A Causal Frequency-Domain Implementation of a Natural Gradient Multichannel Blind Deconvolution and Source Separation Algorithm

Scott C. Douglas[1], *Hiroshi Sawada* [2], *and Shoji Makino* [2]

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

[2]NTT Communications Science Laboratories, NTT Corporation, Kyoto 619-0327 JAPAN

Abstract

Natural gradient adaptation is useful for developing block-based adaptive solutions to convolutive blind source separation tasks, in which frequency-domain fast convolution methods can be exploited for computational simplicity. To maintain causal operation, previously-developed natural gradient convolutive blind source separation algorithms employ approximations to the natural gradient that ultimately limit the separation performances of these schemes. In this paper, we derive a novel causal frequency-domain implementation of a natural gradient algorithm for convolutive blind source separation tasks. Simulations with convolutive speech mixtures indicate that the proposed method provides robust convergence performance for relatively-short separation filters without creating pre- or post-echo artifacts in the extracted sources.

1. Introduction

Blind source separation (BSS) describes techniques for extracting independent signals from linear mixtures of these signals without specific knowledge of the mixing conditions or the signal statistics. This paper considers the convolutive BSS of speech signals, in which a set of m speech signals in the vector sequence $\mathbf{s}(k) = [s_1(k) \ \cdots \ s_m(k)]^T$ is mixed by the linear system

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

in which $\mathbf{A}_l$ is the $(m \times m)$ mixing matrix with entries $\{a_{ijl}\}$ at lag l and $\mathbf{x}(k) = [x_1(k) \ \cdots \ x_m(k)]^T$ contains the m measured speech mixtures $x_i(k)$, $1 \leq i \leq m$. The goal is to adjust the coefficient matrices $\{\mathbf{W}_l\}$ of a multichannel separation system

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

such that each signal in the vector sequence $\mathbf{y}(k)$ contains a filtered version of one unique speech signal within $\mathbf{s}(k)$.

Since the development of the first blind source separation methods for convolutive mixtures in the mid-1990's, numerous researchers have found natural gradient methods useful for developing algorithms for speech separation and other convolutive BSS tasks [1]–[9]. Natural gradient adaptation is a modified gradient search in which the Riemannian metric tensor for the parameter manifold is used to adjust the gradient direction within the coefficient updates [10, 11]. For convolutive BSS tasks using multichannel filters, the natural gradient modification applies the separation filter coefficients $\{\mathbf{W}_l\}$ to terms within the standard gradient using non-causal convolution operations. Such natural gradient procedures can typically be implemented using multiplies and adds, and their structure is usually amenable to block-based computations after truncating the system in (2) to finite-impulse-response (FIR) form.

All existing block-based convolutive BSS procedures that use natural gradient adaptation employ approximations to the natural gradient to make the overall system causal in its operation. These approximations typically include (i) truncating the separation system's impulse response to finite length, (ii) delaying the coefficient update terms to achieve causal operation, and (iii) approximating signal-dependent terms within the updates using past system output values. While simulations indicate that these algorithms can achieve some measure of separation for acoustic mixtures, it is not clear what problems such approximations create. Moreover, our extensive experience with the time-domain methods described in [1, 3, 6] indicate that the separation system can exhibit poor performance in situations where the impulse response of the separation system is inadequate to achieve perfect source isolation. The poor performance is exhibited by annoying pre- and post-echoes in the combined system impulse response that remain even if large data sets and numerous training steps are used. While several researchers have proposed constrained natural gradient adaptation procedures in an attempt to mitigate these effects [4]–[7], these procedures impose additional computational operations and can be difficult to tune properly.

In this paper, we present a novel natural gradient procedure for multichannel blind deconvolution and source separation tasks. The algorithm is a block-based

implementation of the iterative procedure described in [12]. Unlike other similar block-based methods [8], the proposed algorithm modifies the standard gradient of the mutual-information-based separation criterion without introducing delayed signal approximations to simplify the updates. The proposed algorithm is entirely causal in its operation and makes use of fast frequency-domain convolution methods for its efficient implementation. Simulations indicate that the proposed procedure can perform speech separation using relatively-short FIR filters while outperforming a competing block-based natural gradient BSS approach of similar complexity.

2. Algorithm Derivation

The proposed algorithm is designed to iteratively minimize the cost function

$$\mathcal{J}(\mathcal{W}_k(z)) = -\frac{1}{N}\sum_{n=k-N+1}^{k}\sum_{i=1}^{m}\log \widehat{p}(y_{i,k}(n)) - \frac{1}{2\pi j}\oint \log\det|\mathcal{W}_k(z)|z^{-1}dz, \quad (3)$$

where

$$\mathcal{W}_k(z) = \sum_{l=0}^{L}\mathbf{W}_l(k)z^{-l} \quad (4)$$

is the z-transform of the separation system's impulse response at time k, N is the block size, $\widehat{p}_i(y)$ is a model of the p.d.f. of the ith source to be separated, and

$$\mathbf{y}_k(n) = \sum_{l=0}^{L}\mathbf{W}_l(k)\mathbf{x}(n-l). \quad (5)$$

It can be shown that (3) is, up to a constant independent of the separation system, proportional to the mutual information of the output vector signal sequence $\{\mathbf{y}_k(n)\}$ when $\widehat{p}_i(y)$ is the true p.d.f. of the ith source sequence. Minimizing this measure results in set of output signals that are most independent of each other.

The proposed algorithm is a natural gradient modification of the block-based standard gradient procedure defined as

$$\mathbf{W}_l(k+N) = \mathbf{W}_l(k) - \mu\frac{\partial\mathcal{J}(\mathcal{W}_k(z))}{\partial\mathbf{W}_l(k)}, \quad (6)$$

where μ is the algorithm step size. The gradient of $\mathcal{J}(\mathcal{W}_k(z))$ is straightforward to calculate assuming that $\mathcal{W}_k(e^{j\omega})$ has no zero singular values; it is

$$\frac{\partial\mathcal{J}(\mathcal{W}_k(z))}{\partial\mathbf{W}_l(k)} = \frac{1}{N}\sum_{n=k-N+1}^{k}\mathbf{f}(\mathbf{y}_k(n))\mathbf{x}^T(n-l) - \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{W}_k^{-T}(e^{-j\omega})e^{j\omega l}d\omega, \quad (7)$$

where $[\mathbf{f}(\mathbf{y})]_i = -\partial\log\widehat{p}_i(y_i)/\partial y_i$ and we have used the substitution $z = e^{j\omega}$ to transform the contour integral on the right-hand side of (3) into a Fourier integral before taking derivatives of this term with respect to $\mathbf{W}_l(k)$.

The second term on the right-hand side of (7) is difficult to compute, and hence the standard gradient procedure is impractical. We propose to employ the natural gradient modification derived in [1] to this algorithm with careful use of truncation to maintain causality of the coefficient updates. The proposed method is closely related to the coefficient updates

$$\mathbf{W}_l(k+N) = \mathbf{W}_l(k) - \left[\mu\frac{\partial\mathcal{J}(\mathcal{W}_k(z))}{\partial\mathbf{W}_l(k)} * \mathbf{W}_{-l}^T(k) * \mathbf{W}_l(k)\right]_+ \quad (8)$$

where $*$ denotes discrete-time convolution of matrix sequences, $\mathbf{W}_l = \mathbf{0}$ for $l < 0$ and $l > L$, and $[\cdot]_+$ denotes the operation of truncating the matrix sequence within brackets to the causal range from $0 \le l \le L$. Define the block data term $\mathbf{G}_l(k)$ as

$$\mathbf{G}_l(k) = \begin{cases} \frac{1}{N}\sum_{n=k-N+1}^{k}\mathbf{f}(\mathbf{y}_k(n))\mathbf{x}^T(n-l), & 0 \le l \le L \\ \mathbf{0} & \text{otherwise.} \end{cases} \quad (9)$$

Finally, noting that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{W}_k^{-T}(e^{-j\omega})e^{j\omega l}d\omega * \mathbf{W}_{-l}^T(k) \to \mathbf{I}\delta_l \quad (10)$$

as L is increased, we can approximate (8) as

$$\mathbf{W}_l(k+N) = (1+\mu)\mathbf{W}_l(k) - \left[\mu\mathbf{G}_l(k) * \mathbf{W}_{-l}^T(k) * \mathbf{W}_l(k)\right]_+ \quad (11)$$

Equations (5), (9), and (11) define the proposed block-based natural gradient convolutive BSS procedure. Several remarks can be made about this procedure:

1. The output, gradient, and update are all in the form of discrete-time convolution or correlation operations. Hence, they can be efficiently implemented using frequency-domain fast convolution techniques.

2. The structure of $\mathbf{W}_{-l}^T(k) * \mathbf{W}_l(k)$ can be exploited to further reduce system complexity. In particular, the diagonal entries of this multidimensional system impulse response are symmetric, and the resulting FFTs of these sequences are all real-valued.

3. Unlike the approaches in [4]–[7], the proposed method involves *multidimensional* convolutions of matrix sequences. Even so, the new algorithm can be implemented such that its overall complexity is similar to other block-based natural gradient BSS approaches, as described in the next section.

for $i = 1$ to m do
$\quad \mathcal{X}_i(k) = \text{FFT}([\underline{\mathbf{x}}_i(k-L-1) \quad \underline{\mathbf{x}}_i(k)])$
end
for $j = 1$ to m do
$\quad \mathcal{Y}_j(k) = \sum_{i=1}^{m} \mathcal{W}_{ji}(k) \odot \mathcal{X}_i(k)$
$\quad [\bullet \quad \underline{\mathbf{y}}_j(k)] = \text{IFFT}(\mathcal{Y}_j(k))$
$\quad \underline{\mathbf{f}}_j(k) = \mu \mathbf{f}(\underline{\mathbf{y}}_j(k))$
$\quad \mathcal{F}_j(k) = \text{FFT}([\underline{\mathbf{z}} \quad \underline{\mathbf{f}}_j(k)])$
$\quad$ for $i = 1$ to m do
$\quad\quad \mathcal{G}_{ji}(k) = \mathcal{F}_j(k) \odot \mathcal{X}_i^*(k)$
$\quad\quad [\underline{\mathbf{g}}_{ji}(k) \quad \bullet] = \text{IFFT}(\mathcal{G}_{ji}(k))$
$\quad\quad \mathcal{R}_{ji}(k) = \sum_{p=1}^{m} \mathcal{W}_{jp}^*(k) \odot \mathcal{W}_{pi}(k)$
$\quad$ end
$\quad$ for $i = 1$ to m do
$\quad\quad \mathcal{V}_{ji}(k) = \sum_{p=1}^{m} \text{FFT}([\underline{\mathbf{g}}_{jp}(k) \quad \underline{\mathbf{z}}]) \odot \mathcal{R}_{pi}(k)$
$\quad\quad [\underline{\mathbf{v}}_{ji}(k) \quad \bullet] = \text{IFFT}(\mathcal{V}_{ji}(k))$
$\quad\quad \underline{\mathbf{w}}_{ji}(k+L+1) = (1+\mu)\underline{\mathbf{w}}_{ji}(k) - \underline{\mathbf{v}}_{ji}(k)$
$\quad\quad \mathcal{W}_{ji}(k+L+1) = \text{FFT}([\underline{\mathbf{w}}_{ji}(k+L) \quad \underline{\mathbf{z}}])$
$\quad$ end
end

Table 1: Fast frequency-domain implementation of the proposed algorithm for a block size of $N = L + 1$; see text for notational explanation.

4. The proposed method does not appear to suffer from the annoying pre- and post-echo artifacts that are introduced by other block-based approaches. We conjecture that this behavior is due to the use of an accurate signal-truncated representation in (11) of the doubly-infinite non-causal natural gradient update first derived in time-domain form in [1].

3. System Implementation

We now describe an efficient, equivalent frequency-domain implementation of the proposed time-domain algorithm in (5), (9), and (11). This implementation assumes the computationally-efficient block size choice of $N = L + 1$, although the extension of the algorithm to other block sizes is straightforward. For this implementation, define the time-reversed row signal vectors

$$\underline{\mathbf{x}}_i(k) = [x_i(k-L) \ \cdots \ x_i(k)] \quad (12)$$

$$\underline{\mathbf{y}}_j(k) = [y_j(k-L) \ \cdots \ y_j(k)] \quad (13)$$

$$\underline{\mathbf{f}}_j(k) = \mu[f_j(y_j(k-L)) \ \cdots \ f_j(y_j(k))] \quad (14)$$

$$\mathbf{z} = [0 \ \cdots \ 0] \quad (15)$$

Furthermore, let $\text{FFT}(\underline{\mathbf{u}})$ and $\text{IFFT}(\underline{\mathbf{u}})$ denote the FFT and inverse FFT, respectively, of a $2(L+1)$-element row vector $\underline{\mathbf{u}}$, and define $\odot$ as the point-by-point complex multiplication of two vectors $\underline{\mathbf{u}}$ and $\underline{\mathbf{v}}$, such that

$$[\underline{\mathbf{u}} \odot \underline{\mathbf{v}}]_l = u_l v_l. \quad (16)$$

Then, Table 1 lists the fast frequency-domain implementation of the proposed algorithm. In this algorithm, $\bullet$ corresponds to an $(L+1)$-element vector of "don't care" values to be discarded. The notation used here closely follows that used to describe other block-based frequency-domain adaptive filters in the signal processing literature.

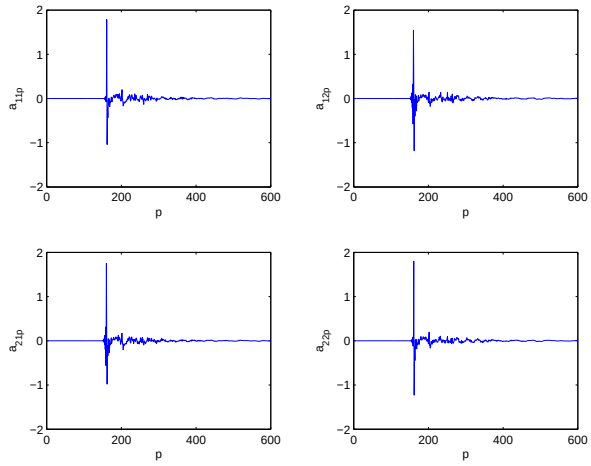

Fig. 1: Impulse responses for the acoustic channel.

The complexity of the proposed algorithm is now considered. For most practical situations in which $L \gg m$, the FFT calculations of a block-based convolutive BSS procedure dominate the system's overall complexity. Our proposed procedure requires $4m^2 + 3m$ FFTs of length $2(L+1)$ for every $(L+1)$ output samples. For comparison, consider the FDMCBD-I algorithm with 75% overlap described in [8], which is a block-based version of the original time-domain natural gradient convolutive BSS procedure derived in [1]. The FDMCBD-I algorithm requires $2m^2 + 3m$ FFTs of length $4(L+1)$ for every $(L+1)$ output samples. Since an M-element FFT requires $M \log_2 M$ complex-valued operations, the two algorithms have a similar complexity to first order. Considering the number of vector dot-multiplies (*e.g.* the $\odot$ symbol in Table 1), the proposed method has $1.5m^3 + 3.5m^2$ such operations, whereas the FDMCBD-I algorithm has $4m^2$ such operations. Considering each algorithm as a whole, the proposed method will be similar in complexity to the FDMCBD-I algorithm except for situations in which the number of sensors or sources m is much larger than two and the filter or block length $(L+1)$ is correspondingly small.

4. Simulations

We now explore the behavior of the proposed procedure via numerical simulations. All of our simulations employ the same two-input, two-output impulse response used in [12] and shown in in Fig. 1. This impulse response was generated from an acoustic laboratory setup of two omnidirectional microphones spaced 4cm apart and mounted in a V-configuration approximately 1.5m from the floor in the center of a 4.45m-by-3.55m-by-2.50m room. A pair of loudspeakers located 1.2m away from the microphones at -40 degrees and $+30$ degrees off-axis

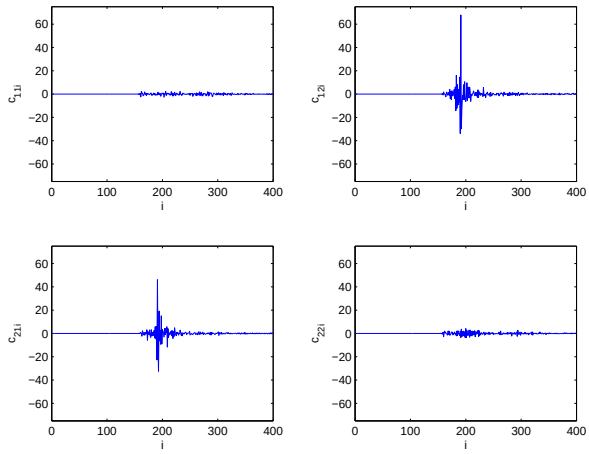

Fig. 2: Combined impulse responses for the new scheme.

served as the acoustic sources. The impulse responses of the loudspeaker-to-microphone acoustic paths were estimated using standard linear estimation techniques at a sampling rate of 8kHz. We then generated the source signal mixtures by digitally filtering recorded male speech signals using these impulse responses. The resulting speech mixtures had an initial signal-to-interference ratio (SIR) of 0.21dB. With this data, we can accurately observe and characterize the combined impulse responses produced by the various algorithms, defined as

$$\mathbf{C}_i(k) \;=\; \sum_{p=0}^{\infty} \mathbf{W}_{i-p}(k)\mathbf{A}_p \qquad (17)$$

All computations were performed using the MATLAB technical computing environment.

Fig. 2 shows the combined impulse responses for the proposed scheme after 1600 block updates (*i.e.* at sample time index $k = 102400$), in which we have chosen $L = 63$, $m = 2$, $\mu = 0.008$, and $\mathbf{W}_p(0) = 10\delta_{p-32}$. As can be seen, the algorithm has suppressed channel crosstalk, and the SIR ratio achieved is approximately 12.5dB. Compare these results to those of Fig. 3 which have been obtained using 4800 block updates of an FFT-based approximate natural gradient algorithm similar to that in [8], except that additional algorithmic simplifications have been employed to allow FFT sizes of $2(L+1)$ without any perceptible loss in overall performance. In this approximate version, the parameters chosen were $L = 63$, $\mu = 0.003$, and $\mathbf{W}_p(0) = 10\delta_{p-32}$. The existing procedure fails to separate the speech signals, achieving a SIR ratio of only 1 to 2dB. The performance of the new scheme is clearly superior in this situation. In addition, the new scheme's performance does not appear to deteriorate under continuous adaptation.

5. Conclusions

In this paper, we have derived a novel causal frequency-domain implementation of a natural gradient multichannel blind deconvolution and source separation algorithm.

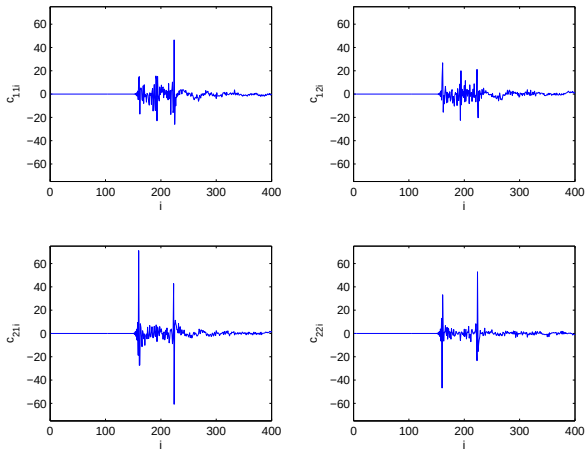

Fig. 3: Combined impulse responses for an existing approximate natural gradient BSS method.

The procedure does not employ delayed coefficient update approximations that appear to cause convergence difficulties for similar approximate natural gradient approaches. The number of FFTs required to implement the algorithm scales almost linearly with the number of FIR filters in the multichannel separation system. Simulations on convolutive speech mixtures indicate that the proposed method provides more robust separation performance than a competing block-based convolutive BSS scheme.

6. References

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

[2] 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.

[3] 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.

[4] 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.

[5] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proc. 3rd Int. Workshop Indep. Compon. Anal. Signal Separation*, San Diego, CA, pp. 722–727, Dec. 2001.

[6] S.C. Douglas, "Blind separation of acoustic signals," in *Microphone Arrays: Techniques and Applications*, M. Brandstein and D. Ward, eds. (New York: Springer-Verlag, 2001), pp. 355-380.

[7] S.C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Communication*, vol. 39, pp. 65-78, Dec. 2002.

[8] M. Joho and P. Schniter, "Frequency domain realization of a multichannel blind deconvolution algorithm based on the natural gradient," *Proc. 4th Int. Workshop Indep. Compon. Anal. Signal Separation*, Nara, Japan, pp. 543-548, Apr. 2003.

[9] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "SIMO-model-based independent component analysis for high-fidelity blind separation of acoustic signals," *Proc. 4th Int. Symp. Indep. Compon. Anal. Signal Separation*, Nara, Japan, pp. 993–998, Apr. 2003.

[10] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251-276, Feb. 1998.

[11] S.C. Douglas and S. Amari, "Natural gradient adaptation," in *Unsupervised Adaptive Filtering, Vol. 1: Blind Source Separation*, S. Haykin, ed. (New York: Wiley, 2000), pp. 13-61.

[12] S.C. Douglas, H. Sawada, and S. Makino, "Natural gradient multichannel blind deconvolution and source separation using causal FIR filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Montreal, Canada, May 2004 (in press).