PAPER

Subband-Based Blind Separation for Convolutive Mixtures of Speech

Shoko ARAKI[†a)], **Shoji MAKINO**[†], ***Members***, **Robert AICHNER**[††], ***Nonmember***, **Tsuyoki NISHIKAWA**[†††], ***and*** **Hiroshi SARUWATARI**[†††], ***Members***

SUMMARY We propose utilizing subband-based blind source separation (BSS) for convolutive mixtures of speech. This is motivated by the drawback of frequency-domain BSS, i.e., when a long frame with a fixed long frame-shift is used to cover reverberation, the number of samples in each frequency decreases and the separation performance is degraded. In subband BSS, (1) by using a moderate number of subbands, a sufficient number of samples can be held in each subband, and (2) by using FIR filters in each subband, we can manage long reverberation. We confirm that subband BSS achieves better performance than frequency-domain BSS. Moreover, subband BSS allows us to select a separation method suited to each subband. Using this advantage, we propose efficient separation procedures that consider the frequency characteristics of room reverberation and speech signals (3) by using longer unmixing filters in low frequency bands and (4) by adopting an overlap-blockshift in BSS's batch adaptation in low frequency bands. Consequently, frequency-dependent subband processing is successfully realized with the proposed subband BSS.

key words: *blind source separation, speech separation, convolutive mixtures, subband processing, frequency dependent processing*

1. Introduction

Blind source separation (BSS) is an approach that estimates original source signals $s_i(n)$ using only information on the mixed signals $x_j(n)$ observed in each input channel. We consider the BSS of speech signals in a real environment, i.e., the BSS of convolutive mixtures of speech. In a real environment, signals are filtered by the acoustic room channel. To separate such complicated mixtures, we need to estimate the unmixing filters of several thousand taps.

Several methods have been proposed for achieving the BSS of convolutive mixtures [1], [2], most of which utilize independent component analysis (ICA). To solve the convolutive BSS problem, algorithms in time and frequency domains have been proposed [3]–[11].

In time-domain BSS, ICA is directly applied to convolutive mixtures, and unmixing FIR filters are directly estimated (e.g. [3]–[5]). Therefore, the independence of outputs can be evaluated directly. However, the convergence of most time-domain BSS algorithms is generally slower than that of frequency-domain methods because the adaptation of such long filters is very complex. Computational complexity is also a problem. Moreover, most time-domain BSS algorithms have another problem: the whitening effect. Since most time-domain BSS algorithms were designed for i.i.d. signals, such algorithms try to make output signals both spatially and temporally independent [3], [7]. When applying such time-domain BSS algorithms to mixtures of speech signals, the output speech signals are whitened and sound unnatural.

By contrast, in frequency-domain BSS, mixtures are converted into the frequency domain, and ICA is applied to instantaneous mixtures in each frequency (e.g. [8]–[11]). Although we can greatly reduce computational complexity by using frequency-domain BSS, frequency-domain BSS algorithms have inherent problems, namely, permutation and scaling problems, which result in the estimated source signal being recovered with a different permutation and gain in different frequency bins. Some solutions have been provided for these problems [8], [11]–[14].

Furthermore, we have shown that performance becomes poor with frequency-domain BSS when using a long frame to estimate a long unmixing filter that can cover realistic reverberation [15], [16]. In a real environment, since impulse response changes momentarily, it is therefore preferable to estimate unmixing filters using adaptation data that are as short as possible. However, when using a longer frame for a few seconds of speech mixtures to convert signals into the frequency domain, the number of samples in each frequency bin becomes small, and therefore, we cannot correctly estimate the statistics in each frequency bin. This means that, in such a case, independence is not evaluated correctly. This is our strongest reason for employing our subband-domain BSS method.

Motivated by these facts, we propose utilizing a BSS method that employs subband processing, hereafter called subband BSS. With subband BSS, observed signals are converted into the subband domain with a filterbank and then separated in each subband using a time-domain BSS algorithm. Then unmixed signals in each subband are synthesized to obtain fullband unmixed signals. With this method, since we can choose a moderate number of subbands, we can maintain a sufficient number of samples in each subband. The subband system also allows us to estimate FIR filters as unmixing filters in each subband. Moreover, as the

Manuscript received April 14, 2005.
Manuscript revised August 2, 2005.
Final manuscript received August 26, 2005.

[†]The authors are with the NTT Communication Science Laboratories, NTT Corporation, Kyoto-fu, 619-0237 Japan.
[††]The author is with the University Erlangen-Nuremberg, Cauerstrasse 7, 91058 Erlangen, Germany.
[†††]The authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630-0192 Japan.
a) E-mail: shoko@cslab.kecl.ntt.co.jp

DOI: 10.1093/ietfec/e88–a.12.3593

unmixing filter length in each subband is shorter than that for time-domain BSS, it is easier to estimate unmixing filters than in time-domain BSS. Therefore, we can obtain unmixing filters long enough to cover reverberation. That is, the subband BSS approach copes with both the frequency-domain approach's difficulty in estimating statistics and the time-domain technique's difficulty in adapting many parameters. Confirming this point is one of our aims of this paper.

Subband BSS has other advantages. First, its permutation problem is less serious than in frequency-domain BSS. This is because the permutation problem does not occur in each subband and, therefore, there are fewer permutation problems in subband BSS. Second, subband BSS can mitigate the whitening effect, which is troublesome in a time-domain BSS algorithm usually designed for i.i.d. signals, by limiting it in each subband.

Previous studies have used subband processing for BSS [17]–[20] to reduce computational complexity. By contrast, our main aim is to maintain the number of samples in each subband so that independence is properly evaluated. Although some authors [19], [20] utilized a scalar coefficient for the unmixing system in each subband, in this paper we use FIR filters for this purpose so as to estimate sufficiently long unmixing filters to cover reverberation.

Furthermore, subband BSS allows us to select a separation method suited to each subband. Using this advantage, we propose an efficient separation procedure that considers the frequency characteristics of room reverberation and speech signals. Generally speaking, an impulse response is usually longer in low frequency bands than in high frequency bands. This makes the separation in low frequency bands difficult. Moreover, because speech signals have high power in low frequency bands, the separation performance in low frequency bands dominates the speech separation performance. Therefore, it is very important to improve separation performance in low frequency bands for speech separation. In this paper, we propose to utilize longer unmixing filters and the overlap-blockshift technique in low frequency bands.

The organization of this paper is as follows. Section 2 presents the framework for BSS of convolutive mixtures of speech. In Sect. 3, we describe the configuration of subband BSS and mention implementation issues. Section 4 reports experiments conducted to confirm the validity of subband BSS. In Sect. 5, we propose ways of improving the low frequency subband performance in which the signal to interference ratio (SIR) is worse than at high frequencies by considering the frequency characteristics of room reverberation and speech signals. The final section concludes this paper.

2. BSS of Convolutive Mixtures

2.1 Model Description

In real environments, signals are affected by reverberation and observed by microphones. Therefore, N_s signals recorded by N_m microphones are modeled as

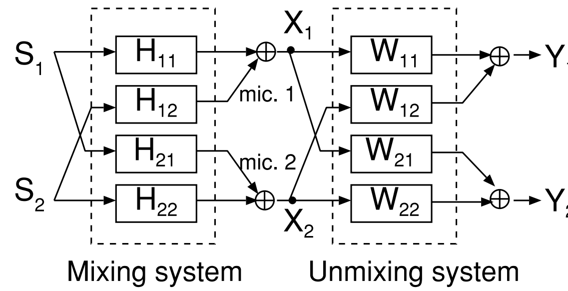

Fig. 1 BSS system configuration (when $N_s = N_m = 2$).

$$x_j(n) = \sum_{i=1}^{N_s} \sum_{k=1}^{P} h_{ji}(k) s_i(n-k+1) \ \ (j = 1, \cdots, N_m), \quad (1)$$

where s_i is the source signal from source i, x_j is the signal observed by microphone j, and h_{ji} is the P taps impulse response from source i to microphone j.

To obtain unmixed signals, we estimate the unmixing filters $w_{ij}(k)$ of Q taps, and the unmixed signals are obtained as below:

$$y_i(n) = \sum_{j=1}^{N_m} \sum_{k=1}^{Q} w_{ij}(k) x_j(n-k+1) \ \ (i = 1, \cdots, N_s). \quad (2)$$

The unmixing filters are estimated so that the unmixed signals become mutually independent.

The BSS block diagram is shown in Fig. 1 for $N_s = N_m = 2$. In this paper we consider the case of $N_s = N_m = N_{sm}$.

2.2 Frequency-Domain BSS and Related Problems

2.2.1 Frequency-Domain BSS

A frequency domain approach to convolutive mixtures transforms the problem into an instantaneous BSS problem in each frequency [8]–[11]. Using T-point short time Fourier transformation for (1), we obtain the approximate time-frequency representation of mixtures,

$$\boldsymbol{X}(\omega, m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega, m) \ \ (m = 0, \cdots, L_m - 1), \quad (3)$$

where ω denotes the frequency bin, m represents the time dependence of the short time Fourier transformation (STFT), L_m is the number of data samples in each frequency bin, $\boldsymbol{S}(\omega, m) = [S_1(\omega, m), \cdots, S_{N_{sm}}(\omega, m)]^T$ is the source signal vector, and $\boldsymbol{X}(\omega, m) = [X_1(\omega, m), \cdots, X_{N_{sm}}(\omega, m)]^T$ is the observed signal vector. We assume that $(N_{sm} \times N_{sm})$ mixing matrix $\boldsymbol{H}(\omega)$ is invertible and that $H_{ji}(\omega) \neq 0$. The STFT is usually executed by applying a window function of length T. In this paper, we call this T the frame size for STFT.

The unmixing process can be formulated in a frequency bin ω:

$$\boldsymbol{Y}(\omega, m) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, m) \ \ (m = 0, \cdots, L_m - 1), \quad (4)$$

where $\boldsymbol{Y}(\omega, m) = [Y_1(\omega, m), \cdots, Y_{N_{sm}}(\omega, m)]^T$ is the estimated source signal vector, and $\boldsymbol{W}(\omega)$ represents an $(N_{sm} \times$

N_{sm}) unmixing matrix at frequency bin ω. Here, we assume that the STFT frame size T is equal to the unmixing filter length Q. The unmixing matrix $\boldsymbol{W}(\omega)$ is determined so that outputs $Y_i(\omega, m)$ become mutually independent. This calculation is carried out independently at each frequency.

2.2.2 Problem of Frequency-Domain BSS

To handle long reverberation, we need to estimate a long unmixing filter $w_{ij}(k)$ of Q taps using learning data that are as short as possible. If the filters are relatively short, we cannot reduce the reverberant components of interferences that are longer than the filters and this has a detrimental effect on the separation performance [21]. On the other hand, with batch adaptation, it is desirable to estimate unmixing filters using adaptation data that are as short as possible, because an impulse response changes momentarily in a real environment. We therefore have to estimate long unmixing filters with speech data of short length.

We have however verified in [16] that when employing a long frame with a fixed long frame shift for several seconds of data to prepare an unmixing filter long enough to cover reverberation, it becomes difficult to maintain a sufficient number of data samples to estimate the statistics in each frequency bin. This makes the estimation of statistics difficult. In particular, independence assumption between source signals seems to collapse [16]. Therefore, we cannot obtain sufficient separation performance with a long frame with frequency-domain BSS.

3. Subband Based BSS

Motivated by the above frequency-domain BSS problem, we propose utilizing subband BSS. With subband BSS, we can choose a moderate number of subbands and therefore maintain a sufficient number of samples in each subband. Subband BSS also allows us to estimate short FIR filters as unmixing filters in each subband, due to the down-sampling procedure at the subband analysis stage. Therefore, we should be able to obtain an unmixing filter long enough to cover reverberation. Moreover, as the unmixing filter length in each subband is shorter than that for time-domain BSS, it is easier to estimate unmixing filters than in time-domain BSS. That is, the subband BSS approach offers a compromise between a time-domain technique, which is usually computationally complex and usually converges slower than a frequency-domain counterparts, and a frequency domain technique, which has difficulty estimating statistics.

3.1 Configuration of Subband BSS

The subband BSS system is composed of three parts: a subband analysis stage, a separation stage, and a subband synthesis stage (Fig. 2) [22], [23]. For the subband analysis/synthesis system, we utilize a polyphase filterbank [24] with oversampling [25] and single side band (SSB) modulation, which is widely used in echo canceller area [26], [27]. Thanks to the oversampling we can reduce the aliasing distortion, and thanks to the SSB modulation we can utilize any existing real-number BSS algorithm in each subband.

First, in the subband analysis stage, input signals $x_j(n)$ are divided into N subband signals $X_j(k, m)$ $(k = 0 \cdots, N-1)$, where k is the subband index, m is the time index, and N is the number of subbands (0–2π). Here we used a polyphase filterbank [24], that has the form of a generalized discrete Fourier transform filterbank [28]. Furthermore, to execute BSS on real-valued signals, we also used single sideband (SSB) modulation/demodulation [28] in the analysis/synthesis stage. Since signals are band-limited in each subband, we can employ decimation at the down-sampling rate R. To reduce the aliasing problem, we used a down-sampling rate of $R < N$. In this paper, SSB-modulated subband signals were not critically sampled, but two-times oversampled. That is, the down-sampling rate R was given by $R = \frac{N}{4}$. The low-pass filter used in the analysis filterbank was $f(n) = \mathrm{sinc}(\frac{n\pi}{N/2})$ of length $6N$. By using SSB modulation, we obtain SSB modulated real-valued signals $X_j^{SSB}(k, m)$ in each subband.

Then, time-domain BSS is executed on $X_j^{SSB}(k, m)$ in each subband in the separation stage. As SSB modulation is performed in the analysis stage, we can implement a time-domain BSS algorithm without expanding it into a complex value version. Since we employ down-sampling, short FIR filters of length Q/R are sufficient to separate the subband signals in each subband. Thus SSB modulated unmixed signals $Y_i^{SSB}(k, m)$ are obtained in each subband.

Finally, in the subband synthesis stage, unmixed signals $y_i(n)$ are obtained by synthesizing each unmixed signal $Y_i^{SSB}(k, m)$. The low-pass filter used in the synthesis filterbank was $g(n) = \mathrm{sinc}(\frac{n\pi}{R/2})$ of length $6R$.

3.2 Time-Domain BSS Implementation for a Separation Stage

We can use any time-domain BSS algorithm for subband BSS. Here, we describe the algorithm used in our experiment. In addition, this section describes how to design the initial value of time-domain BSS for each subband and how to solve the scaling and permutation problems.

3.2.1 Time-Domain BSS Algorithm

In this paper, we used an algorithm based on time-delayed decorrelation for non-stationary signals [4], [29], [30]. Relying on the non-stationarity and non-whiteness of the source signals, this algorithm simultaneously minimizes the cross-correlation of output signals for some time lags for all analysis blocks. We estimate FIR filters as the separation filters $w_{ij}^k(m)$ in each subband k. We write them in a matrix form $\boldsymbol{w}^k(m)$ where its ij component is $w_{ij}^k(m)$ for convenience. The adaptation rule of the i-th iteration is

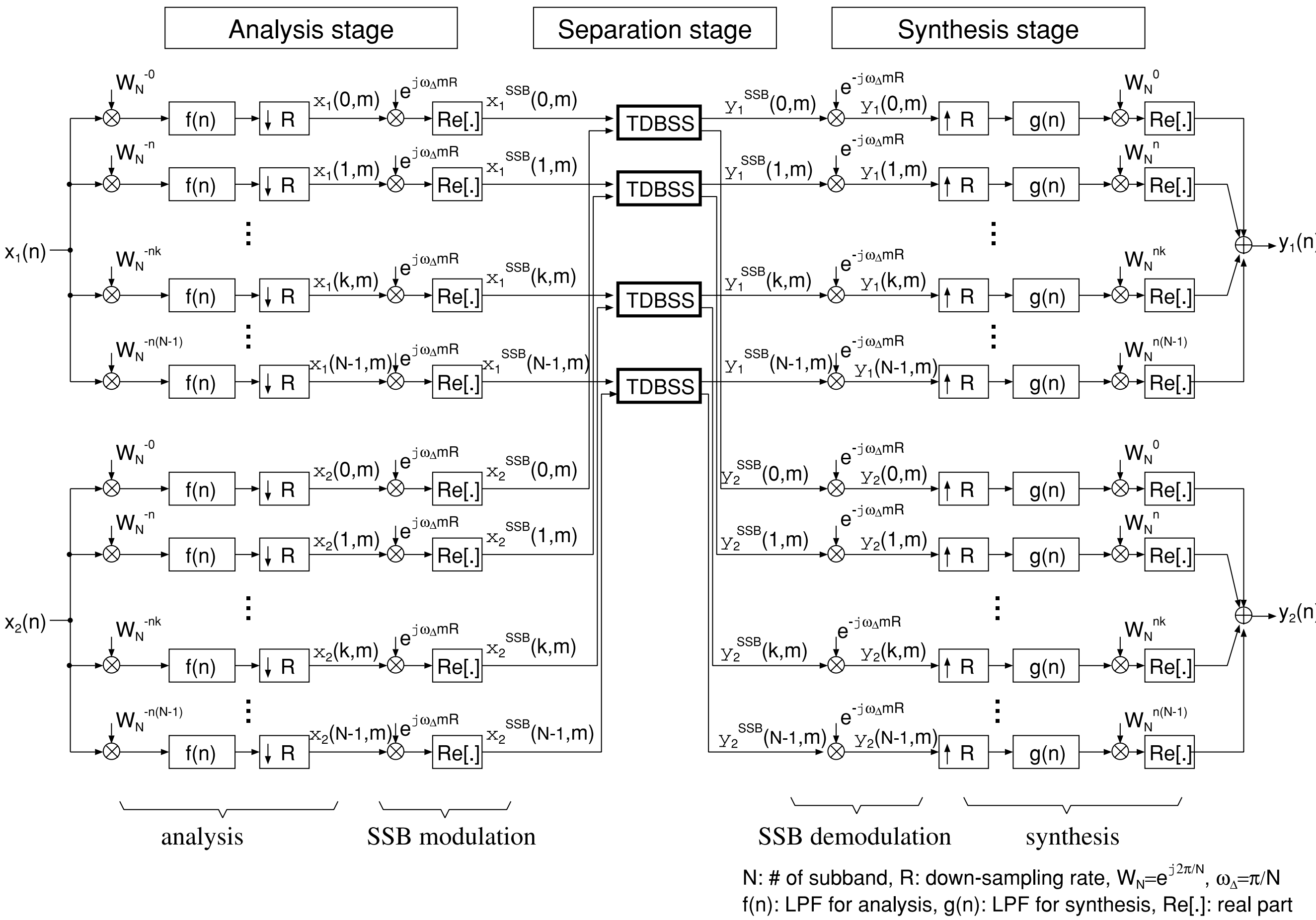

Fig. 2 System configuration of subband BSS. TDBSS denotes time-domain BSS. LPF denotes low pass filter. A 2×2 case is depicted.

$$\begin{aligned}\boldsymbol{w}_{i+1}^{k}(m) &= \boldsymbol{w}_{i}^{k}(m) \\ &\quad + \frac{\alpha}{BS}\sum_{b=0}^{BS-1}\{(\mathrm{diag}\boldsymbol{R}_{y}^{b}(0))^{-1}(\mathrm{diag}\boldsymbol{R}_{y}^{b}(m)) \\ &\quad - (\mathrm{diag}\boldsymbol{R}_{y}^{b}(0))^{-1}\boldsymbol{R}_{y}^{b}(m)\} * \boldsymbol{w}_{i}^{k}(m), \qquad (5)\end{aligned}$$

where $\boldsymbol{R}_y^b(\tau)$ represents the covariance matrix of outputs $\boldsymbol{y}(m) \equiv [Y_1^{SSB}(k,m), \cdots, Y_{N_{sm}}^{SSB}(k,m)]^T$ in the b-th (b=0, $\cdots, B-1$) analysis block with time delay τ, [i.e., $\boldsymbol{R}_y^b(\tau) = \frac{1}{L}\sum_{t=1}^{L}\boldsymbol{y}(b\frac{L}{S}+t)\boldsymbol{y}^T(b\frac{L}{S}+t-\tau)$], α denotes a step-size parameter, $*$ denotes a convolution operator, L is the block length, and S is the blockshift rate.

Note that the algorithm we used here is a *batch* algorithm, i.e., the algorithm runs by using all the data on each iteration.

3.2.2 Initial Value Design of Unmixing Filters

The initial value of the unmixing filters is very important for the convergence of time-domain BSS. Moreover, this initialization mitigates the permutation problem in frequency and subband BSS. As the initial value of the unmixing filters $\boldsymbol{w}$, we can use constraint null beamformers [31]. This is based on the fact that the BSS solution behaves as adaptive beamformers, which form a spatial null towards a jammer direction [32]. Based on this, we design null beamformers towards possible sound directions and utilize them as our initial values for BSS adaptation.

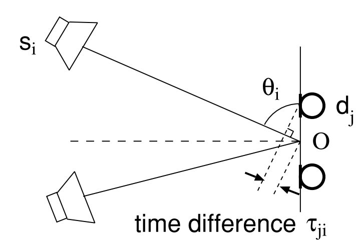

Fig. 3 Setup of a null beamformer.

Here, we assume a linear microphone array with a known microphone spacing. First, we assume that the mixing system $\boldsymbol{H}(\omega)$ represents only the time difference of sound arrival τ_{ji} with respect to the midpoint between the microphones (Fig. 3). This $\boldsymbol{H}(\omega)$ is written in the frequency domain as follows:

$$\boldsymbol{H}(\omega) = \boldsymbol{H}(2\pi f) = \begin{bmatrix} \exp(j2\pi f\tau_{11}) & \cdots & \exp(j2\pi f\tau_{1N_{sm}}) \\ \vdots & \ddots & \vdots \\ \exp(j2\pi f\tau_{N_{sm}1}) & \cdots & \exp(j2\pi f\tau_{N_{sm}N_{sm}}) \end{bmatrix}, \quad (6)$$

where $\tau_{ji} = \frac{d_j}{c}\sin\theta_i$, d_j is the position of the j-th microphone, θ_i is the direction of the i-th source as an initial value, and c is the speed of sound. Note that these d_j values need not be precise because this $\boldsymbol{H}(\omega)$ is used only for the initialization of BSS. Note also that the precise directions of sources, which are not given in a blind scenario, are not required for initialization. That is, θ_i values can be set at very rough approximations, e.g., ±60° for the 2×2 case (i.e., left or right position, for example).

Then we calculate the inverse of $\boldsymbol{H}(\omega)$ at each frequency, $\boldsymbol{W}(\omega) = \boldsymbol{H}^{-1}(\omega)$ and convert the elements $W_{ij}(\omega)$ of this $\boldsymbol{W}(\omega)$ into the time domain, $w_{ij}(n) = \mathrm{IFFT}(W_{ij}(\omega))$. This is the null beamformer that forms nulls towards θ_i, and this is the initial value for time-domain BSS. We then obtain the initial value in each subband by using subband analysis on these $w_{ij}(n)$.

3.2.3 Solving Permutation and Scaling Problems

Thanks to the initial value mentioned in Sect. 3.2.2, we did not encounter the permutation problem in our experiments. If it arises, it can be solved by reordering the row of estimated unmixing filters $\boldsymbol{w}^k(m)$ so that the null of the directivity pattern obtained by $\boldsymbol{w}^k(m)$ is sorted and forms a null in almost the same direction in all subbands [12], [33]. We can also solve the permutation problem by sorting the row of the estimated unmixing filter $\boldsymbol{w}^k(m)$ so that the cross-correlation of separated signals in adjacent subbands is maximized.

The scaling problem did occur in our experiments. That is, the estimated source signal components had different gain in different subbands. To solve it, we can also use the directivity pattern calculated with unmixing filters [34]. Our scaling method was as follows:

i) Synthesize $\boldsymbol{w}^k(m)$ to obtain $\boldsymbol{w}(n)$ in the time-domain and then obtain $\boldsymbol{W}(\omega)$ using a discrete Fourier transform (DFT).
ii) Draw the directivity gain pattern of $\boldsymbol{W}(\omega)$ [34] and obtain the estimated signal directions θ_i ($i = 1, \cdots, N_{sm}$) from the minimum of each directivity pattern. When $N_{sm} \geq 3$, they can be easily estimated using the method proposed in [35]. If the permutation problem is observed, solve it by reordering the $\boldsymbol{W}(\omega)$ row so that the θ_i values are sorted.
iii) Make null beamformers by using (6) with the estimated θ_i in step ii) and by calculating $\boldsymbol{W}(\omega) = \boldsymbol{H}^{-1}(\omega)$. We call this null beamformer $\boldsymbol{W}_{\mathrm{NBF}}(\omega)$.
iv) Calculate the inverse DFT of $\boldsymbol{W}_{\mathrm{NBF}}(\omega)$ and perform subband analysis to obtain $\boldsymbol{w}^k_{\mathrm{NBF}}(m)$.
v) Rescale $\boldsymbol{w}^k(m)$ so that $\|w^k_{ij}(m)\| = \|w^k_{\mathrm{NBF}ij}(m)\|$, where $\|x(m)\|$ means $\sum_m^{Q_k} x^2(m)$ and Q_k is the unmixing filter length in the k-th subband.

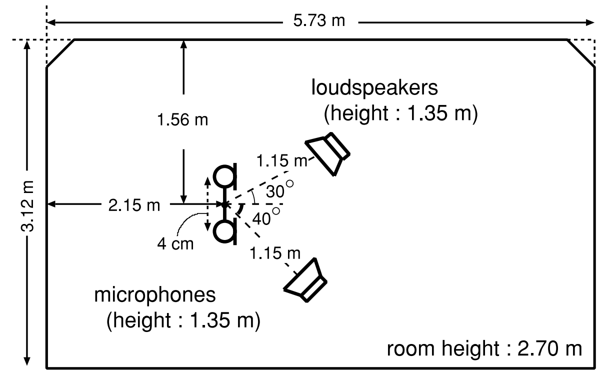

Fig. 4 Layout of room used in experiments.

4. Basic Experiments for Subband BSS

4.1 Experimental Setup

We undertook separation experiments using speech data convolved with impulse responses measured in a real environment for the 2×2 case. The impulse responses were measured in the room shown in Fig. 4. Reverberation time T_{R} was 300 ms. Since the sampling rate was 8 kHz, 300 ms corresponds to 2400 taps. As original speech, we used two sentences spoken by two male and two female speakers. Investigations were carried out for six combinations of speakers. The lengths of these mixed speech signals were about eight-seconds each. We used the first three seconds of the mixed data for learning, and we separated the entire eight second data.

To evaluate the performance, we used the signal to interference ratio (SIR), defined as

$$\mathrm{SIR}_i = \mathrm{SIR}_{\mathrm{O}i} - \mathrm{SIR}_{\mathrm{I}i} \quad (7)$$
$$\mathrm{SIR}_{\mathrm{O}i} = 10\log\frac{\sum_n \mathrm{y}^2_{is_i}(n)}{\sum_n(\sum_{j\neq i}\mathrm{y}_{is_j}(n))^2}$$
$$\mathrm{SIR}_{\mathrm{I}i} = 10\log\frac{\sum_n \mathrm{x}^2_{ks_i}(n)}{\sum_n(\sum_{j\neq i}\mathrm{x}_{ks_j}(n))^2},$$

where y_{is_j} is the output of the whole system at y_i when only s_j is active, and $\mathrm{x}_{ks_i} = \mathrm{h}_{ki} * \mathrm{s}_i$ ($*$ is a convolution operator, $k = i$ in our experiments). SIR is the ratio of a target-originated signal to jammer-originated signals.

4.2 Subband System

For subband analysis and synthesis, we used a polyphase filterbank [24] with single sideband (SSB) modulation/demodulation [28], which was mentioned in Sect. 3.1. Here, the number of subbands N was 64 and the down-sampling rate R was $16(= N/4)$. We chose this number of subbands N so that the down-sampling rate of subband BSS corresponded to that of the conventional frequency-domain BSS of frame size $T = 32$ with half frame shift (see Sect. 4.3).

For the time-domain algorithm used in subband BSS, we estimated unmixing filters w_{ij} of 64 and 128 taps in each subband. Step-size for adaptation α was 0.02, and the number of blocks B was fixed at 20 for three seconds of speech. We adopted θ_i=±60° as the initial values of the unmixing filters (see Sect. 3.2.2).

We measured the signal to distortion ratio (SDR) to evaluate the subband analysis-synthesis system. SDR is defined as

$$\mathrm{SDR} = 10\log\frac{\sum_t^{L_\delta} b^2(t-D)}{\sum_t^{L_\delta}\{b(t-D)-a(t)\}^2}\ [\mathrm{dB}], \tag{8}$$

where system input $b(t) = \delta(t - \frac{L_\delta}{2})$, L_δ is the length of the delta function, D is the delay caused by low pass filters (LPF) in the analysis and synthesis stages, and $a(t)$ is the output (impulse response) of the subband analysis-synthesis system. SDR was 59.2 dB. This distortion caused by subband analysis and synthesis can be ignored because the separation performance SIR (7) is at most 15 dB (see Sect. 4.5.1) and thus masks this distortion.

4.3 Conventional Frequency-Domain BSS

The frequency-domain BSS iteration algorithm was a natural gradient based algorithm [3]

$$\Delta \boldsymbol{W}_i(\omega) = \eta\ \left[\mathrm{diag}\left(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right]\boldsymbol{W}_i(\omega),$$

where $\boldsymbol{Y}$=$\boldsymbol{Y}(\omega, m)$, superscript H denotes a conjugate transpose, and $\langle x(m)\rangle$ denotes the time-average with respect to time m: $\frac{1}{L_m}\sum_{m=0}^{L_m-1} x(m)$. Subscript i expresses the value of the i-th step in the iterations, η is a step-size parameter, and $\Phi(\cdot)$ is a nonlinear function. As the nonlinear function $\Phi(\cdot)$, we used $\Phi(\boldsymbol{Y}) = \tanh(g \cdot \mathrm{abs}(\boldsymbol{Y}))e^{j\arg(\boldsymbol{Y})}$ [36], where g is a parameter to control nonlinearity and we utilized $g = 100$. As an initial value of the unmixing matrix, we utilized $\boldsymbol{W}(\omega) = \boldsymbol{H}^{-1}(\omega)$ with θ_i=±60° (see Sect. 3.2.2).

We fixed the frame shift at half the STFT frame size T so that the number of samples in the time-frequency domain were the same. To solve the scaling and permutation problems, we used the blind beamforming algorithm proposed by Kurita et al. [33]: first, from the directivity pattern obtained by $\boldsymbol{W}(\omega)$, we estimate the source directions and reorder the row of $\boldsymbol{W}(\omega)$ so that the directivity pattern forms a null toward the same direction in all frequency bins, and then we normalize the row of $\boldsymbol{W}(\omega)$ so that the gains of the target directions become 0 dB.

Note that we used a time-average of $\boldsymbol{Y}(\omega, m)$ of three seconds for adaptation, i.e., we used a *batch* algorithm. It should also be noted that if we fix the data length and frame shift at half the frame size, the number of samples L_m of sequences $\boldsymbol{Y}(\omega, m)$ in each frequency bin depends on frame size T: roughly speaking, $L_m \propto$ (data length)$/T$.

4.4 Conventional Fullband Time-Domain BSS

We also examined fullband time-domain BSS. The algorithm was the same as that used in subband BSS, i.e., (5).

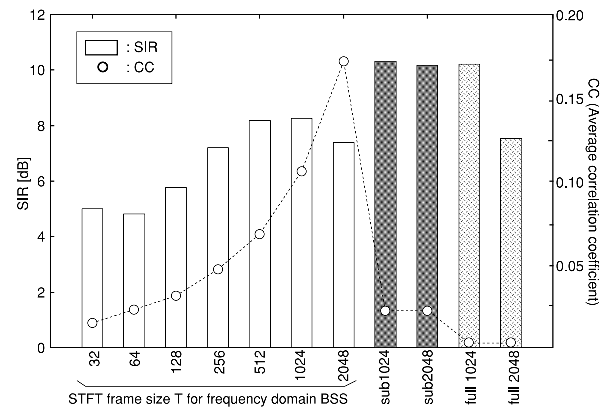

Fig. 5 Separation performance of frequency-domain BSS (white bars), subband BSS (black bars) and fullband time-domain BSS (gray bars). "CC": average correlation coefficient. T_R = 300 ms.

In this case, output signal vector $\boldsymbol{y}(n)$ was the signal in time domain $[y_1(n), \cdots, y_{N_{sm}}(n)]^T$. We used values of $\alpha = 0.002$ and $B = 20$. To obtain the initial condition of the unmixing filters, we also utilized $\boldsymbol{W}(\omega) = \boldsymbol{H}^{-1}(\omega)$ with θ_i=±60° and converted it into time domain (see Sect. 3.2.2).

In fullband time-domain BSS, output speech signals are distorted and whitened (see [37] and Sect. 4.6). We evaluated the SIR values after compensating for this whitening effect [31].

4.5 Results

4.5.1 Separation Performance of Subband BSS

In order to confirm the superiority of subband BSS over frequency- and time-domain BSS, we compared the separation performance of subband BSS with that of frequency-domain BSS.

Figure 5 shows the separation result SIR and the value of the average correlation coefficient between source signals $\mathrm{CC}(N) = \frac{1}{N}\sum_{k=1}^{N}|r_k|$, where N is the number of subbands for subband BSS or frame size for frequency-domain BSS and r_k is the correlation coefficient between source signals of a k-th frequency/subband.

For frequency-domain BSS, the parameter was the STFT frame size T. In Fig. 5, T is shown by the x axis. For subband BSS, we used unmixing filters $\boldsymbol{w}^k(m)$ of 64 and 128 taps in each subband, corresponding to 1024 and 2048 taps in a fullband, respectively. They are shown as "sub1024" and "sub2048" in Fig. 5, respectively. N = 64 subbands with decimation R = 16 correspond to T = 32 in frequency-domain BSS with regard to the down-sampling rate. The number of learning data samples in the time-frequency domain was the same for subband and frequency-domain BSS.

With frequency-domain BSS, CC becomes large and the independent assumption seems to collapse as frame size T becomes large because the number of samples in each frequency bin becomes small. Therefore, the performance degraded when we used unmixing filters of 2048 taps (i.e., frame size T = 2048). With fullband time-domain BSS, on the other hand, the CC was very small, and we obtained good results when the unmixing filter length was 1024 (see "full1024" in Fig. 5). However, when we utilized the unmixing filter length of 2048 (see "full2048" in Fig. 5), it became difficult to estimate unmixing filters, and performance was degraded.

In contrast, we achieved better separation performance in subband BSS even when we estimated unmixing filters of 2048 taps. Moreover, in subband BSS, we were able to confirm that the CC value was sufficiently small. From the CC values, we argue that the independence assumption held well in subband BSS. Another possible reason for the superior performance of subband BSS is that the permutation problem does not arise in the subbands. This point will be discussed in the next subsection.

4.6 Discussion

In the experiments, we saw that we can maintain the number of samples in each subband and obtain better separation performance.

Moreover, using subband BSS, we obtained separated signals with less distortion than when using fullband time-domain BSS. When using the usual time-domain BSS algorithm, the output signal spectrum is flattened [37] because we are removing the time dependence of the speech signals. These whitened speech signals sound unnatural. In contrast, because this whitening effect is limited to each subband, it can be diminished by subband BSS. Figure 6 shows an example of separated speech with time-domain BSS and subband BSS. The separated signal is whitened using time-domain BSS, while the shape of the spectrum holds well using subband BSS.

Furthermore, in general, the permutation problem occurs in frequency-domain BSS and subband BSS; spectral components of sources are recovered in a different order at different frequencies, although we did not face such a problem in our experiments due to the initialization with null beamformers. This makes the time domain reconstruction of separated signals difficult. However, this problem is less serious in subband BSS than in frequency-domain BSS because the permutation problem does not occur in each subband as the separation procedure is executed in each subband. Therefore, we face smaller number of permutation problems than with frequency-domain BSS. In particular, subband BSS encounters very few permutation problems in low frequency bands, where it is difficult to solve the problems with frequency-domain BSS [12]. Moreover, we can use a wider band signal than frequency-domain BSS to solve the permutation problem in between subbands. Therefore, we can use more information on separated signals and unmixing filters and can solve the problem more easily than in frequency-domain BSS.

Finally, we discuss computational cost. Because the calculation of convolution and correlation in time domain (5) is very expensive, we calculate them in the frequency domain. As discussed in [17], [18], subband processing reduces computational cost. By considering the decimation R, computational cost for N subband per time is reduced to about $(N/2 + 1)/(R \times R)$ times that of fullband time-domain BSS. As $R = N/4$ in our case, we can reduce computational cost by about $2/R$.

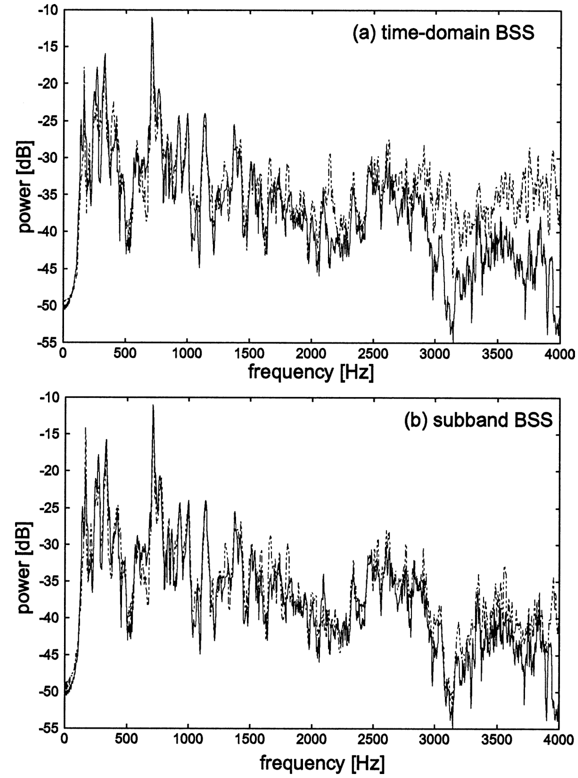

Fig. 6 Example of a spectrum of a separated signal with (a) time-domain BSS and (b) subband BSS (broken lines). Solid lines show the spectrum of the original speech.

5. Further Improvement with Frequency Appropriate Processing

In subband BSS, we can use different separation methods to estimate unmixing filter for different subbands. In this section, we propose concentrating on low frequency bands.

The SIR is generally worse in low frequency bands as shown in Fig. 7, which plots the SIR values of separated signals for each subband. One reason for poor performance at low frequencies is that the impulse response is usually longer (see Fig. 8), and therefore it is more difficult to separate signals in low frequency bands than in high frequency bands. Moreover, since speech signals have high power in low frequency bands, the performance in these bands dominates the overall speech signal separation performance. Therefore, it is important to improve separation performance in low frequency bands to obtain better overall separation performance.

From a beamforming point of view, the resolution of a spatial cancellation is proportional to the frequency. Therefore, the small phase difference between the observations at the microphones is another reason for poor performance in

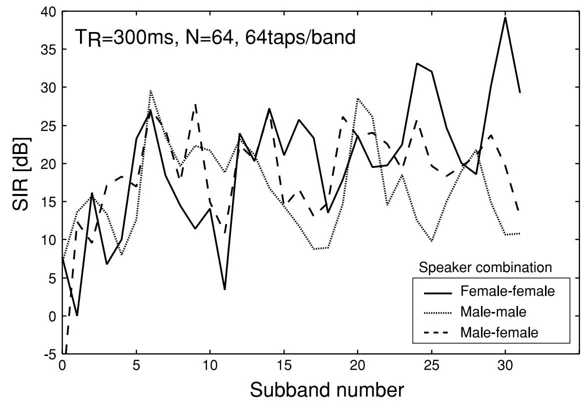

Fig. 7 SIR of separated signals in each subband. SIR is poor in low frequency bands for every speaker combination.

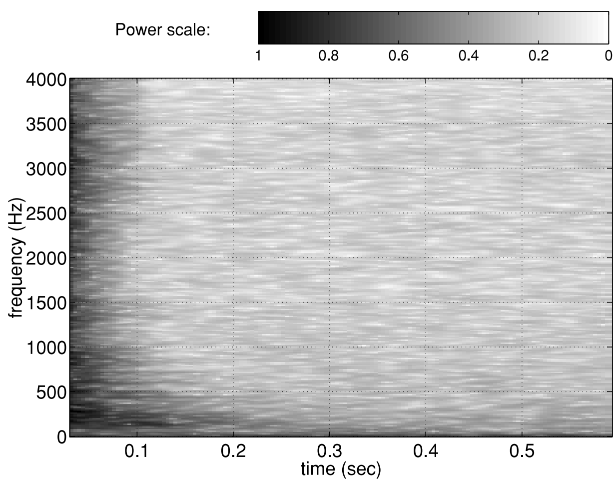

Fig. 8 Example of acoustic impulse response of a room. Black indicates high power and white indicates low power. Reverberation is longer at low frequencies than at high frequencies.

Table 1 Separation performance of subband BSS. (A)–(F): the overlap-blockshift was executed only for bands 0-5; and (G) and (H): the overlap-blockshift was executed for *all* subbands.

	# of taps		S I R [dB]		
	band 0-5	band 6-32	no-overlap	overlap (x2)	overlap (x4)
(A)	32	32	6.0		
(B)	64	32	9.9	9.8	
(C)	128	32	9.5	10.1	10.4
(D)	64	64	10.3	10.8	10.7
(E)	128	64	10.5	11.4	12.2
(F)	128	128	10.1	11.0	11.7
(G)	64	64	10.3	10.7	10.7
(H)	128	128	10.1	11.2	12.2

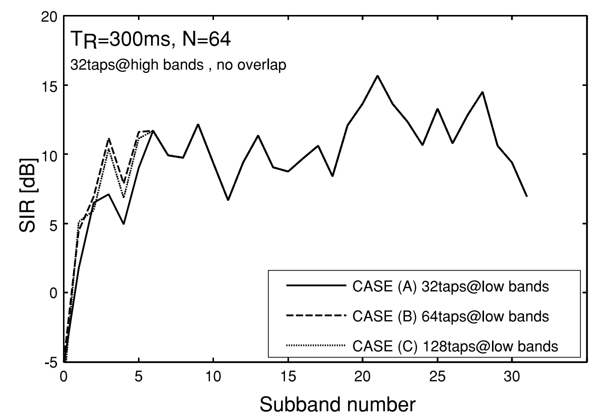

Fig. 9 Effect of the filter length for the low frequency bands.

low frequency bands. We may be able to use different microphone pairs for each subband having appropriate spacing for each subband. In this paper, however, we consider the case of N_m microphones whose number and spacing are fixed and ignore multiple spacing microphone case.

5.1 Longer Unmixing Filters in Low Frequency Bands

One possible way to improve SIR in low frequency bands is to estimate longer unmixing filters in these bands to cover reverberation. We therefore propose using longer unmixing filters for low frequency bands (bands 0-5). Figure 8 shows that the reverberation is long below about 600 Hz. Therefore, we used long filters for these frequency bands. The column labeled "no-overlap" in Table 1 shows the separation performance for each unmixing filter length condition.

In Table 1 (A)–(C), we used a 32 tap separation filter for high frequency bands, and changed the filter length for low frequency bands (bands 0-5). It is conceivable that a 32 tap long unmixing filter cannot cover reverberation in low frequency bands [see Table 1 (A)]. When we used long unmixing filters only in low frequency bands [Table 1 (B)], separation performance was greatly improved. However, when we used 128 taps in low frequency bands, separation performance degraded [see Table 1 (C)]. Figure 9 shows SIR for cases (A)–(C): the performance of (C) is worse than (B) because the number of samples in each subband is too small to allow us to precisely estimate a 128 tap unmixing filter. The proposal in the next section (Sect. 5.2) will overcome this problem.

5.2 Overlap-Blockshift in Low Frequency Bands

Another possible way to improve SIR in low frequency bands is to utilize a fine overlap-blockshift in the time-domain BSS stage for low frequency bands. Using the fine overlap-blockshift, we can outwardly increase the number of samples in each subband and estimate the unmixing filters more precisely. Since our time-domain BSS algorithm (5) divides signals into B blocks to utilize signal non-stationarity, we can divide signals into blocks with an overlap as long as non-stationarity is expressed among blocks. Note that this overlap-blockshift is executed in the separation stage, i.e., after the decimation for subband analysis.

In Table 1 [(B)–(F)], the columns show SIR obtained by the overlap-blockshift only for low frequency bands (bands 0-5). Overlap (×2) and overlap (×4) denotes the

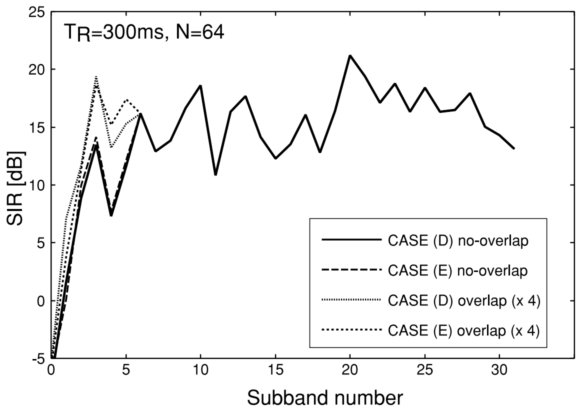

Fig. 10 Effect of the overlap-blockshift only in low frequency bands.

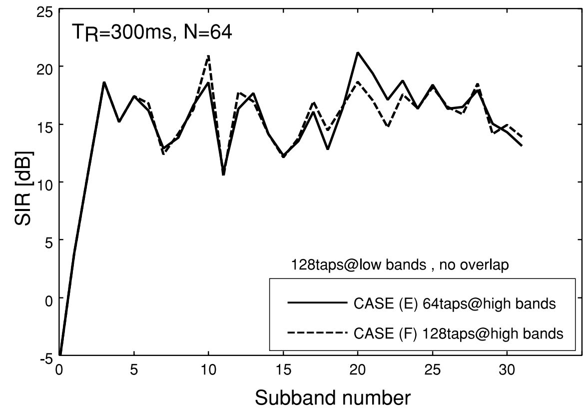

Fig. 11 Example of SIR in each subband using a long filter in *all* frequency bands.

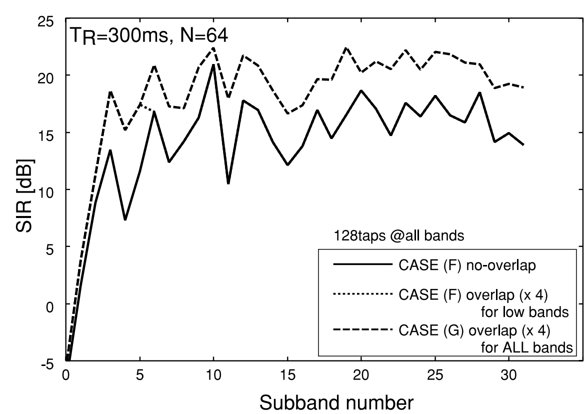

Fig. 12 Example of SIR in each subband obtained with the overlap-blockshift in *all* subbands.

block-shift rate $S = 2$ and 4 in (5), respectively. (D) and (E) in Table 1 show that when we used the overlap-blockshift only for low frequency bands, we obtained better separation performance. With a fourfold overlap-blockshift, we were able to estimate the unmixing filters of 128 taps in low frequency bands, and we obtained the best separation performance (underlined in Table 1). Figure 10 shows the effect of the fine overlap-blockshift in low frequency bands.

5.3 Discussion

Even when using 128 taps for all the frequency bands [(F) in Table 1], the performance is no better than when we used 128 taps only for the low frequency bands [(E) in Table 1]. Figure 11 shows SIR in each subband for cases (E) and (F). The use of long unmixing filters is not so effective in high frequency bands. Sometimes, short filters achieve better separation performance than long filters in high frequency bands. It is a waste of effort to use 128 taps in all subbands.

When the overlap-blockshift was used in all subbands [see (G) and (H) in Table 1], the increase in SIR was very small compared with the SIR for (D) and (F) in Table 1. The improvement in separation performance provided by the overlap-blockshift is shown in Fig. 12. The overlap-blockshift is also effective in high frequency bands. However, the contribution of the improvement in high frequency bands to SIR is not dominant because the original power of the high frequency components of the speech signal is smaller than that of the low frequency components. Therefore, we conclude that the use of a fine overlap-blockshift only in low frequencies is sufficient to obtain improved performance.

6. Conclusion

Subband processing was applied to BSS for convolutive mixtures of speech. This was motivated by the fact that separation performance is degraded when a long frame size is used for several seconds of speech in frequency-domain BSS. We showed that subband BSS can (1) maintain a sufficient number of samples to estimate the statistics in each subband and (2) estimate an unmixing filter long enough to cover reverberation. We confirmed in experiments that subband BSS is effective.

Furthermore, by efficiently using subband processing, i.e., employing an appropriate separation method for each frequency band, we showed that (3) we can improve separation performance with long unmixing filters and (4) with the overlap-blockshift technique in low frequency bands. By using long unmixing filters and the fine overlap-blockshift technique only in low frequency bands, we can efficiently separate convolutive mixtures of speech. Such frequency-dependent processing is impossible with time-domain BSS and complicated with frequency-domain BSS. Moreover, we can save computation cost without degrading separation performance by limiting the use of long unmixing filters and the fine overlap-blockshift only to low frequency bands. Subband BSS is a powerful separation tool when source signals s_i or the impulse response of system h_{ji} have different characteristics in different frequency bands.

References

[1] S. Haykin, Unsupervised Adaptive Filtering, John Wiley & Sons, 2000.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
[3] S. Amari, S.C. Douglas, A. Cichocki, and H.H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications, pp.101–104, April 1997.
[4] M. Kawamoto, A.K. Barros, A. Mansour, K. Matsuoka, and N. Ohnishi, "Real world blind separation of convolved non-stationary signals," Proc. ICA'99, pp.347–352, Jan. 1999.
[5] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics," IEEE Trans. Speech Audio Process., vol.13, no.1, pp.120–134, Jan. 2005.
[6] T.W. Lee, Independent component analysis -Theory and applications, Kluwer, 1998.
[7] S.C. Douglas, "Blind separation of acoustic signals," Microphone Arrays: Techniques and Applications, ed. M. Brandstein and D.B. Ward, pp.355–380, Springer, Berlin, 2001.
[8] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol.22, pp.21–34, 1998.
[9] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," Proc. ICA'99, pp.365–370, Jan. 1999.
[10] M.Z. Ikram and D.R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," Proc. ICASSP2000, pp.1041–1044, June 2000.
[11] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," Proc. ICA2000, pp.215–220, June 2000.
[12] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust approach to the permutation problem of frequency-domain blind source separation," Proc. ICASSP2003, pp.381–384, April 2003.
[13] K. Rahbar and J.P. Reilly, "A new fast-converging method for BSS of speech signals in acoustic environments," WASPAA2003, pp.21–24, 2003.
[14] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," Proc. ICA2001, pp.722–727, Dec. 2001.
[15] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," Proc. ICASSP2001, pp.2737–2740, May 2001.
[16] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," IEEE Trans. Speech Audio Process., vol.11, no.2, pp.109–116, 2003.
[17] J. Huang, K.C. Yen, and Y. Zhao, "Subband-based adaptive decorrelation filtering for co-channel speech separation," IEEE Trans. Speech Audio Process., vol.8, no.4, pp.402–406, July 2000.
[18] F. D-Beaulieu and B. Champagne, "Fast convolutive blind speech separation via subband adaptation," Proc. ICASSP2003, pp.513–516, 2003.
[19] N. Grbic, X.J. Tao, S.E. Nordholm, and I. Claesson, "Blind signal separation using overcomplete subband representation," IEEE Trans. Speech Audio Process., vol.9, no.5, pp.524–533, July 2001.
[20] Y. Qi, P.S. Krishnaprasad, and S. Shamma, "The subband-based independent component analysis," Proc. ICA2000, pp.199–204, June 2000.
[21] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Evaluation of separation and dereverberation performance in frequency domain blind source separation," Acoustical Science and Technology, vol.25, no.2, pp.119–126, March 2004.
[22] S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H. Saruwatari, "Blind source separation for convolutive mixtures of speech using subband processing," Proc. SMMSP2002 (International Workshop on Spectral Methods and Multirate Signal Processing), pp.195–202, Sept. 2002.
[23] S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H. Saruwatari, "Subband based blind source separation for convolutive mixtures of speech," Proc. ICASSP2003, pp.509–512, April 2003.
[24] M.R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," IEEE Trans. Acoust. Speech Signal Process., vol.24, no.3, pp.243–248, June 1976.
[25] S.L. Gay and R.J. Mammone, "Fast converging subband acoustic echo cancellation using RAP on the WE DSP16A," Proc. ICASSP90, pp.1141–1144, April 1990.
[26] P.L. Chu, "Weaver SSB subband acoustic echo canceller," Proc. IWAENC93 (International Workshop on Acoustic Echo Control), pp.173–176, Sept. 1993.
[27] S. Makino, J. Noebauer, Y. Haneda, and A. Nakagawa, "SSB subband echo canceller using low-order projection algorithm," Proc. ICASSP96, pp.945–948, May 1996.
[28] R. Crochiere and L. Rabiner, Multirate Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[29] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," IEICE Trans. Fundamentals, vol.E86-A, no.4, pp.846–858, April 2003.
[30] M. Kawamoto, K. Matsuoka, and N. Ohnishi, "A method of blind separation for convolved non-stationary signals," Neurocomputing, vol.22, pp.157–171, 1998.
[31] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming," NNSP2002, pp.445–454, 2002.
[32] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures," EURASIP Journal on Applied Signal Processing, vol.2003, no.11, pp.1157–1166, 2003.
[33] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," Proc. ICASSP2000, pp.3140–3143, June 2000.
[34] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," Proc. ICASSP2001, pp.2733–2736, May 2001.
[35] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," Seventh International Symposium on Signal Processing and its Applications (ISSPA 2003), pp.411–414, 2003.
[36] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," Proc. ICASSP2002, pp.1001–1004, May 2002.
[37] X. Sun and S. Douglas, "A natural gradient convolutive blind source separation algorithm for speech mixtures," Proc. ICA2001, pp.59–64, Dec. 2001.

Shoko Araki received the B.E. and the M.E. degrees in mathematical engineering and information physics from the University of Tokyo, Japan, in 1998 and 2000, respectively. In 2000, she joined NTT Communication Science Laboratories, Kyoto. Her research interests include array signal processing, blind source separation applied to speech signals. She received the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Best Paper Award of the IWAENC in 2003 and the 19th Awaya Prize from Acoustical Society of Japan (ASJ) in 2001. She is a member of the IEEE and the ASJ.

Shoji Makino received the B.E., M.E., and Ph.D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981. He is now an Executive Manager at the NTT Communication Science Laboratories. He is also a Guest Professor at the Hokkaido University. His research interests include adaptive filtering technologies and realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech. He received the TELECOM System Technology Award of the TAF in 2004, the Best Paper Award of the IWAENC in 2003, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002, the Achievement Award of the IEICE in 1997, and the Outstanding Technological Development Award of the ASJ in 1995. He is the author or co-author of more than 200 articles in journals and conference proceedings and has been responsible for more than 150 patents. He is a member of the Conference Board of the IEEE SP Society and an Associate Editor of the IEEE Transactions on Speech and Audio Processing. He is also an Associate Editor of the EURASIP Journal on Applied Signal Processing. He is a member of the Technical Committee on Audio and Electroacoustics of the IEEE SP Society as well as the Technical Committee on Blind Signal Processing of the IEEE Circuits and Systems Society. He is also a member of the International ICA Steering Committee and the Organizing Chair of the ICA2003 in Nara. He is the General Chair of the IWAENC2003 in Kyoto. He was a Vice Chair of the Technical Committee on Engineering Acoustics of the IEICE and the ASJ. He is an IEEE Fellow, a council member of the ASJ, and a member of the EURASIP.

Robert Aichner received the Dipl.-Ing. (FH) degree in electrical engineering from the University of Applied Sciences, Regensburg, Germany in 2002. From 2001 to 2002 he researched at the Speech Open Lab of the Research and Development division of the Nippon Telegraph and Telephone Corporation (NTT) in Kyoto, Japan. There he was working on time-domain blind source separation of audio signals. Since 2002 he is a member of the research staff at the Chair of Multimedia Communications and Signal Processing at the University of Erlangen-Nuremberg, Germany. His current research interests include multichannel adaptive algorithms for hands-free human-machine interfaces and its application to blind source separation, noise reduction, source localization, adaptive beamforming and acoustic echo cancellation.

Tsuyoki Nishikawa received the B.E. degree in electronic system and information engineering from Kinki University in 2000, the M.E. and Ph.D. degree in information and science from Nara Institute of Science and Technology (NAIST) in 2002, 2005, respectively. In 2005, he joined Audio Group, Matsushita Electric Industrial CO.,LTD. His research interests include acoustics signal processing, sensor array processing, and blind source separation. He received TELECOM System Technology Award and TELECOM System Technology Student Award from the Telecommunications Advancement Foundation in 2004, C&C Young Best Paper Award of the Foundation for C&C Promotion in 2005, and Funai Information Technology Award for Young Researchers of the Funai Foundation for Information Technology in 2005. He is a member of the the Acoustical Society of Japan.

Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO.,LTD., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Award from IEICE in 2001. He is a member of IEEE, and the Acoustical Society of Japan.