f}(\mathbf{y}(k-L-i) \mathbf{y}^T(k-L-l). \quad (23)$$

The above approximation assumes that the extracted speech signals are statistically-stationary over the prediction interval, which is a reasonable assumption. With this approximation, we obtain the coefficient updates as

$$\mathbf{W}_l(k+1) = \mathbf{W}_l(k) + \mathbf{M}(k)\big[\mathbf{W}_l(k) - \mathbf{g}(k-L)\mathbf{u}^T(k-l)\big] \quad (24)$$

$$\mathbf{g}(k) = \mathbf{f}(\mathbf{y}(k-L)) - \sum_{q=1}^{M} \overline{\mathbf{D}}_l \mathbf{g}(k-q) \quad (25)$$

$$\mathbf{u}(k) = \sum_{q=0}^{L} \mathbf{W}_{L-q}(k) \mathbf{y}_D(k-q) \quad (26)$$

$$\mathbf{y}_D(k) = \mathbf{y}(k-L) + \sum_{q=0}^{M-1} \overline{\mathbf{D}}_{M-q} \mathbf{y}(k-q) \quad (27)$$

$$\mathbf{y}(k) = \mathbf{y}_S(k) + \sum_{l=1}^{M} \overline{\mathbf{D}}_l \mathbf{y}_S(k-l) \quad (28)$$

$$\mathbf{y}_S(k) = \sum_{l=0}^{N} \mathbf{W}_l(k) \mathbf{x}(k-l) \quad (29)$$

where (24) uses the result of (23), (25) employs the autoregressive property of the inverse of any linear prediction filter, (27) is a temporary expression used in (26), and $\mathbf{y}(k)$ in (28) and $\mathbf{y}_S(k)$ in (29) are shown in Figure 1. This algorithm is particularly simple, requiring $4mn(N+1) + 3mM + m$ muliply/accoumulates (MACs) at each time step to compute, or approximately four MACs per adaptive filter coefficient when $N \gg M$.

To see why this algorithm is well-suited to speech separation, we return to Figure 1. In this block diagram, $\mathcal{D}_{inv}(z) = \mathcal{D}^{-1}(z)$ is the system function for the speech production model, and $\overline{D}(z)$ is given by

$$\overline{\mathcal{D}}(z) = \mathbf{\Phi} \mathcal{D}(z) \mathbf{\Phi}^T. \quad (30)$$

Thus, the overall system function from the independent sample sequence $\mathbf{p}(k)$ to the system output $\mathbf{y}(k)$ is

$$\mathcal{C}(z) = \mathbf{\Phi} \mathcal{D}(z) \mathbf{\Phi}^T \mathcal{W}(z) \mathcal{A}(z) \mathcal{D}^{-1}(z). \quad (31)$$

The algorithm that we have constructed is, ignoring truncation effects, identical to the multichannel blind deconvolution algorithm in (8)–(10). Thus, at convergence, the combined system function has the approximate form

$$\mathcal{C}(z) \approx \mathbf{\Phi} \mathcal{E}(z), \quad (32)$$

where $\mathcal{E}(z)$ is a diagonal matrix whose diagonal elements are $\varepsilon_{jj\Delta_j} z^{-\Delta_j}$ and $\mathbf{\Phi}$ is a permutation matrix. Combining (31) and (32), we obtain

$$\mathcal{D}(z) \mathbf{\Phi}^T \mathcal{W}(z) \mathcal{A}(z) \mathcal{D}^{-1}(z) = \mathcal{E}(z). \quad (33)$$

Pre- and post-multiplying both sides of the above equation by $\mathbf{\Phi}\mathcal{D}^{-1}(z)$ and $\mathcal{D}(z)$, respectively, gives

$$\mathcal{W}(z)\mathcal{A}(z) = \mathbf{\Phi}\mathcal{D}^{-1}(z)\mathcal{E}(z)\mathcal{D}(z) \quad (34)$$
$$= \mathbf{\Phi}\mathcal{E}(z), \quad (35)$$

where the last simplification follows from the diagonal natures of $\mathcal{D}^{-1}(z)$, $\mathcal{E}(z)$, and $\mathcal{D}(z)$. In other words, the outputs in $\mathbf{y}_S(k)$ are exactly the speech signals in $\mathbf{s}(k)$, up to scaling, order permutation, and arbitrary delay factors. This solution is exactly what is desired; the speech signals remain separated but not deconvolved.

4. SIMULATION RESULTS

This section compares the separation results of various convolutive BSS algorithms in several real-world acoustic BSS tasks. The real-world signal mixtures used for these evaluations have been taken from data set made available by several authors [8, 9, 10]. Three of the examples have been recorded at a f_s = 16kHz sampling rate representing wideband speech, whereas one example was recorded at a f_s = 12kHz sampling rate.

We compare the performances of the multichannel blind deconvolution (MBD) algorithm in (8), the nonholonomic convolutive BSS (NH-CBSS) algorithm in [12], and the linear-prediction-based convolutive BSS (LP-CBSS) algorithm in (24)–(29), in which all updates have been implemented in block form using FFT-based fast convolution methods. Each block-based update calculates the convolution terms using L-sample sums, so that all filter coefficients are updated every L time instants. Similar block-based methods have been used in other adaptive filtering tasks [19]. To calculate the linear predictor coefficients for the LP-CBSS algorithm, we used the `lpc` command in MATLAB as applied to each of the measured mixtures. This choice assumes that the spectral content of each mixture is largely controlled by a dominant speech signal that is then extracted at that particular system output. Successive refinements of the linear predictor coefficients could have been calculated during adaptation but were avoided for complexity reasons.

As for other algorithm parameters, the diagonal entries of $\mathbf{M}(k)$ were chosen as

$$\mu_i(k) = \frac{\mu_0}{L\left(\beta + \sum_{p=L+1}^{2L} y_i(k-p)f(y_i(k-p))\right)} \quad (36)$$

where the constant $\beta = 0.01$ was used to avoid a divide-by-zero condition. This "quasi-normalized" step size strategy makes the updates scale-independent and generally improves algorithm robustness. For all separation tasks, the chosen filter lengths were $L = N = 1024$ and $M = 50$. Several passes through each recorded segment–between 30 and 50–were allowed in order to achieve convergence of every algorithm, with step sizes chosen $\mu_0 = 1$. Detailed convergence rate studies are the subject of current efforts.

To evaluate each algorithm's performance, we employed a strategy based on identifiable signal content. First, temporal portions of each recording were identified that only contained a single source, and the variances of these portions were computed. Let $\widehat{\sigma}_{ij}^2$ correspond to the estimated variance of source i in signal j from these portions. Next, portions of each signal containing no sources were found to estimate the noise powers $\widehat{\rho}_j$ for each channel. Then, the signal-to-interference ratio (SIR) and signal-to-noise ratio (SNR) for the jth channel were computed as

$$\mathrm{SIR}_j = \frac{\widehat{\sigma}_{jj}^2 - \widehat{\rho}_j}{\widehat{\sigma}_{ij}^2 - \widehat{\rho}_j}, \quad \mathrm{SNR}_j = \frac{\widehat{\sigma}_{jj}^2 - \widehat{\rho}_j}{\widehat{\rho}_j}, \quad (37)$$

where $i \neq j$. In situations where one signal is persistently-exciting and the other signal is intermittent, we only give SIR and SNR values for the intermittent signal. In addition, we compute the power spectral densities (PSDs) of the original and extracted signals to gauge the temporal effects that each algorithm imposes on the extracted signals.

Table 1 lists the SIRs and signal-to-noise ratios (SNRs) for the original and separated signals for four different two-channel signal separation tasks. For comparison, we also calculate the SIRs and SNRs produced by each authors' proposed approach. Comparing these results, we see that the multichannel blind deconvolution algorithm generally does not provide the best separation results, and it can fail to provide any reasonable amount of separation. The nonholonomic and linear-prediction-based convolutive BSS methods generally provide much better performance. In fact, the LP-CBSS algorithm generally provided the best performance in terms of both SIR and SNR, even as compared to each author's proposed method.

To gauge the listenability of each separation result, we calculated the power spectral densities (PSDs) of the input and output signals from the various algorithms. Shown in Figure 2 are the corresponding PSD curves for the Lee Number example. As can be seen, the MBD algorithm tends to "flatten" the spectral content of the measured signals in the separated outputs, making the resulting signals sound unnatural. By contrast, both the NH-CBSS and LP-CBSS algorithms largely maintain the spectral contents of the original signal mixtures. Moreover, the LP-CBSS algorithm generally provided the best signal-to-noise ratios of all methods, such that any environmental noise was not significantly enhanced in the separated system's outputs.

5. CONCLUSIONS

In this paper, we propose a novel convolutive BSS algorithm that is specifically designed to separate mixtures of speech signals as measured by multiple sensors. Employing linear prediction filters within the adaptive process, we effectively translate an existing multichannel blind deconvolution algorithm based on information-theory to the convolutive BSS task under a widely-assumed model for speech production. Numerical evaluations indicate the abilities of the approaches to separate two-channel speech signal mixtures recorded in real-world environments.

REFERENCES

[1] S. Haykin, ed. *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation*, John Wiley & Sons, 2000.

[2] E. Weinstein, M. Feder, and A.V. Oppenheim, "Multichannel signal separation by decorrelation," *IEEE Trans. Signal Processing,* vol. 1, pp. 405-413, July 1993.

[3] S. Van Gerven and D. Van Compernolle, "Signal separation by symmetric adaptive decorrelation: Stability, convergence, and uniqueness," *IEEE Trans. Signal Processing,* vol. 43, pp. 1602-1612, July 1995.

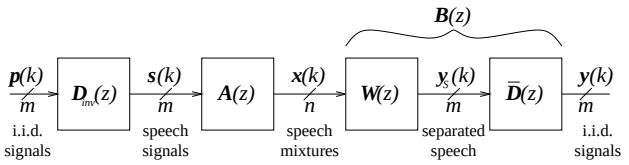

Figure 1. Block diagram of speech separation system using autoregressive models for speech producti

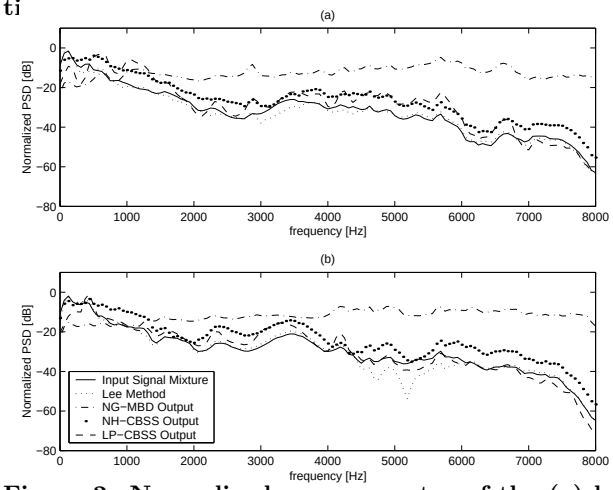

Figure 2. Normalized power spectra of the (a) left and (b)right signal channels in the Lee Number example.

Table 1. Signal-to-interference and signal-to-noise ratios in the numerical examples.

	Original Mixture	Author's Outputs	NG-MBD Outputs	NH-CBSS Outputs	LP-CBSS Outputs
Lee Number					
SIR (Left)	0.5 dB	21.0 dB	15.1 dB	16.9 dB	20.3 dB
SIR (Right)	0.9 dB	6.0 dB	10.5 dB	3.1 dB	9.4 dB
SNR (Left)	24.1 dB	21.0 dB	19.9 dB	25.2 dB	26.9 dB
SNR (Right)	27.2 dB	25.3 dB	19.8 dB	19.8 dB	27.4 dB
Lee News					
SIR (Left)	5.1 dB	12.8 dB	17.9 dB	14.1 dB	17.8 dB
SIR (Right)	0.9 dB	10.9 dB	1.8 dB	6.5 dB	9.0 dB
SNR (Left)	21.4 dB	21.4 dB	22.0 dB	17.0 dB	22.9 dB
SNR (Right)	18.3 dB	16.2 dB	16.7 dB	15.4 dB	14.8 dB
Parra					
SIR (Left)	3.0 dB	9.3 dB	8.0 dB	9.2 dB	11.9 dB
Anemuller					
SIR (Left)	1.3 dB	12.4 dB	6.2 dB	14.5 dB	14.6 dB

[4] R.H. Lambert, "Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures," Ph.D. dissertation, Univ. Southern California, Los Angeles, CA, May 1996.

[5] S. Amari, S. Douglas, A. Cichocki, H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," *Proc. 1st IEEE Workshop Signal Processing Adv. Wireless Commun.,* Paris, France, pp. 101–104, Apr. 1997.

[6] R.H. Lambert and A.J. Bell, "Blind separation of multiple speakers in a multipath environment," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Munich, Germany, vol. 1, pp. 423-426, Apr. 1997.

[7] S. Amari, S.C. Douglas, A. Cichocki, and H.H. Yang, "Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach," *Proc. 11th IFAC Symp. Syst. Ident.,* Kitakyushu City, Japan, vol. 3, pp. 1057-1062, July 1997.

[8] T.-W. Lee, A. Ziehe, R. Orglmeister and T. J. Sejnowski, "Combining time-delayed decorrelation and ICA: towards solving the cocktail party problem," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Seattle, WA., pp. 1089–1092, May 1998.

[9] L. Parra, C. Spence, and B. De Vries, "Convolutive blind source separation based on multiple decorrelation," *Proc. IEEE Workshop on Neural Networks Signal Processing,* Cambridge, UK, pp. 23–32, Sept. 1998.

[10] J. Anemuller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," *Proc. 2nd IEEE Int. Workshop Indep. Compon. Anal. Signal Separation,* Helsinki, Finland, pp. 215–220, June 2000.

[11] X. Sun and S. C. Douglas, "Multichannel blind deconvolution of arbitrary signals: adaptive algorithms and stability analyses," *Proc. 34th Asilomar Conf. Signals, Syst., Comput.,* Pacific Grove, CA, vol. 2, pp. 1412-1416, Nov. 2000.

[12] S.C. Douglas, "Blind signal separation and blind deconvolution." In *Handbook of Neural Networks Signal Processing,* Y.-H. Hu and J.-N. Hwang, eds. (Boca Raton, FL: CRC Press, 2001), Chap. 7.

[13] W. Davenport, Jr., "A study of speech probability distributions," Tech. Rep. 148, MIT Research Laboratory of Electronics, Cambridge, MA, 1950.

[14] A.J. Paulraj and C.B. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Mag.,* vol. 14, no. 6, pp. 49-83, Nov. 1997.

[15] D.-T. Pham, "Mutual information approach to blind separation of stationary sources," *Proc. 1st Workshop Indep. Compon. Anal. Signal Separation,* Aussois, France, pp. 215-220, Jan. 1999.

[16] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251-276, 1998.

[17] S.C. Douglas and S. Amari, "Natural gradient adaptation." In *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation,* S. Haykin, ed. (New York: Wiley, 2000), pp. 13–61.

[18] J. D. Gibson, "Speech Signal Processing." In *The Electrical Engineering Handbook,* ed. R. C. Dorf (Boca Raton, FL: CRC Press, 1993), pp. 279-314.

[19] J.J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Mag.,* vol. 9, pp. 14-37, Jan. 1992.