TIME DOMAIN BLIND SOURCE SEPARATION OF NON-STATIONARY CONVOLVED SIGNALS BY UTILIZING GEOMETRIC BEAMFORMING

Robert Aichner[†*], Shoko Araki[†], Shoji Makino[†],
Tsuyoki Nishikawa[‡] and Hiroshi Saruwatari[‡]
[†]NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
aichner@LNT.de, {shoko, maki}@cslab.kecl.ntt.co.jp
[‡]Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara 630-0101, Japan

Abstract. We propose a time-domain BSS algorithm that utilizes geometric information such as sensor positions and assumed locations of sources. The algorithm tackles the problem of convolved mixtures by explicitly exploiting the non-stationarity of the acoustic sources. The learning rule is based on second-order statistics and is derived by natural gradient minimization. The proposed initialization of the algorithm is based on the null beamforming principle. This method leads to improved separation performance, and the algorithm is able to estimate long unmixing FIR filters in the time domain due to the geometric initialization. We also propose a post-filtering method for dewhitening which is based on the scaling technique in frequency-domain BSS. The validity of the proposed method is shown by computer simulations. Our experimental results confirm that the algorithm is capable of separating real-world speech mixtures and can be applied to short learning data sets down to a few seconds. Our results also confirm that the proposed dewhitening post-filtering method maintains the spectral content of the original speech in the separated output.

INTRODUCTION

Blind source separation (BSS) refers to the problem of recovering signals from several observed linear mixtures. The adjective "blind" stresses the fact that the source signals are not observed and that no information on the mixing process is available. The lack of *a priori* knowledge of the mixing system is compensated by a statistically strong but physically plausible assumption of independence. The weakness of the prior information is precisely the strength of the BSS model. Thus BSS has received considerable attention in the last few years, and many algorithms have been proposed [4, 5, 6, 10]. However, the separation of broadband signals in reverberant environments remains a challenging problem.

*This work was performed while the author was with the University of Applied Sciences Regensburg. Currently the author is at the Telecommunications Laboratory, University of Erlangen-Nuremberg.

0-7803-7616-1/02/$17.00 (c) 2002 IEEE

In this paper we consider the BSS of convolutive mixtures of speech. Many researchers have proposed frequency domain algorithms, which in general transform the convolutive mixture problem in the time domain to multiple instantaneous mixtures in the frequency domain. However, in [3] it was shown that for long reverberation the separation performance of frequency domain BSS degrades for long FFT frame sizes. This was our motivation in approaching the convolutive BSS problem in the time domain.

CONVOLUTIVE BSS MODEL

In real environments, where sound travels slowly compared to the distances in a typical acoustic environment, the signal arrives at the sensors with different time delays. This scenario is referred to as a multi-path environment and can be described as a finite impulse response (FIR) convolutive mixture:

$$\mathbf{x}(t) = \sum_{k=0}^{P-1} \mathbf{H}(k)\,\mathbf{s}(t-k) \tag{1}$$

where $\mathbf{s}(t) = [s_1(t), ..., s_n(t)]^T$ are the mutually independent source signals and $\mathbf{x}(t) = [x_1(t), ..., x_n(t)]^T$ are the mixed signals obtained by the microphones. The superscript ^{T} denotes transposition. The mixing system $\mathbf{H}$ is a $n \times n$ matrix consisting of channel impulse responses $h_{ji}(k)$ $(i, j = 1, ..., n)$ that are modeled by FIR filters with $k = 0, ..., P-1$. P denotes the length of the corresponding mixing filter, e.g. for a reverberation time of 300 ms and a sampling frequency of 8 kHz to $P = 2400$ taps. Equation (1) can be compactly written as

$$\mathbf{x}(t) = \mathbf{H} * \mathbf{s}(t) \tag{2}$$

where * denotes the convolution.

To obtain the estimated sources $\mathbf{y}(t)$, we seek a $n \times n$ matrix of FIR filters operating on the sensor measurements $\mathbf{x}(t)$, such that the components of the output vector $\mathbf{y}(t)$ are statistically independent:

$$\mathbf{y}(t) = \sum_{k=0}^{M-1} \mathbf{W}(k)\,\mathbf{x}(t-k) \tag{3}$$

We introduce $\overline{\mathbf{W}}(z)$ as the z-transform of the unmixing filter coefficient $\mathbf{W}(k)$ with $k = 0, ..., M-1$:

$$\begin{aligned} \overline{\mathbf{W}}(z) &= \textstyle\sum_{k=0}^{M-1} \mathbf{W}(k)\, z^{-k} \\ &= \begin{bmatrix} \sum_{k=0}^{M-1} w_{11}(k)\, z^{-k} & \cdots & \sum_{k=0}^{M-1} w_{1n}(k)\, z^{-k} \\ \vdots & \ddots & \vdots \\ \sum_{k=0}^{M-1} w_{n1}(k)\, z^{-k} & \cdots & \sum_{k=0}^{M-1} w_{nn}(k)\, z^{-k} \end{bmatrix} \end{aligned} \tag{4}$$

where M denotes the length of the unmixing filter and z^{-1} is used as the unit-delay operator for convenience, i.e., $z^{-k} \cdot x(t) = x(t-k)$. Therefore (3) can be

expressed as

$$\mathbf{y}(t) = \overline{\mathbf{W}}(z)\,\mathbf{x}(t) \tag{5}$$

The mutual independence of the sources $\mathbf{s}(t)$ is not affected by permutation or filtering. Thus, $\mathbf{y}(t)$ can only approximate $\mathbf{s}(t)$ up to an unknown permutation and filtering operation. In this paper we consider a two speaker, two microphone BSS scenario as shown in Fig. 1.

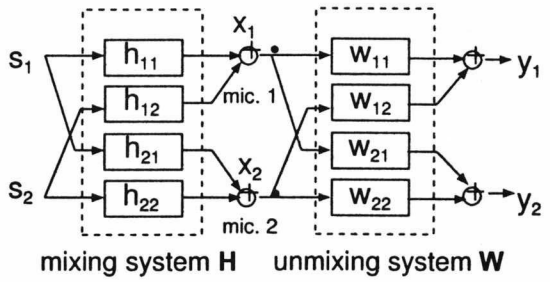

Figure 1: BSS model.

ALGORITHM DERIVATION

The assumption of independence causes the correlation matrix $\mathbf{R}_{ss}$ of the sources $\mathbf{s}(t)$ to become a diagonal matrix. Acoustic sources are assumed to be non-stationary, i.e., the auto-correlations of sources change independently with time t. Hence the correlation matrix of the outputs $\mathbf{R}_{yy}$ also varies with time t. Thus if we force estimated outputs $\mathbf{y}(t)$ to be uncorrelated at every time point t, we obtain a much stronger condition than simple decorrelation, and thus we are able to separate the sources.

We use the cost function proposed by Kawamoto [8] as a measure of uncorrelatedness. For off-line learning, this cost function has been modified with a block averaging technique:

$$Q\left(b, \overline{\mathbf{W}}(z)\right) = \frac{1}{2B}\sum_{b=1}^{B}\left\{\log\left(\det \operatorname{diag}\mathbf{R}_{yy}^{(b)}(0)\right) - \log\left(\det \mathbf{R}_{yy}^{(b)}(0)\right)\right\} \tag{6}$$

where B denotes the number of local analysis blocks, diag $\{\mathbf{X}\}$ are the diagonal elements of the matrix $\mathbf{X}$, and $\mathbf{R}_{yy}^{(b)}(\tau)$ represents the correlation matrix of $\mathbf{y}(t)$ in the b-th analysis block with time delay τ. The correlation matrix is defined by $\mathbf{R}_{yy}^{(b)}(\tau) = E_{(b)}\left[\mathbf{y}(t)\,\mathbf{y}(t+\tau)\right]$, where the expectation value $E_{(b)}[x]$ denotes the time average of x for the b-th block.

The segmentation of the observed mixtures into blocks ensures that we are calculating the cross-correlations at multiple times. The non-negative cost function becomes zero only when $y_i(t)$ and $y_j(t)$ are uncorrelated for all of the local analysis blocks, i.e., $E_{(b)}\left[y_i(t)\,y_j(t)\right] = 0$ $(i, j = 1, ..., n; i \neq j, b = 1, ..., B)$.

We use the natural gradient method, as introduced by Amari [1], to minimize the cost function.

$$\Delta\mathbf{W}(k) \propto -\frac{\partial Q\left(b, \overline{\mathbf{W}}(z)\right)}{\partial \mathbf{W}(k)}\overline{\mathbf{W}}(z)^{T}\,\overline{\mathbf{W}}(z) \quad (k = 0, ..., M-1) \tag{7}$$

where the symbol $\propto$ means "proportional to". This minimization leads to the adaptation rule [11]

$$\begin{aligned}\mathbf{W}_{i+1}(k) &= \tfrac{\alpha}{B}\sum_{b=1}^{B} \left\{\left(\left(\mathbf{R}_{yy}^{(b)}(0)^{-1}\right)^{T} - \left(\operatorname{diag}\mathbf{R}_{yy}^{(b)}(0)\right)^{-1}\right) \right. \\ &\quad \left. \cdot\mathbf{R}_{yy}^{(b)}(-k)\right\}\overline{\mathbf{W}}_i(z) + \mathbf{W}_i(k)\end{aligned} \tag{8}$$

where α is the step size parameter. Equation (8) converges if the off-diagonal components of $\mathbf{R}_{yy}^{(b)}(0)$ are minimized for all blocks. We confirmed in experiments that by considering only time-delay $\tau = 0$, we cannot achieve separation in a reverberant environment. Therefore, we expand (8) to the following equation to evaluate the off-diagonal components of $\mathbf{R}_{yy}^{(b)}(-k)$ for all time delays $k = 0, ..., M-1$ [11].

$$\begin{aligned}\mathbf{W}_{i+1}(k) = \; & \tfrac{\alpha}{B}\sum_{b=1}^{B}\left\{\left(\operatorname{diag}\mathbf{R}_{yy}^{(b)}(-k) - \mathbf{R}_{yy}^{(b)}(-k)\right)\right. \\ & \left.\cdot\left(\operatorname{diag}\mathbf{R}_{yy}^{(b)}(0)\right)^{-1}\right\}\overline{\mathbf{W}}_i(z) \\ & +\mathbf{W}_i(k),\ (k = 0, ..., M-1)\end{aligned} \tag{9}$$

This iterative algorithm is based only on second-order statistics and uses the property of non-stationarity to achieve the separation of the sources.

UTILIZATION OF GEOMETRIC BEAMFORMING

The convergence and the result of the separation of gradient-based algorithms is very much influenced by the initial value. We propose a new approach to calculate the initial value by adding geometric information on the positions of the microphones and the assumed positions of the speakers. In a recent paper of Parra and Alvino [12] they used the same approach of geometric initialization for frequency domain BSS.

The equivalence of adaptive beamformers and BSS shown in [2] was our motivation to use a beamformer technique for initializing the adaptation algorithm. As we assume the sources to be spatially separated we can employ a null beamformer with beams that place spatial zeros at the orientations of interfering sources. The performance of a null beamformer is depending on the ratio of direct to reverberant sound energy. We assume that the contribution of the direct sound prevails. For the two-speaker, two-microphone scenario, we assume two sources with angles of $\theta_i = \pm 60^o$, measured with respect to the normal of the microphone array. Figure 2 shows the configurations for calculating the null beamformer when S_1 and S_2 is the target signal, respectively. The delay D is calculated for both configurations with

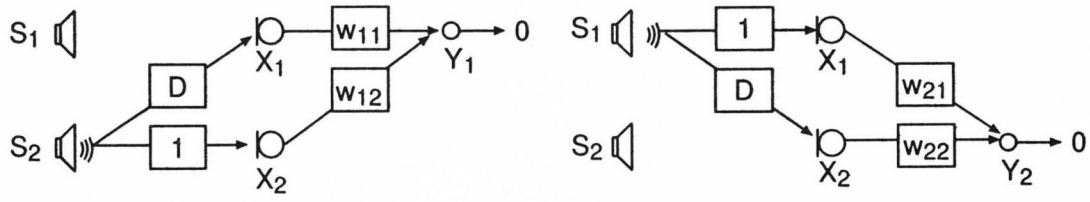

Figure 2: Null beamformer system configurations.

respect to the angle of the interfering source and the microphone positions. These delays are then used as cross path filters w_{12} and w_{21}, respectively, and subtracted from the straight path. As these delays are fractions of the sampling time, we are using a sinc function for representation. The filter in the straight path is initialized as a dirac function. The small components besides the dirac (Fig. 3) originate from neglecting frequency components smaller than 62.5 Hz. This is done because we cannot calculate a sharp spatial null for low frequencies due to the small microphone spacing of 4 cm.

When applying the system to real-world signals, we have to consider that the mixing system can be non-minimum phase. Consequently, when calculating the inverse of the mixing matrix **H**, we will not obtain a stable causal filter. However, there exists a stable *non-causal* inverse of the system. By time-delaying the initial values of the unmixing matrix **W** by M/2 taps, our algorithm accesses both future and past values of the observed signals, and thus a non-causal filter is incorporated. It should be note that the unmixing filters in BSS can only be estimated as the inverse filters of the mixing system subjected to an arbitrary filtering. Figure 3 shows the initial values of the unmixing matrix for a filter length of $M = 64$ taps. Conventionally, a unity matrix initialization ($\mathbf{W}(k) = \mathbf{I}$, $k = M/2$ and $\mathbf{W}(k) = \emptyset$, $k \neq M/2$) would be used. In the following performance comparison, we show that our new initialization method improves the separation performance significantly and that we can estimate long unmixing FIR filters in the time domain that cover the entire reverberation.

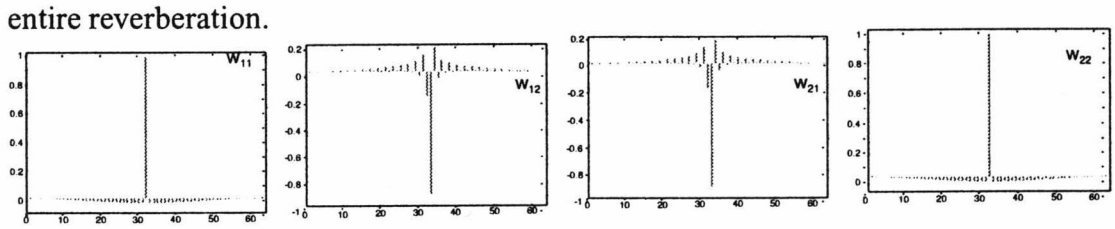

Figure 3: Initial value **W** for an unmixing filter length of $M = 64$ taps.

Note that as the initial value of the unmixing matrix **W**, we can also use the constraint null beamformer, i.e., the null beamformer which can make a spatial null towards a jammer signal and maintain the gain and phase of a target signal.

Using this initial value, the whitening effect mentioned in the following section is not so strong. Even if we do not apply our following dewhitening method, the separated signals do not have big distortions.

DEWHITENING OF THE OUTPUT SIGNALS

In the BSS of convolutive mixtures, Sun and Douglas clearly distinguished multichannel blind deconvolution from the convolutive BSS [13]. Multichannel blind deconvolution tries to make the output both spatially and temporally independent. The sources are assumed to be independent from channel to channel and from sample to sample. On the other hand, convolutive BSS tries to make the output mutually independent without deconvolution. Since speech is temporally correlated, convolutive BSS is appropriate for the task of speech separation. If we apply multichannel blind deconvolution for speech, it imposes undesirable constraints on the output, causing undesirable spectral equalization, flattening, or whitening.

It can be observed that the spectrum of the output signals is flattened due to our new initialization method. Additionally, the algorithm contributes to this whitening of the spectrum because we are minimizing all off-diagonal components of $\mathbf{R}_{yy}^{(b)}(-k)$ for $k = 0, ..., M-1$, i.e., we are removing the time-dependencies of the speech signals. This whitening of the spectrum causes the speech signals to sound unnatural.

To overcome this problem, we apply post-filters to our system. These filters are based on the method of removing the ambiguity of amplitude in frequency domain BSS proposed by Ikeda [7]. To the authors' knowledge, no post-processing method for time domain BSS using this principle has been proposed. The general idea is to transfer the separated output signals $\mathbf{y}(t)$ into the frequency domain and then solve the problem of irregular amplitude for each frequency bin. The observed mixtures $\mathbf{X}(\omega)$ are described by $\mathbf{X}(\omega) = \mathbf{H}(\omega) \cdot \mathbf{S}(\omega)$ and $\mathbf{X}(\omega) = \mathbf{W}^{-1}(\omega) \cdot \mathbf{Y}(\omega)$. We can assume that, when sources are set at almost the same distance from a microphone array, the amplitudes of all elements of the mixing filter matrix $\mathbf{H}(\omega)$ to be equal because the attenuation of all observed sound signals are nearly equal due to the small microphone spacing. The inverse of the unmixing matrix $\mathbf{W}(\omega)$ is denoted as $\mathbf{W}^{-1}(\omega) = \left[W_{ij}^{-1}(\omega)\right]$. For the two-speaker, two-microphone scenario we obtain:

$$X_1(\omega) = W_{11}^{-1}(\omega) Y_1(\omega) + W_{12}^{-1}(\omega) Y_2(\omega) \tag{10}$$

$$X_1(\omega) = 1 \cdot S_1(\omega) + 1 \cdot S_2(\omega) \tag{11}$$

$$X_2(\omega) = W_{21}^{-1}(\omega) Y_1(\omega) + W_{22}^{-1}(\omega) Y_2(\omega) \tag{12}$$

$$X_2(\omega) = 1 \cdot S_1(\omega) + 1 \cdot S_2(\omega) \tag{13}$$

We can now rescale the output signal $Y_1(\omega)$ in each frequency bin by multiplying $Y_1(\omega)$ with the factor $W_{11}^{-1}(\omega)$ so that the amplitude of $X_1(\omega)$ in (10) is equal to the amplitude of $X_1(\omega)$ in (11). The same process is applied to the output signal $Y_2(\omega)$ so that the amplitude of $X_2(\omega)$ in (12) and (13) are the same. Thus we obtain a dewhitening matrix $\mathbf{V}(\omega)$ which is defined as $\mathbf{V}(\omega) = \operatorname{diag} \mathbf{W}^{-1}(\omega)$. The dewhitened output signals $\mathbf{Y}(\omega)$ are obtained by

$$\mathbf{Y}(\omega) = \mathbf{V}(\omega) \mathbf{W}(\omega) \mathbf{X}(\omega). \tag{14}$$

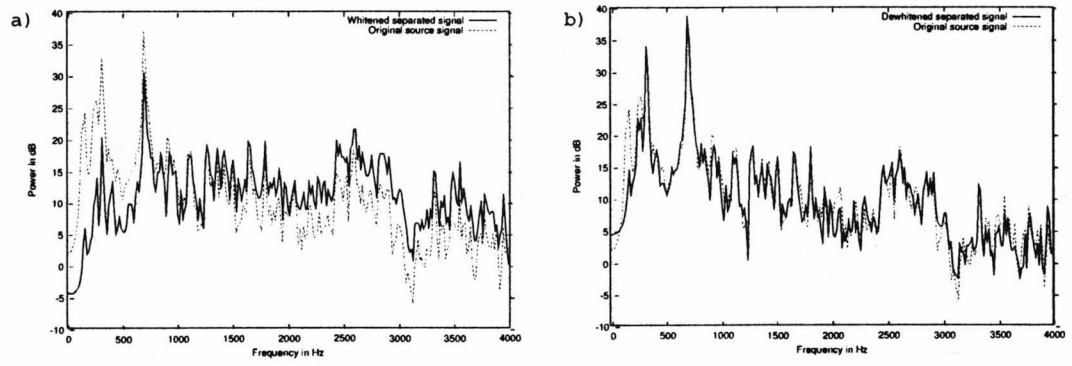

Figure 4: a) Output signal with whitened spectrum. b) Spectrum of output signal after post-processing.

The post-processing filter emphasizes low frequencies to restore the original spectral content of the source signals (Fig. 4). The algorithm without dewhitening returns too large SIR values because due to the emphasis of high frequencies, the smaller separation capability of our BSS system for low frequencies is not taken into account. Thus the dewhitening process is correcting the too large signal-to-interference ratio. We can observe that the obtained spectrum complies well with the original spectral content.

EXPERIMENTS AND RESULTS

Conditions for experiments

The experiments were conducted using speech data convolved with the room impulse responses of a real room (Fig. 5), with the reverberation time $T_R = 300$ ms. Since the sampling frequency was 8 kHz, the reverberation time corresponds to a room impulse response of 2400 taps. We used a two-element microphone array with inter-element spacing of 4 cm. The speech signals arrived from two different directions, -30^o and 40^o. Two sentences spoken by two male and two female speakers selected from the ASJ continuous speech corpus were used as source signals [9]. We used six combinations of speakers and varied the length of the signals from one to eight seconds. We used the entire length of the mixed data for learning, and the same length for separation. In order to evaluate the performance, we used the signal-to-interference ratio (SIR), defined as the ratio of the signal power of the target signal to the signal power stemming from the jammer signal. All SIR values, except the section examining the block length, were calculated by using the *dewhitened* separated output signals.

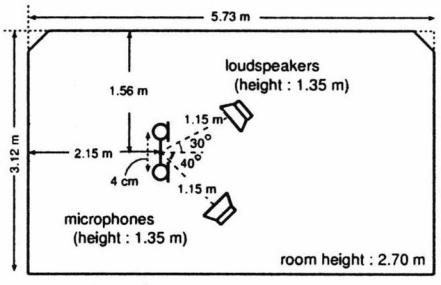

Figure 5: Layout of the reverberant room used in the experiments.

Comparison of initial values

We compared the performance of our proposed null beamforming initialization method with the conventionally used unity matrix initialization. The number of blocks B was set to 20 and the step size α was 0.1. The length of the speech data was three seconds.

It is generally known that time domain BSS works only in the case of mixtures with a short-tap FIR filter. Our experimental results in Fig. 6 confirm that the performance of conventionally initialized time domain BSS degrades for long unmixing filters. Our new null beamforming initialization method overcomes these difficulties and shows good results even when using long unmixing FIR filters. Hence we can cover the entire reverberation and achieve better separation results.

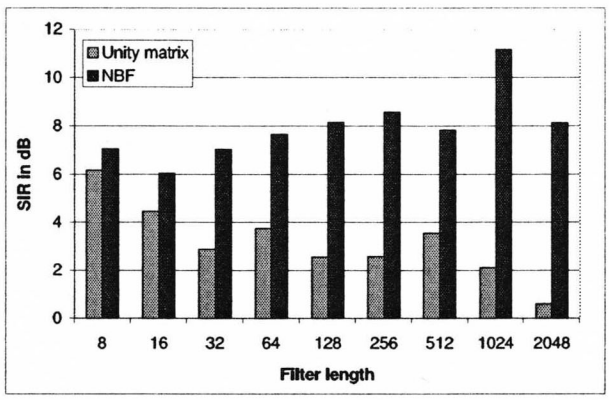

Figure 6: Relationship between SIR and filter length M for initialization with unity matrix and null beamformer (NBF).

Relationship between separation performance and block length

The adaptation rule in (9) converges only if the off-diagonal elements of $\mathbf{R}_{yy}^{(b)}(-k)$ are minimized for all $k = 0, ..., M-1$ and for all blocks $b = 1, ..., B$. As the source signals are non-stationary, we obtain a new correlation matrix $\mathbf{R}_{yy}^{(b)}$ for every block and $\mathbf{R}_{yy}^{(b)}$ has to be diagonalized for all blocks b simultaneously. Hence the separation performance depends on the number of blocks B. The update rule can be seen as an independent estimation of the unmixing matrix in each analysis block:

$$\begin{aligned}\mathbf{W}_{i+1}(k) = \tfrac{\alpha}{B} \quad & \{\Delta\mathbf{W}_{b=1}(k) + \\ & + \Delta\mathbf{W}_{b=2}(k) + \\ & \vdots \\ & + \Delta\mathbf{W}_{b=B}(k)\} + \mathbf{W}_i(k)\end{aligned} \tag{15}$$

After estimating each $\Delta\mathbf{W}_b(k)$, we take the ensemble average over all local analysis blocks. An increase in the number of blocks B has the effect that the ensemble average will be estimated over more samples, but the time average used to compute the correlation matrix in each block will be estimated over fewer time samples, due to the smaller block length. We carried out the experiments with a data length of eight seconds, an unmixing filter of $M = 1024$ taps, and a step size set to 0.01. The obtained results in Fig. 7 show that performance increases with the number of blocks until it is saturating when the block length becomes smaller than 1200 samples (i.e., B=50). This confirms that we obtain a new condition for every block. We observed that the saturation effect of the SIR is independent of data length and filter length and appears always when the block length becomes smaller than 1200 samples. It should be noted that the values in Fig. 7 correspond to the SIR *before* the post-processing step of dewhitening.

Relationship between separation performance and unmixing filter length

We investigated the influence of the unmixing FIR filter length on the separation performance. The optimum value for the number of blocks B was chosen according

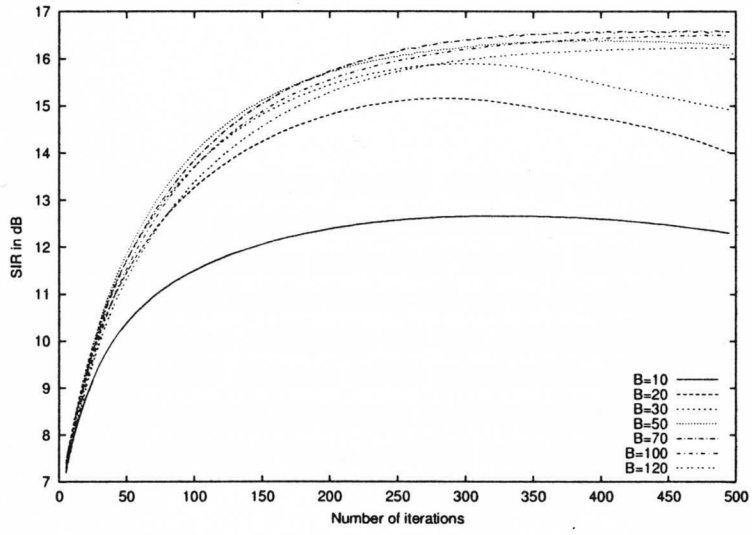

Figure 7: Relationship between separation performance and number of blocks B.

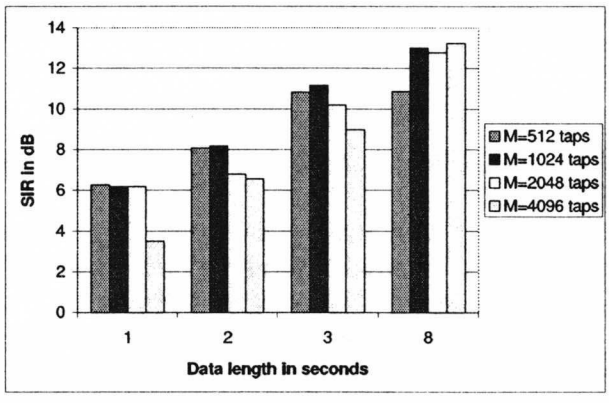

Figure 8: Relationship between separation performance and unmixing filter length for various data lengths.

to the results in the preceding section, and the step size was set to 0.01.

In Fig. 8 it can be seen that the separation performance and the optimum filter length depend on the length of the observed mixtures. This was expected because shorter data means that the mixtures can be segmented into fewer blocks. Therefore, we obtain a smaller number of correlation matrices that can be diagonalized simultaneously and hence the separation performance degrades. However, we are still able to achieve reasonable separation performance for a data length of a few seconds.

When estimating long filters, we need more conditions, i.e. more analysis blocks, to determine all filter taps. Thus for a short data length of a few seconds, a shorter filter length is more suitable.

In conventional time domain BSS, the length of the unmixing FIR filter is usually restricted to a few hundred taps. To the authors' knowledge, no algorithm has been presented which can use very long unmixing filters to deal with highly reverberant environments. Our method is capable of estimating FIR filters on the order of a couple of thousand taps. Figure 8 shows that if the data length is sufficient, we can achieve better performance with long filters that cover the entire reverberation.

CONCLUSION

In this paper we propose a new time domain convolutive BSS algorithm for non-stationary sources that utilizes geometric beamforming information. The novel initialization method provides superior separation performance and allows the use of long unmixing FIR filters to cover the reverberation time. Due to the whitening effect caused by the initialization and the algorithm, we applied a post-processing method to restore the spectral content of the sources. Our simulation results show that the algorithm is highly capable of separating real-world speech signals under highly reverberant conditions and achieves good performance with short training data of a length down to a few seconds.

ACKNOWLEDGMENTS

We would like to thank Dr. Shigeru Katagiri for his continuous encouragement. The authors are also grateful to Prof. Dr.-Ing. Walter Kellermann for useful discussions.

References

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

[2] S. Araki, S. Makino, R. Mukai and H. Saruwatari, "Equivalence between Frequency Domain Blind Source Separation and Frequency Domain Adaptive Null Beamformers," in **Proc. Eurospeech '01**, Sept. 2001, pp. 2595–2598.

[3] S. Araki, S. Makino, R. Mukai and H. Saruwatari, "Fundamental limitation of frequency domain Blind Source Separation for convolutive mixture of speech," in **Proc. ICASSP '01**, May 2001.

[4] J.-F. Cardoso, "Blind signal separation: statistical principles," in **Proc. IEEE '98**, Oct. 1998, vol. 9, pp. 2009–2025.

[5] S. Haykin (ed.), **Unsupervised Adaptive Filtering**, Volume 1 Blind Source Separation, John Wiley & Sons, 2000.

[6] A. Hyvaerinen, J. Karhunen and E. Oja, **Independent Component Analysis**, John Wiley & Sons, 2001.

[7] S. Ikeda and N. Murata, "A Method of ICA in time-frequency domain," in **Proc. ICA '99**, Jan. 1999, pp. 365–371.

[8] M. Kawamoto, K. Matsuoka and N. Ohnishi, "A method of blind separation for convolved non-stationary signals," **Neurocomputing**, vol. 22, pp. 157–171, 1998.

[9] T. Kobayashi, S. Itabashi, S. Hayashi and T. Takezawa, "ASJ continuous speech corpus for research," **J. Acoust. Soc. Jpn.**, vol. 48, no. 12, pp. 888–893, 1992, (in Japanese).

[10] T. Lee, **Independent Component Analysis**, Kluwer Academic Publishers, 1998.

[11] T. Nishikawa, H. Saruwatari and K. Shikano, "Blind source separation based on multi-stage ICA using frequency-domain ICA and time-domain ICA," in **Proc. ICFS 2002**, Mar. 2002, (accepted).

[12] L. Parra and C. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," in **Proc. Neural Networks for Signal Processing**, 2001.

[13] X. Sun and S. Douglas, "A natural gradient convolutive Blind Source Separation algorithm for speech mixtures," in **Proc. ICA '01**, Dec. 2001.