THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.24, NO.2E 2005. 6 pp. 54-60

A Frequency-Domain Normalized MBD Algorithm with Unidirectional Filters for Blind Speech Separation

Hye-Jin Kim*, Seung-Hyon Nam*
*Department of Electronic Engineering, Paichai University
(Received October 28 2004; revised November 23 2004; accepted January 26 2005)

Abstract

A new multichannel blind deconvolution algorithm is proposed for speech mixtures. It employs unidirectional filters and normalization of gradient terms in the frequency domain. The proposed algorithm is shown to be approximately nonholonomic. Thus it provides improved convergence and separation performances without whitening effect for nonstationary sources such as speech and audio signals. Simulations using real world recordings confirm superior performances over existing algorithms and its usefulness for real applications.

Keywords: *Blind Multichannel Deconvolution, Blind Source Separation, Natural Gradient, Whitening Effect, Spectrum Normalization*

I. Introduction

Blind speech separation (BSS) in a room environment is very attractive for many practical applications such as robust speech recognition, echo cancelation, and object-based audio processing. The multichannel blind deconvolution (MBD) algorithm is one practical method for blind source separation. It is based on the fact that original sources are statistically independent in nature. There are many ways to measure statistical independence. Mutual information is one of such measures widely used in blind source separation[1,2]. If sources are nonstationary, on the other hand, multiple decorrelations at different lags are utilized[3].

In convolutive mixing, the mixed signal at the sensor j is given by

$$x_j(k) = \sum_{i=1}^{n} \sum_{p=-\infty}^{\infty} a_{ji,p} s_i(k-p),\ j = 1, \cdots, m \qquad (1)$$

where $s_i(k), i = 1, \cdots, n$, are source signals and $a_{ji,p}$ is the $(j,i)^{th}$ element of the mixing system $A(z) = \sum_{p=-\infty}^{\infty} A_p(z) z^{-1}$ at lag p. Similarly, the i^{th} unmixing signal is given by

$$u_i(k) = \sum_{j=1}^{m} \sum_{p=-\infty}^{\infty} w_{ij,p}(k) x_j(k-p),\ i = 1, \cdots, n \qquad (2)$$

where $w_{ij,p}(k)$ is the $(j,i)^{th}$ element at lag p of the unmixing system at time k, ie. $W(z,k) = \sum_{p=-\infty}^{\infty} W_p(k) z^{-p}$. The number of sensors m is assumed to be equal to or greater than the number of sources n in general. Since statistical independence is invariant under ordering, scaling, and filtering, it is well known that there exist fundamental indeterminacy on order, scaling, and filtering of unmixed signals.

Various frequency-domain and time-domain MBD algorithms are proposed for blind speech separation[1-3]. In the frequency-domain algorithms, separation is performed in each frequency bin and permutation in each frequency bin occur. Although various solutions to the permutation problem have been proposed, significant performance degradation is inevitable. Time-domain algorithms, on the other hand, suffer from whitening effect such that spectrum of unmixed signals are flattened. The whitening effect is a major obstacle to blind speech separation in which quality of unmixed speech signals is of prime interest.

Corresponding author: Seung-Hyon Nam (shnam@pcu.ac.kr)
Department of Electronic Engineering, Paichai University
439-6, Doma-dong, Seo-ku, Daejeon
(Phone: 042-520-5607)

Furthermore, time-domain algorithms do not provide enough separation performance for speech signals in real world noisy environments.

In this paper, structures and properties of existing time-domain algorithms are investigated. Then a solution to this problem is investigated in a single channel case, and the solution is extended to multichannel case. Simulations using a real world data confirms superior performances of the proposed algorithm in terms of convergence, separation, and speech quality. In addition, a new performance measure is devised to compare performances of various algorithms in real situations.

II. Structures of Existing MBD Algorithms

In[1], the MBD algorithm with the natural gradient (NGMBD) is then presented as

$$\Delta \boldsymbol{W}_p(k) = \boldsymbol{W}_p(k) - \boldsymbol{y}(k-L)\ \boldsymbol{v}^T(k-p) \tag{3}$$

where $\boldsymbol{y}(k) = f(\boldsymbol{u}(k))$ for some monotonic nonlinear function $f(\cdot)$ and

$$\boldsymbol{v}(k) = \sum_{q=0}^{L-1} \boldsymbol{W}_{L-q}^T(k)\ \boldsymbol{u}(k-q) \tag{4}$$

where L is the filter length. Notice that (4) is backward filtering of unmixed signals caused by bidirectional nature of unmixing filters. This bidirectional nature introduces many approximations in the derivation of the MBD algorithm[1].

Although the NGMBD algorithm works very well for sources uncorrelated in time, it is well-known to suffer performance degradation for highly correlated nonstationary sources such as speech and audio. Performance degradation of the NGMBD algorithm for acoustical mixtures is twofold. One is slow convergence due to approximations/delays involved and large eigenvalue spread of the cross-correlation matrix between $\boldsymbol{y}(k-L)$ and $\boldsymbol{u}(k-p)$. Approximations and delays increase with filter length. The other factor is whitening of unmixed sources since the NGMBD algorithm (3) has equilibrium points

$$E\{y_i(k)u_j(k-l)\} = \delta_{ij}\delta_l \tag{5}$$

Consequently quality of the unmixed acoustic signal is generally poor although it is still intelligible. The whitening problem has been treated by some researchers - post processing [2], a nonholonomic algorithm[4], and a linear predictive method [5]. Post processing is limited because of spectra of original sources are unknown. In case of the linear predictive method, there is a danger of predicting room impulse responses as vocal track transfer functions. Finally, exact nonholonomic constraints cannot be implemented in (3) due to backward filtering (4).

Another interesting MBD algorithm with natural gradient has been proposed in[2] as

$$\Delta \boldsymbol{W}^f(b) = \left\{\bar{\boldsymbol{I}} - \boldsymbol{y}^f(b)(\boldsymbol{u}^f(b))^H\right\} \boldsymbol{W}^f(b) \tag{6}$$

where the superscript f denotes quantity in the frequency domain and $\bar{\mathbf{I}}$ is a matrix that is formed by repeating an identity matrix along the frequency domain. Notice that the computation of (6) is performed in element-wise in each frequency f. In addition, linear convolutions and correlations are performed in the frequency domain using circular convolutions and correlations, respectively. Thus it is necessary to remove aliased parts properly. This algorithm is based on the FIR polynomial matrix algebra developed by Lambert[6]. It is worth to mention that in (6) mixing/unmixing processes are assumed to be circulant which is not true in general for nonstationary sources[7]. For this reason, separation and convergence performances of the algorithm are generally poor if sources are not white.

Investigation of these two MBD structures reveals that an exact MBD algorithm using bidirectional filters is not possible. This is the main reason why MBD algorithms with bidirectional filters do not provide sufficient performance in real world environments. Therefore, we need to seek a new structure for better performance. A solution to this problem is obtained by investigating a single channel algorithm.

III. An Alternative Form of the MBD Algorithm for Normalization in the Frequency Domain

3.1. A Single Channel Case

Consider a single channel Bussgang deconvolution algorithm with natural gradient[1]. In a single channel Bussgang deconvolution algorithm, the filter $w(k)$ is assumed to be finite

and the output is expressed as

$$u(k)=\sum_{p=0}^{L-1} w_p(k)\,x(k-p) \tag{7}$$

The Bussgang algorithm with natural gradient is then obtained by applying $w(z^{-1},k)w(z,k)$ to the standard gradient as

$$\Delta w_p(k)=y(k)\sum_{q=0}^{L-1}\sum_{r=0}^{L-1}x(k-p+q-r)w_q(k)\,w_r(k) \tag{8}$$

where $y(k)=f(u(k))$ is the output of Bussgang nonlinearity to $u(k)$. If we assume that $w_r(k-p+q)\approx w_r(k)$ for $0\le p,q\le L-1$, (8) can be rewritten approximately as

$$\begin{pmatrix}\Delta w_0(k)\\ \vdots\\ \Delta w_{L-1}(k)\end{pmatrix}=\begin{pmatrix}y(k)u(k) & \cdots & y(k)u(k-L+1)\\ \vdots & \ddots & \vdots\\ y(k)u(k-L+1) & \cdots & y(k)u(k)\end{pmatrix}\begin{pmatrix}\Delta w_0(k)\\ \vdots\\ \Delta w_{L-1}(k)\end{pmatrix} \tag{9}$$

Since future samples of $u(k)$ are involved in (9), $u(k)$ and $y(k)$ are delayed by L samples as in (3). The Bussgang algorithm is usually initialized with $w_p(0)=\delta_{p-q}$ for $0\le q\le L-1$. Then the position of the leading tap w_q affects convergence of the algorithm. If $0\le q\le L-1$, the converged filter $w_p(0)=\delta_{p-q}$ would be a delayed version of a bidirectional nonminimum phase filter.

The cross-correlation matrix between $y(k)$ and $u(k)$ is diagonal if the source signals are nearly white which is true in general for telecommunication signals. If the source signals are correlated in time, however, the cross-correlation would not be diagonal and eigenvalue spread would be large. Although natural gradient provides faster convergence than standard gradient, large eigenvalue spread would be very harmful to convergence of the algorithm.

To demonstrate this adverse effect, we examined the trajectories of equalizers with both standard and natural gradients. We used the same experimental setup as in [8] except that speech sources as well as noise sources are used. Figure 1 shows trajectories of algorithms that start from six different initial values to the optimal point $w_{opt}=[1,\ -0.95]^T$. For white noise sources, trajectories Fig. 1(c) of natural gradient are direct than trajectories Fig. 1(a) of standard gradient as demonstrated in[8]. For a speech source, however, trajectories Fig. 1(d) of natural gradient become similar to trajectories Fig. 1(b) of standard gradient.

For better convergence, it is necessary to introduce normalization of the gradient term in the frequency domain[9]. To demonstrate the effect of normalization, we compared three single channel algorithms: with a bidirectional and a unidirec- tional filter both normalized in the time domain and with a unidirectional filter normalized in the frequency domain. Figure 2 shows clear advantage of frequency-domain normalization over time-domain normalization. In case of time-domain normalization, performance of the bidirectional filter is turned out to be better than that of unidirectional filter since filter length is only two. As the filter length increases, however, error due to approximation/ delay would affect adversely.

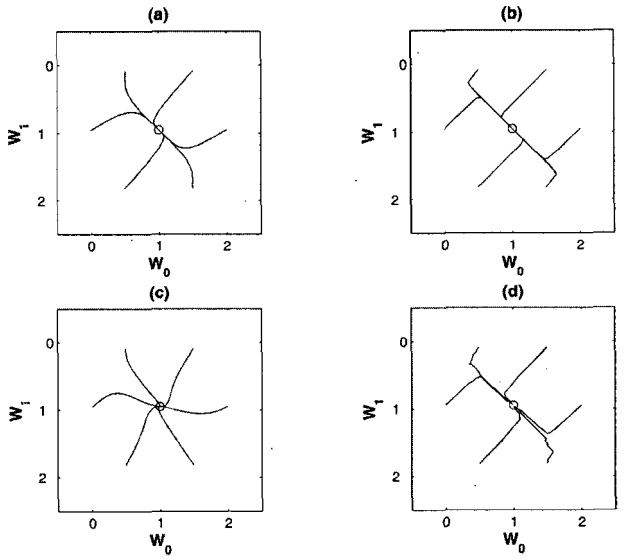

Fig. 1. Trajectories of single channel equalizers: (a) standard/white, (b) standard/speech, (c) natural/white, and (d) natural/ speech.

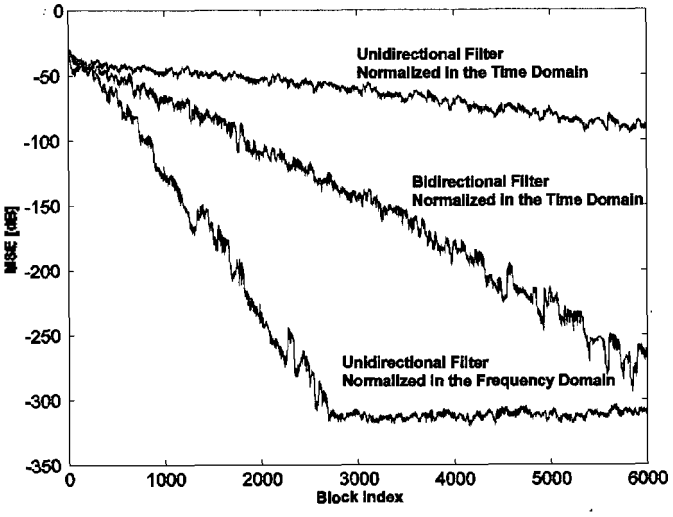

Fig. 2. Convergence of three single channel algorithms for speech source.

3.2. A MBD Algorithm with Unidirectional Filters

It is noted that direct normalization of the gradient term in the frequency domain may not always provide satisfactory results because of backward filtering (4). To avoid backward filtering, future samples in (3) are simply ignored rather than being delayed. Then the single channel algorithm (9) becomes

$$\begin{pmatrix} \Delta w_0(k) \\ \vdots \\ \Delta w_{L-1}(k) \end{pmatrix} = \begin{pmatrix} y(k)u(k) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ y(k)u(k-L+1) & \cdots & y(k)u(k) \end{pmatrix} \begin{pmatrix} \Delta w_0(k) \\ \vdots \\ \Delta w_{L-1}(k) \end{pmatrix} \tag{10}$$

or

$$\Delta w_p(k) = \sum_{q=0}^{p} y(k)u(k-p+q)\, w_q(k) \tag{11}$$

If we extending (11) into the multichannel case, the corresponding multichannel algorithm can be written as

$$\Delta \boldsymbol{W}_p(k) = \sum_{q=0}^{p} \{ \boldsymbol{I} - \boldsymbol{y}(k)\, \boldsymbol{u}^T(k-p+q) \}\, \boldsymbol{W}_q(k) \tag{12}$$

Notice that (12) does not include backward filtering and unmixing filters are unidirectional and causal. The correlation term $\boldsymbol{y}(k)\boldsymbol{u}^T(k-p+q)$ can be easily normalized in the frequency domain. In fact, this is the same causal MBD algorithm that has been derived on the basis of the geometrical structures of the FIR manifolds[10]. The minimum phase algorithm (12) is known to have good convergence properties: equivariant property in the Lie group sense and nonsingularity of $\boldsymbol{W}_0$ [10]. Notice that the causal MBD algorithm (12) still suffer from the whitening effect for nonstationary sources. There is no literature that reports the application of the causal MBD algorithm (12) to blind speech separation although a nonholonomic version is mentioned[11].

IV. A Normalized Form of the Causal MBD Algorithm with Unidirectional Filters

The MBD algorithms are in general implemented in the frequency domain using FFTs in an overlap-save manner[12-14]. Nevertheless, any existing MBD algorithm has never been presented in an normalized form in the frequency domain for its structural inadequacy.

The causal MBD algorithm (12) with unidirectional filters can be effectively normalized in the frequency domain as

$$\Delta \boldsymbol{W}^f(b) = \left\{ \bar{\boldsymbol{I}} - \Lambda_y^{-1}(b)\, \boldsymbol{y}^f(b) (\boldsymbol{u}^f(b))^H\, \Lambda_u^{-1}(b) \right\} \boldsymbol{W}^f(b) \tag{13}$$

where the superscript ^{f} denotes the Fourier transform. Here, $\boldsymbol{W}^f(b)$, $\bar{\boldsymbol{I}}$, $\Lambda_y(b)$, $\Lambda_u(b)$ are (n x n x N) matrices, and $\boldsymbol{y}^f(b)$ and $\boldsymbol{u}^f(b)$ are (n x N) matrices where N is the FFT size. Computation of (13) is performed in element-wise in the frequency domain. Proper time-domain constraints are imposed to computations of linear convolutions and correlations using circular convolutions and correlations, respectively. The heart of the algorithm is normalization in the frequency domain using extended diagonal matrices $\Lambda_y(b)$ and $\Lambda_u(b)$ whose diagonal elements $\sqrt{P_{y_i}(b)}$ and $\sqrt{P_{u_j}(b)}$ are given by power spectrum of $\boldsymbol{y}^f(b)$ and $\boldsymbol{u}^f(b)$, respectively. The power spectra are updated for each frequency at each block time b as follows:

$$P_{y_i}(b) = (1-\gamma)\, P_{y_i}(b-1) + |\, y_i^f(b)|^2 \tag{14a}$$

$$P_{u_j}(b) = (1-\gamma)\, P_{u_j}(b-1) + |\, u_j^f(b)|^2 \tag{14b}$$

where $0 \le \gamma \le 1$.

The proposed algorithm has many excellent properties. Immediately, we can observe that, at steady states, the update rule (13) has equilibrium points

$$\frac{E\{ y_i^f(b)\, (u_j^f(b))^* \}}{\sqrt{E\{|y_i^f(b)|^2\} E\{|u_j^f(b)|^2\}}} = \delta_{ij} \tag{15}$$

Clearly, equilibrium points (15) do not impose any compulsory constraints on the spectrum of unmixed signals whereas (5) forces whitening of unmixed sources. Therefore, the proposed algorithm is approximately nonholonomic. The exact holonomicity can be achieved also if we set all the diagonal components of the gradient terms to zero. Such flexibility is one advantage of the algorithm. Normalization of gradient terms in the frequency domain also improves convergence since it provides the same step size for all frequency bins.

V. Simulations

5.1. Performance Measure

Usually intersymbol interference (ISI) or signal-to-interference ratio (SIR) are used as performance measures for BSS algorithms.

These performance measures, however, require special knowledge in experimental setups. That is, mixing filters should be known to compute ISI values and only one source should be active at a time to compute SIR values.

In this paper, we use the normalized off-diagonal power of the cross-correlation matrix of unmixed signals. Note that diagonalization of the cross-correlation matrix of the unmixed signals is utilized to separate mixed signals in the multiple decorrelation algorithm based on the second order statistics[15]. The normalized off-diagonal power is computed in the frequency domain and averaged over the entire signal at each iteration as follows:

$$P_{off}(b) = \frac{\sum_{i}\sum_{j\neq i}\sum_{f} |u_i^f(b)(u_j^f(b))^*|}{\sum_{i}\sum_{f} |u_i^f(b)|^2} \tag{16a}$$

$$\bar{P}_{off}(iter) = \frac{1}{K}\sum_{b=1}^{K} 10\log_{10} P_{off}(b) \tag{16b}$$

where *iter* is the iteration index and k is the total number of blocks in an entire signal.

5.2. Results with Real Recordings

To confirm the performance of the proposed algorithm, experiments on blind speech separation in a normal office and a car have been performed. We compare the proposed algorithm (13) with the NGMBD algorithm (3) and the causal MBD algorithm (12). Notice that the proposed algorithm uses normalization in the frequency domain, while the other two algorithms use normalization in the time domain. The MBD algorithm (6) which uses circulant condition as mentioned, is not compared here since we are concentrated on the modified versions of the MBD algorithm (3).

a. Real Recordings in a Normal Office

We use the English-Spanish data sampled at 16kHz, which counts from 1 to 10 recorded in a normal office, avaliable in a web-site[16]. Filter length, block length, and frame length are set to 128, 256, 512, respectively, so that 50% overlap-save is used. Figure 3 shows $\bar{P}_{off}(iter)$ values computed at each iteration during learning of speech data of 7.8sec long. The proposed algorithm performs far better than the other two. Separation performance of the proposed algorithm is also investigated by comparing the unmixed outputs. Figure 4 and 5 show that the proposed algorithm significantly improves separation performance without whitening effect. Listening to unmixed outputs clearly confirms the results.

b. Real Recordings in a Car

Recordings in a car are collected using two AT831b (by Audio-Technica) pin microphones attached under both sides of the center rear-view mirror. Mixed voices of the driver and the passenger are recorded at 44.1kHz and downsampled to 8kHz. The car is turned on and all windows are closed.

Filter length, block length, and frame length are set to 256, 512, 1024, respectively. Figure 6 show convergence of algorithms as data of 4.6sec long is iterating 20 times. The proposed algorithm performs far better than the other two algorithms. It is also interesting to see from Fig. 3 and 6 that the MBD algorithms normalized in the time domain use step size values quite different for each experiment. In case of the proposed

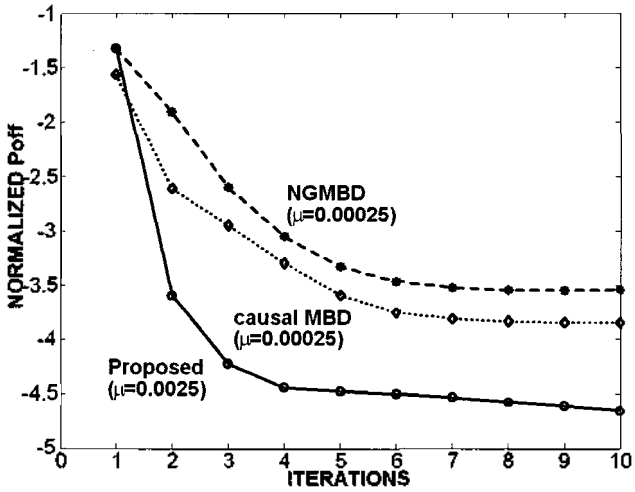

Fig. 3. Convergence of MBD algorithms for the recordings in a normal office.

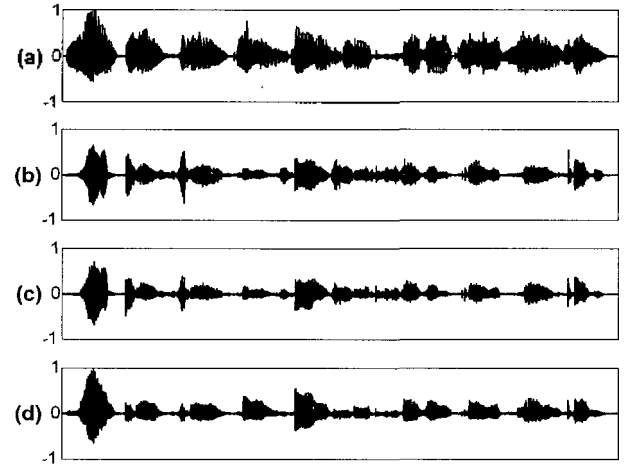

Fig. 4. Mixed and unmixed signals (channel 1): (a) mixed signal and unmixed signal from (b) the NGMBD (c) the causal MBD (d) the proposed algorithm.

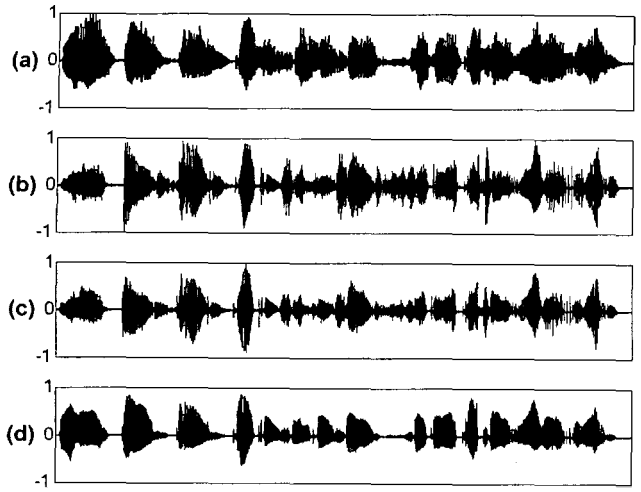

Fig. 5. Mixed and unmixed signals (channel 2): (a) mixed signal and unmixed signal from (b) the NGMBD (c) the causal MBD (d) the proposed algorithm.

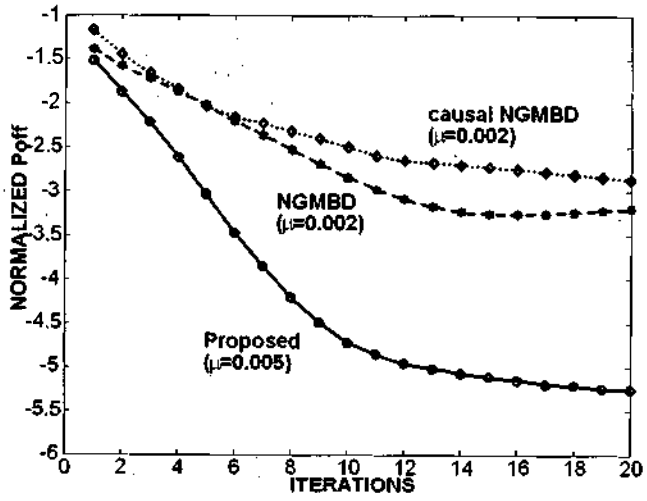

Fig. 6. Convergence of MBD algorithms for the recordings in a car.

algorithm, however, the same step size is used for both experiments with respect to the frame length. This implies uniform convergence behavior of the proposed algorithm.

VI. Conclusions

Existing MBD algorithms for blind speech separation in real world environments are reviewed. After investigating a single channel algorithm, an efficient way to overcome shortcomings of the existing algorithms is proposed. The proposed MBD algorithm uses unidirectional filters and employs normalization of gradient terms in the frequency domain. It improves convergence property without whitening effect so that it is very attractive for real world applications. A new performance measure is devised to compare BSS algorithms in real world situations without any special setup. Simulation results confirm the superior performances of the proposed algorithm over existing MBD algorithms.

Acknowledgement

This work was supported by grant No.R05-2004-000-10290-0 from Korea Science & Engineering Foundation

References

1. S. C. Douglas, A. Cichocki, S. I. Amari,, and H. H Yang, "Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach," Proc. IEEE 11th IFAC Symposium on System Identification, kitakyushu, Japan, 1057-1062, 1997.
2. T. W. Lee, A. Bell, and R. Orglmeister, "Blind source separation of real world signals," Proc. IEEE Int. Conf. Neural Networks, Houston, 2129-2135, June 1997.
3. C. Fancourt and L. Parra, "Coherence function as a criterion for blind source separation," Proc. IEEE International Workshop on Neural Networks and Signal Processing, 303-312, 2001.
4. S. I. Amari, T. P. Chen, and A. Cichocki, *Nonholonomic orthogonal learning algorithms for blind source separation*, (Neural Computation), 12, 2000.
5. X. Sun and S. Douglas, "A natural gradient convolutive blind source separation algorithm for speech mixtures," Proc. Int. Workshop on Independent Component Analysis, 2001, San Diego, California, 59-64, 2001.
6. R. H. Lambert, *Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures*, (Ph.D dissertation, University of Southern California, Los Angels, CA, 1996).
7. S. C. Douglas and S. Haykin, "Relationship between blind deconvolution and blind source separation," Unsupervised Adaptive Filtering: 2 Blind Deconvolution, Ed. S. Haykin, Wiley, 113-145, 2000.
8. S. C. Douglas, C. A. Cichocki, and S. Amari, "Self-whitening algorithm for adaptive equalization and deconvolution," IEEE Trans. Signal Processing, 47, 1161-1165, 1999.
9. S. Haykin, Adaptive Filter Theory, 4th ed., Prentice-Hall, 2002.
10. L. Zhang, A. Cichocki, and S. I. Amai, "Geometrical structures of FIR manifold and their application to multichannel blind deconvolution," Proc. IEEE Workshop on Neural Networks for Signal Processing, Madison, Wisconsin, 303-312, 1999.
11. A. Cichocki and S. I. Amari, *Adaptive Blind Signal and Image Processing: learning algorithms and applications*, (Wiley, 2002).
12. E. R. Ferrara, "Fast implementation of LMS adaptive filter," IEEE Trans. on Acoustics Speech and Signal Processing, 28, 1980.
13. K. Na, S. Kang, K. Lee, and S. Chae, "Frequency domain implementation of block adaptive filters for ica-based multichannel blind deconvolution," Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, 1999.
14. M. Joho and P. Schniter, "Frequency domain realization of a multichannel blind deconvolution algorithm based on the natural

gradient," Proc. Int Workshop on Independent Component Analysis and Signal Separation, 2003.
15. L. Parra and C. Spence, "Convolutive blind source separation of nonstationary sources," IEEE Trans. Speech and Audio Processing, 8 (3), 320-327, May 2000.
16. http://inc2.ucsd2.edu/~tewon/

[Profile]

•Hye-Jin Kim

Hey-Jin Kim received the B.S and MS. degrees from the Paichai University in 2003 and 2005, respectively, both in electrical engineering. In 2005, he joined Electro- Optical System as a research staff.

•Seung-Hyon Nam

Seung Hyon Nam received the B.S. degree from Sogang University, Korea, in 1980, the M.S. degree from the University of Alabama, Huntsville, in 1987, and the Ph.D. degree from Texas A&M University, College Station TX, in 1992. From 1979 and 1985, he was with Agency for Defense Development in Korea as a Research Engineer. Since 1993, he has been with Paichai University, Taejon, Korea, where he is currently a Professor in the Department of Electronic Engineering.

His research interests include speech and audio processing, data compression, adaptive filter theory.