

Beamforming Initialization and Data Prewhitening in Natural Gradient Convolutional Blind Source Separation of Speech Mixtures

Malay Gupta and Scott C. Douglas

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

Abstract. Successful speech enhancement by convolutional blind source separation (BSS) techniques requires careful design of all aspects of the chosen separation method. The conventional strategy for system initialization in both time- and frequency-domain BSS involves a *diagonal center-spike* FIR filter matrix and no data preprocessing; however, this strategy may not be the best for any chosen separation algorithm. In this paper, we experimentally evaluate two different approaches for potentially-improving the performance of time-domain and frequency-domain natural gradient speech separation algorithms – *prewhitening* of the signal mixtures, and *delay-and-sum* beamforming initialization for the separation system – to determine which of the two classes of algorithms benefit most from them. Our results indicate that frequency-domain-based natural gradient BSS methods generally need geometric information about the system to obtain any reasonable separation quality. For time-domain natural gradient separation algorithms, either beamforming initialization or prewhitening improves separation performance, particularly for larger-scale problems involving three or more sources and sensors.

1 Introduction

Convolutional blind source separation (CBSS) refers to the separation of signals that have been mixed through a dispersive environment using signal processing procedures that do not have specific knowledge of the source properties or the mixing conditions. Due to the dispersive nature of the channel, CBSS algorithms must attempt to undo both spatial and temporal mixing effects. As a result, CBSS algorithms tend to be more complicated than their spatial-only BSS counterparts.

Frequency-domain approaches to CBSS transform the measured mixtures into the discrete frequency-domain via the short-time Fourier transform (STFT) and apply spatial-only (instantaneous) BSS algorithms in each frequency component of the mixtures individually [1]. After separation in the frequency-domain, these signals must be carefully reconstructed before being inverse-Fourier-transformed to recover the time-domain signals. This reconstruction process requires estimating the permutation and scaling ambiguities for all the frequency

components of the separated sources. Prior information about the array geometry and directions-of-arrival (DOAs) of the sources at the sensor array is often assumed. Several researchers have offered ways to use this information in the reconstruction process [2]–[6]. Post-processing permutation resolution can be computationally-demanding if more than two sources are being separated. In many cases, a closed form solution is not possible [4].

In contrast, time-domain CBSS algorithms adapt the impulse response of a multichannel linear filter using only as many output signals as the number of sources that are being extracted [7]–[10]. Because they use time-domain convolutions instead of frequency-domain multiplications, these methods tend to be more difficult to code. Note that their computational complexities can be made to be similar to those of frequency-domain approaches through block processing [8]. Since the algorithms employ a separation criterion whose number of outputs equals the number of sources being estimated, time-domain CBSS approaches do not appear to have severe source permutation problems over different extracted frequencies. These time-domain methods tend to converge more slowly, however, if careful strategies for algorithm implementation are not considered. In [11] one simple way to improve convergence performance for the time-domain method in [8] has been described.

In this paper, we compare the use of two well-known strategies for improving the performance of CBSS algorithms: (1) beamforming initialization [5,12], and (2) multichannel prewhitening [8]. Both time-domain and frequency-domain versions of the well-known natural gradient CBSS method are evaluated and their performances compared to other competing approaches using data collected from a controlled laboratory measurement setup. These numerical experiments show that (a) beamforming initialization is required for frequency-domain natural gradient CBSS methods if no other technique is used to resolve permutation ambiguities, (b) prewhitening alone does not improve the performance of frequency-domain natural gradient CBSS methods, and (c) the performance of time-domain natural gradient algorithms improves with either signal prewhitening or beamforming coefficient initialization, and this improvement is significant when dealing with mixtures of more than two sources.

2 Time- and Frequency-Domain Signal Models

For multichannel acoustic recordings, the n -dimensional signal mixtures at time k , $\mathbf{x}(k) = [x_1(k) \cdots x_n(k)]^T$ can be modeled as

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

where $\{\mathbf{A}_l\}$ denotes a sequence of $n \times m$ mixing matrices, $\mathbf{A}(z) = \sum_{l=-\infty}^{\infty} \mathbf{A}_l z^{-l}$ is the multichannel system transfer function, and $\mathbf{s}(k) = [s_1(k) \cdots s_m(k)]^T$ is the m -dimensional signal vector at time k . All CBSS algorithms attempt to find a time-varying separating or demixing system $\mathbf{B}(k, z)$ to process the signal

mixtures $\mathbf{x}(k) = \mathbf{A}\{\mathbf{s}(k)\}$ such that $\mathbf{y}(k) = \mathbf{B}\{\mathbf{x}(k)\}$ contains the estimates of each of the sources in $\mathbf{s}(k)$ without repetition. Mathematically, this can be represented as

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

In practice, a truncated causal approximation to (2) is often employed, where L is a positive integer and

$$\mathbf{y}(k) = \sum_{l=0}^L \mathbf{B}_l(k)\mathbf{x}(k-l). \quad (3)$$

Frequency-domain CBSS algorithms use the STFT to transform the time-domain data into the frequency-domain, whereby a separate complex-valued instantaneous demixing system is found for each of the frequency components of the mixed signals. The input data in the l^{th} frequency bin ω_l is given by

$$\mathbf{x}(\omega_l, k) = \mathbf{A}(\omega_l)\mathbf{s}(\omega_l, k), \quad (4)$$

where k denotes the time dependence of the STFT, $\mathbf{s}(\omega_l, k)$ is the transformed source signal vector, and $\mathbf{A}(\omega_l)$ denotes the mixing matrix for the l^{th} frequency bin. As such, the demixing process in each frequency bin is formulated as

$$\mathbf{y}(\omega_l, k) = \mathbf{B}(\omega_l, k)\mathbf{x}(\omega_l, k), \quad (5)$$

where $\mathbf{y}(\omega_l, k)$ and $\mathbf{B}(\omega_l, k)$ are the estimated source signal vector and the demixing matrix, respectively, in the l^{th} frequency bin at time k .

3 Beamforming vs. Prewhitening in Convolutional Blind Source Separation

Both beamforming and prewhitening attempt to solve part of the goal achieved by successful application of CBSS methods in certain contexts.

Beamforming and CBSS attempt to suppress interferences caused by spatially-distinct sources to extract individual source signals when operating on data collected from a uniform linear array. Beamforming methods work by providing maximum gain in the direction of the desired user. CBSS based methods have been observed to place spatio-temporal nulls in the directions of interfering users in some environments [12].

Beamforming methods typically assume a working knowledge of the sensor array manifold and the directions-of-arrival (DOAs). CBSS methods, on the other hand, typically assume no known signal or measurement structure other than a linear dispersive channel for the mixing process. Researchers have suggested the merger of beamforming with CBSS to include prior information about the array manifold and DOAs within CBSS algorithms [3]. These techniques are primarily designed to remove permutation difficulties that lead to lower separation

performance. In situations where DOA information is not available *a priori*, researchers have suggested procedures for estimating DOAs as part of the CBSS algorithm being developed [3,6]. In narrowband beamforming, the directional vector associated with a frequency ω for a source impinging on the array from a direction θ is given as

$$\mathbf{d}(\omega, \theta) = [\exp(j\omega\tau_0(\theta)) \cdots \exp(j\omega\tau_{m-1}(\theta))]^T, \tag{6}$$

where $\tau_l(\theta) = ld\sin(\theta)/c$ is the time delay associated with the l^{th} sensor with respect to the reference sensor, m is the number of sensors in the array, d is the array element separation, and c is the speed of sound. Signals with significant frequency content (*e.g.* audio signals) received at a sensor array will typically have directional vectors associated with each of their frequency components. Thus, clustering of the directional vectors may be needed to obtain a consistent estimate of the source DOAs.

Perhaps the simplest way to employ DOA knowledge to improve CBSS convergence performance is to initialize the separation system coefficients $\{\mathbf{B}_l(0)\}$ or their frequency counterparts $\{\mathbf{B}(\omega_l, 0)\}$ to a series of fixed beamformers in which the mainlobe of each of the beampatterns in each frequency bin for the i th separation system points toward a talker. In this case, we would choose

$$\mathbf{B}(\omega_l, 0) = [\mathbf{d}(\omega_l, \theta_1) \ \mathbf{d}(\omega_l, \theta_2) \ \dots \ \mathbf{d}(\omega_l, \theta_m)]^H. \tag{7}$$

For time-domain algorithms, we can compute the appropriate initial coefficients by taking the inverse FFTs of the frequency-domain responses in (7) about their points of symmetry. Initializing CBSS algorithms in this way does not modify the algorithm's operation other than choosing its initial state. The main alternative to this coefficient initialization is center-spike initialization, in which

$$\mathbf{B}_l(0) = \begin{cases} \mathbf{I}, & l = \frac{L}{2} \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{8}$$

For frequency-domain algorithms, we can compute the appropriate initial coefficients by taking the FFT of the time-domain responses in (8) about their points of symmetry. Center-spike initialization makes no assumption about the source-sensor array geometry.

Prewhitening is a preprocessing strategy whereby the measured signals $\{x_i(k)\}$, $1 \leq i \leq n$ are linearly filtered such that the filtered signals $\{v_i(k)\}$ approximately satisfy

$$E\{v_i(k)v_j(k-l)\} \approx E\{|v_i(k)|^2\}\delta_{i-j}\delta_l, \tag{9}$$

where δ_l is the Kronecker delta function. These prewhitened signals are used in place of $\{x_i(k)\}$ in the separation system. Examples of prewhitening algorithms include the linear phase adaptive procedure in [13] and the least-squares multichannel linear predictor described in [8]. When block processing is used, one can use successive filtering operations to process $\{\mathbf{x}(k)\}$ to produce $\mathbf{v}(k)$, which is likely the most computationally efficient method.

Prewhitening solves part of the CBSS task, as decorrelation is a necessary but not sufficient condition for source separation of mixtures of statistically-independent signals. Hence, it is reasonable to use prewhitening as a preprocessing step to remove any signal correlations contained in the data prior to performing separation with any CBSS method. In this case, the input signals $\{x_i(k-l)\}$ used in the separation system are replaced by the prewhitened signals $\{v_i(k-l)\}$ obtained at the outputs of the prewhitening system.

One can view beamforming initialization and prewhitening as two simple but competing approaches for improving the performances of CBSS methods that do not require significant alteration of the separation algorithm. Note that prewhitening effectively alters the DOAs seen by the separation system within the prewhitened data, so using both prewhitening and beamforming initialization does not make sense unless special constraints are placed on the prewhitening task. It is unclear without performing experiments which procedure is to be preferred, and whether both frequency-domain-based and time-domain-based CBSS algorithms benefit from such procedures. The goal of this paper is to explore these issues through experimental evaluation on real-world speech signal mixtures to see what classes of algorithms benefit most from them.

4 Numerical Experiments

We now present numerical evaluations to illustrate the separate effects that beamforming initialization and data prewhitening have on the behaviors of one class of CBSS algorithms. In order to minimize any performance effects due to choice of separation criterion, we focus on the natural gradient algorithms presented in [4] and [8]. The algorithms attempt to minimize the mutual information of the extracted signals using frequency-domain and time-domain system structures, respectively. For comparison, we show the performance of two other algorithms on this data: one employing decorrelation with geometric beamforming constraints [3], and one using contrast-based optimization with prewhitening [9,10]. These latter algorithms incorporate either beamforming or prewhitening within their structures and are not claimed to work without such pre-processing.

Data for these evaluations was generated in an acoustically-isolated laboratory environment with three loudspeakers playing recordings of talkers (one female and two male) as the sources. The sources were located 127 cm away from the three omnidirectional microphones and were spaced at angles of -30° , 0° , and 27.5° , respectively, from the array normal. The inter-sensor spacing of the microphone array was 4 cm. Acoustic foam was placed on the walls of the room to obtain a reverberation time of 300 ms for the environment. All recordings were made using 7 seconds of data per channel and a 48 kHz sampling rate and were downsampled to an 8 kHz sampling rate for processing. Fig. 1 shows the impulse responses of the loudspeaker/microphone paths for these mixing conditions.

The various algorithms were applied to this measured microphone data for two- and three-source mixtures, whereby the 0° source was omitted for the two-source mixture. After separation, least-squares methods were used to estimate

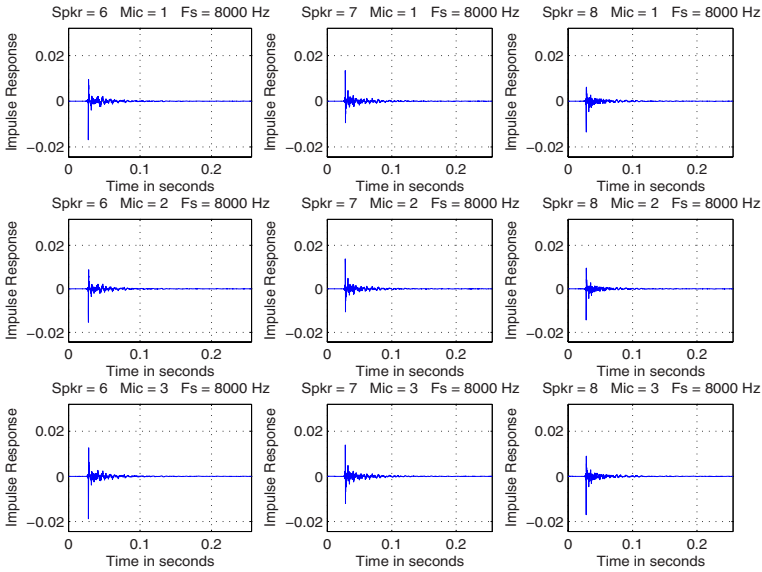


Fig. 1. Acoustic chamber impulse response in a three source, three microphone setup. Room conditions correspond to a reverberation time (RT) of 300 msec.

the contributions of the source recordings to each of the recorded mixtures as well as the output signals from each algorithm. By calculating power ratios from these least-squares estimates, we can compute the average improvement in signal-to-interference-plus-noise ratio (SINR) for each algorithm in each case.

For the normalized natural gradient algorithm in the frequency domain [11], the parameters chosen were $L = 512$ and $\mu = 0.35$, and 200 passes of the algorithm through the data have been used to adapt the filter. For the natural gradient time-domain algorithm [8], we used $L = 512$ and a step size schedule of $\mu = .0009$ for 150 data passes followed by $\mu = 0.0001$ for a single data pass followed by $\mu = 0.00001$ for a second single data pass. The data nonlinearity used in each algorithm was $f(y) = y/|y|$, where y in this case corresponds to the i th frequency bin output or the i th time-domain filter output, respectively.

Table 1 shows the SINR improvements obtained by the various algorithms for the various processing strategies on the two-source mixture data. As can be seen, the frequency-domain natural gradient method does not perform well either with center-spike initialization or with data prewhitening. With beamforming initialization, the algorithm achieves good performance on this data that closely matches the time-domain natural gradient algorithm. The latter algorithm's performance is quite good for center-spike initialization on this data, but improvements of 1.0dB and 2.7dB are obtained with beamforming initialization and data prewhitening, respectively. Shown for comparison are the behaviors of the decorrelation-based method in [3] as well as the contrast-based method with prewhitening in [11]. As can be seen, the time-domain natural gradient

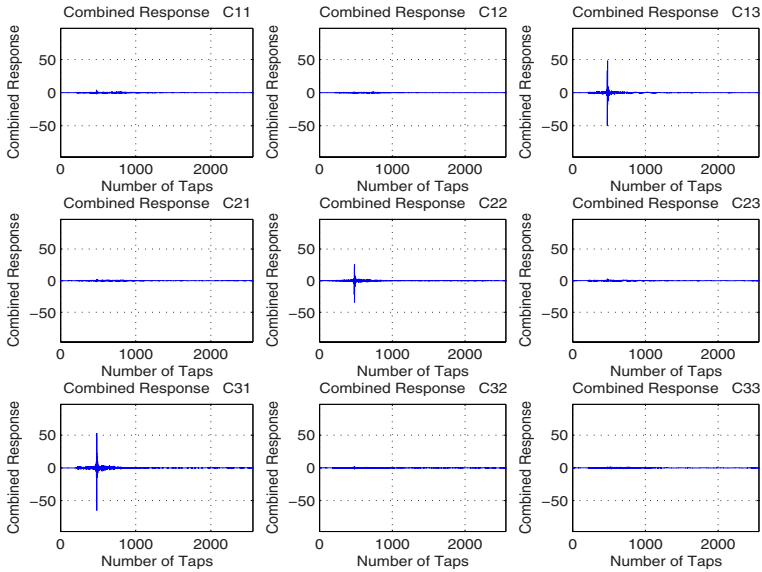


Fig. 2. Combined impulse response of the time-domain truncated natural gradient with delay-and-sum beamforming initialization

Table 1. Improvement in average SINR [dB]; RT=300 ms

Algorithm	TWO SOURCE CASE			THREE SOURCE CASE		
	Center Spike	w/Beamforming	w/Prewhitening	Center Spike	w/Beamforming	w/Prewhitening
SNGFD[11]	0.25	13.56	1.52	3.33	12.55	4.55
NGTD[8]	12.63	13.60	15.34	10.89	17.07	16.80
Parra-GBSSII[3]	–	7.95	–	–	5.42	–
STFICA-Symm[9,10]	–	–	11.23	–	–	12.66

method outperforms both of these competing methods when using the same spatial knowledge of the environment or data pre-processing.

Also shown in Table 1 are the SINR improvements obtained by the various algorithms for the various processing strategies on the three-source mixture data. Similar performance relationships as in the two-source data case are observed in this case. The frequency-domain natural gradient algorithm obtains adequate separation only with beamforming initialization, whereas the time-domain natural gradient algorithm can separate the source mixtures with any of the three strategies employed. The best performance is obtained with beamforming initialization, although separation using data prewhitening is nearly as good. Fig. 2 shows the combined impulse responses at convergence for the natural gradient time-domain algorithm with beamforming initialization when applied to this data, indicating that separation has occurred. It should be noted that

prewhitening-based processing strategies can still be used if knowledge of the source-sensor array geometry is not available.

5 Conclusions

In convolutive blind source separation of speech signal mixtures, beamforming initialization and prewhitening are two simple strategies for improving the performance of any separation algorithm not already leveraging this structural knowledge. This paper evaluates the behaviors of two versions of the well-known natural gradient algorithm as implemented in the time- and frequency-domains, respectively, when using each of these strategies. Experiments indicate that the frequency-domain natural gradient algorithms rely on the spatial structure of the source-microphone mixing conditions, and they cannot adequately separate sources without using knowledge of the directions-of-arrival within the algorithm. Prewhitening alone does not help the performance of frequency-domain algorithms. Time-domain natural gradient algorithms can separate without directions-of-arrival knowledge; however, their performances are improved when either beamforming initialization or data prewhitening is employed.

References

1. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. *Neurocomputing* 22(1-3), 21–34 (1998)
2. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. *IEEE Trans. Speech Audio Processing* 8, 320–327 (2000)
3. Parra, L., Alvino, C.: Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Trans. Speech Audio Processing* 10(6), 352–362 (2002)
4. Mitianoudis, N., Davies, M.E.: Audio source separation of convolutive mixtures. *IEEE Trans. Speech Audio Processing* 11, 489–497 (2003)
5. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. Speech Audio Processing* 12, 530–538 (2004)
6. Saruwatari, H., Kawamura, T., Nishikawa, T., Lee, A., Shikano, K.: Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. *IEEE Trans. Audio Speech Language Processing* 14, 666–678 (2006)
7. Amari, S., Douglas, S.C., Chichocki, A., Yang, H.H.: Multichannel blind deconvolution and equalization using the natural gradient. In: *Proc. IEEE Workshop Signal Proc. Adv. Wireless Comm. Paris, France, April 1997*, pp. 101–104. IEEE Computer Society Press, Los Alamitos (1997)
8. Douglas, S.C., Sawada, H., Makino, S.: Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters. *IEEE Trans. Speech Audio Processing* 13, 92–104 (2005)
9. Douglas, S.C., Sawada, H., Makino, S.: A spatio-temporal FastICA algorithm for separating convolutive mixtures. In: *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Philadelphia, PA, vol. 5*, pp. 165–168 (March 2005)

10. Douglas, S.C., Gupta, M., Sawada, H., Makino, S.: Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures. *IEEE Trans. Speech Audio Language Processing*, 15(5) (July 2007)
11. Douglas, S.C., Gupta, M.: Scaled natural gradient algorithms for instantaneous and convolutive blind source separation. In: *IEEE Int. Conf. Acoust. Speech, Signal Processing*, Honolulu, HI (April 2007) (to appear)
12. Araki, S., Makino, S., Hinamoto, Y., Mukai, R., Nishikawa, T., Saruwatari, H.: Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures. *EURASIP J. Applied Signal Processing* 2003(11), 1157–1166 (2003)
13. Douglas, S.C., Cichocki, A.: Neural networks for blind decorrelation of signals. *IEEE Trans. Signal Processing* 45, 2829–2842 (1997)