BLIND SOURCE SEPARATION FOR CONVOLUTIVE MIXTURES OF SPEECH USING SUBBAND PROCESSING

Shoko Araki † *Shoji Makino* † *Robert Aichner* †‡
*Tsuyoki Nishikawa** *and Hiroshi Saruwatari**
† NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
e-mail: shoko@cslab.kecl.ntt.co.jp
‡ University of Applied Sciences Regensburg
Seybothstrasse 2, 93053 Regensburg, Germany
*Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara 630-0101, Japan

ABSTRACT

Subband processing is applied to blind source separation (BSS) for convolutive mixtures of speech in order to overcome the drawback of frequency-domain BSS. In frequency-domain BSS, we cannot use a long frame size to cover long reverberation for several seconds of speech. This is because we cannot correctly estimate the statistics in each frequency bin if we use a long frame with short observed speech signals. In subband based BSS, (1) we can maintain the number of samples needed to estimate the statistics in each subband by using a moderate number of subbands, and (2) we can cover long reverberation by using FIR filters in each subband. In the proposed subband BSS, the permutation problem can be solved more easily than in the frequency-domain BSS. Moreover, we can avoid the whitening effect of separated signals which occurs in time-domain BSS.

1 INTRODUCTION

Blind source separation (BSS) is an approach that estimates original source signals $s_i(n)$ using only the information of the mixed signals $x_j(n)$ observed in each input channel. This technique is applicable to the realization of noise robust speech recognition and high-quality hands-free telecommunication systems. It may also become a clue to auditory scene analysis.

In this paper, we consider the BSS of speech signals in a real environment, *i.e.*, BSS of convolutive mixtures of speech. To achieve BSS of convolutive mixtures, several methods have been proposed [1, 2]. Some approaches consider the impulse responses of a room h_{ji} as FIR filters and estimate those filters in the time domain [3, 4]; other approaches transform the problem into the frequency domain to solve an instantaneous BSS problem for every frequency simultaneously [5, 6, 7].

In a real environment, an impulse response is not stable for several seconds. Therefore, we have to estimate unmixing filters with short mixed speech signals. We have shown, however, that the performance becomes poor with frequency-domain BSS when we use a long frame to estimate a long unmixing filter which can cover realistic reverberation [8]. This is because when we use a longer frame for a few seconds of speech mixtures, the number of samples in each frequency bin becomes small and, therefore, we cannot correctly estimate the statistics in each frequency bin. On the other hand, the performance of time-domain BSS for convolutive mixtures in a real environment is not good enough, either. This is because the adaptation of an unmixing filter is too complex to estimate a filter which is long enough to cover the reverberation, and because there are many local minima [9].

In this paper, we propose a method of blind source separation using subband processing in order to overcome these problems. Hereafter, we call this method subband BSS. First, we divide observed signals into a relatively small number of subbands in order to maintain a sufficient number of samples in each subband. Then, in each subband, we estimate unmixing filters which are short enough to estimate using a time-domain BSS method.

Previous work has used subband processing for BSS to deal with the frequency crossover between the adjacent frequency bins in frequency-domain BSS [10]. Moreover, subband BSS was used to reduce computational complexity [11]. Our aim, on the other hand, is to maintain the number of samples in each subband. Furthermore, some authors [10, 12] utilized a scalar coefficient for the

unmixing system in each subband, but we use FIR filters as the unmixing system in each subband so as to estimate a long enough unmixing filter to cover the reverberation.

The organization of this paper is as follows. In section 2, the framework of BSS of convolutive mixtures of speech is presented. In section 3, we explain the problems of frequency-domain BSS. Then, so as to solve these problems, we propose a blind source separation method with subband processing (subband BSS) in section 4. Experiments are conducted to confirm the validity of this method in section 5. In addition, we discuss the characteristics of subband BSS in section 6. The final section concludes this paper.

2 BLIND SOURCE SEPARATION OF CONVOLUTIVE MIXTURES OF SPEECH

2.1 Mixed signals model

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

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

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

2.2 Unmixed signals model

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

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

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

In this paper, we consider a two-input, two-output convolutive BSS problem, *i.e.*, $N = M = 2$ (see Fig. 1).

3 FREQUENCY-DOMAIN BSS and RELATED PROBLEMS

3.1 Frequency-domain BSS

The frequency domain approach to convolutive mixtures transforms the problem into an instantaneous BSS problem in the frequency domain [5, 6]. Using T-point short-time Fourier transformation for (1), we obtain

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

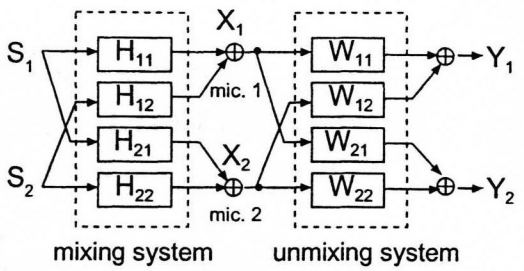

Figure 1: BSS system configuration.

where ω denotes the frequency, m represents the time-dependence of the short-time Fourier transformation, L_m is the number of data samples in each frequency bin, $\boldsymbol{S}(\omega,m) = [S_1(\omega,m), S_2(\omega,m)]^T$ is the source signal vector, and $\boldsymbol{X}(\omega,m) = [X_1(\omega,m), X_2(\omega,m)]^T$ is the observed signal vector. We assume that the (2×2) mixing matrix $\boldsymbol{H}(\omega)$ is invertible and that $H_{ji}(\omega) \neq 0$.

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

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

where $\boldsymbol{Y}(\omega,m) = [Y_1(\omega,m), Y_2(\omega,m)]^T$ is the estimated source signal vector, and $\boldsymbol{W}(\omega)$ represents a (2×2) unmixing matrix at frequency bin ω. Here, we assume that the DFT frame size T is equal to the unmixing filter length Q. $\boldsymbol{W}(\omega)$ is determined so that $Y_1(\omega,m)$ and $Y_2(\omega,m)$ become mutually independent. This calculation is carried out at each frequency independently.

3.2 Problem of frequency-domain BSS

In real environments, it is quite impossible to assume that an impulse response does not change for a period of several seconds. We therefore have to estimate unmixing filters with speech data of short length. Moreover, in order to handle long reverberation, we need to estimate a long unmixing filter $w_{ij}(k)$ of Q-taps using as short learning data as we can. We have verified, however, that the separation performance decreases when we use a long frame for several seconds of speech signals [8]. We show this result and describe the problems of frequency-domain BSS in the rest of this section.

3.2.1 Experimental setup

Separation experiments were conducted using speech data convolved with impulse responses recorded in a real environment. The layout of the room we used to measure the impulse responses is shown in Fig. 2. The reverberation time T_R was 300 ms. Since the sampling rate was 8 kHz, 300 ms corresponds to 2400 taps. As the original speech, we used two sentences spoken by

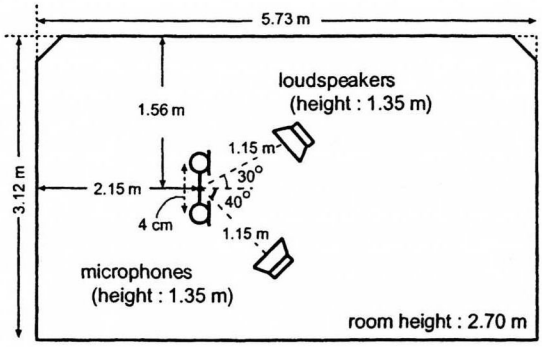

Figure 2: Layout of room used in experiments.

two male and two female speakers. Investigations were carried out for six combinations of speakers. The data length for adaptation was three seconds or about eight seconds, and the data length for separation was about eight seconds.

The frequency-domain BSS algorithm was

$$\boldsymbol{W}_{i+1}(\omega) = \boldsymbol{W}_i(\omega) + \eta\left[\mathrm{diag}\left(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right]\boldsymbol{W}_i(\omega), \quad (5)$$

where $\boldsymbol{Y}=\boldsymbol{Y}(\omega,m)$, superscript H denotes 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 is used to express the value of the i-th step in the iterations, η is a step-size parameter, and $\Phi(\cdot)$ is a non-linear function. Here we used $\Phi(\boldsymbol{Y}) = \{1+\exp(-\boldsymbol{Y}^{(\mathrm{R})})\}^{-1} + j\{1+\exp(-\boldsymbol{Y}^{(\mathrm{I})})\}^{-1}$, where $\boldsymbol{Y}^{(\mathrm{R})}$ and $\boldsymbol{Y}^{(\mathrm{I})}$ are the real and the imaginary parts of $\boldsymbol{Y}$, respectively. We fixed the frame shift as half of the frame size T, so as to make the number of data samples in the time-frequency domain equal. Note that we used the time-average of $\boldsymbol{Y}(\omega,m)$ of three seconds for adaptation, *i.e.,* we used a *batch* algorithm. Note also that if we fix the data length and frame shift as half of the frame size, the number of samples L_m of sequences $\boldsymbol{Y}(\omega,m)$ in each frequency bin depends on the frame size T: roughly speaking, $L_m \propto$ (data length)$/T$.

In order to evaluate the performance, we used the *signal to interference ratio* (SIR), defined as follows:

$$\mathrm{SIR}_i = \mathrm{SIR}_{\mathrm{O}i} - \mathrm{SIR}_{\mathrm{I}i} \quad (6)$$

$$\mathrm{SIR}_{\mathrm{O}i} = 10\log\frac{\sum_\omega |A_{ii}(\omega)S_i(\omega)|^2}{\sum_\omega |A_{ij}(\omega)S_j(\omega)|^2},$$

$$\mathrm{SIR}_{\mathrm{I}i} = 10\log\frac{\sum_\omega |H_{ii}(\omega)S_i(\omega)|^2}{\sum_\omega |H_{ij}(\omega)S_j(\omega)|^2}$$

where $\boldsymbol{A}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{H}(\omega)$ and $i \neq j$. SIR means the ratio of a target-originated signal to a jammer-originated signal.

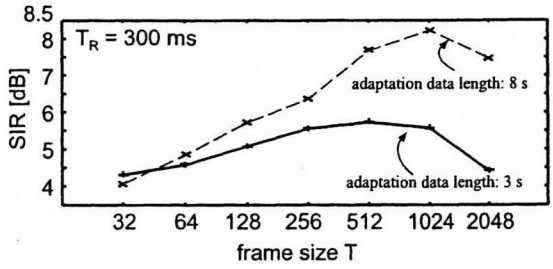

Figure 3: Example of the relationship between the frame size T and separation performance. T_{R}=300 ms.

3.2.2 Result and discussion

Figure 3 shows the relationship between the frame length T and separation performance. From Fig. 3, we can see that there is a decrease in the performance when we use a long frame size T.

The reason for this degradation of separation performance can be explained as follows. In the frequency-domain BSS framework, the signal we can use is not $x(n)$, but $\boldsymbol{X}(\omega,m)$. When we fix the frame shift (for example, at half of the frame size T) and when the frame size T is long, the number of samples in each frequency bin becomes small. This makes the estimation of statistics, like the zero mean and independent assumptions, difficult. Here, when the number of samples is too small to estimate statistics correctly, we say "the independence assumption is not held" or "the independence decreases/collapses."

To investigate the independence of two signals, we evaluated the average of the correlation coefficients r_ω over all frequency bins:

$$J(T) = \frac{1}{T}\sum_{\omega=1}^{T} |r_\omega|, \quad (7)$$

where

$$r_\omega = \frac{\langle (U_1(\omega,m) - \overline{U_1(\omega)})(U_2(\omega,m) - \overline{U_2(\omega)})\rangle}{\sqrt{\langle\{U_1(\omega,m) - \overline{U_1(\omega)}\}^2\rangle}\sqrt{\langle\{U_2(\omega,m) - \overline{U_2(\omega)}\}^2\rangle}}. \quad (8)$$

$\overline{U}$ represents a mean value, U is the source signal S, observed signal X, or separated signal Y and $\langle.\rangle$ denotes the time-average.

The solid lines of Fig. 4 show the relationship between the frame size T and $J(T)$ for a male-male speaker pair of three seconds. Note that the data length is fixed and that the frame shift is also fixed at half of the frame size; the number of data samples L_m in each frequency is different for each frame size. The independence decreases as the frame length T increases. The collapse of the independence assumption has an adverse effect on the adaptation [13].

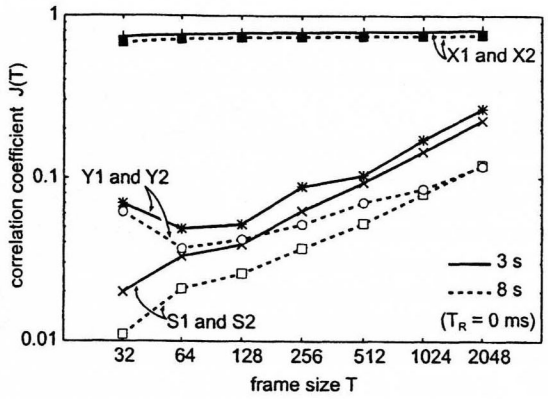

Figure 4: Relationship between the frame size T and the correlation coefficient.

In frequency-domain BSS, it is difficult to design an unmixing filter of sufficient length to cover reverberation with several seconds of speech.

4 SUBBAND BASED BLIND SOURCE SEPARATION

In the previous section, we pointed out the problems of frequency-domain BSS. In the frequency domain, when we use a longer frame in order to prepare an unmixing filter long enough to cover reverberation, it is difficult to maintain a sufficient number of data samples to estimate the statistics of the source signals in each frequency bin.

We showed that if we use longer learning data of eight seconds, we can obtain better performance using a longer frame size (dashed line in Fig. 3). However, in practice, it is quite impossible to assume that the impulse response does not change for a period of several seconds. Accordingly, we have to estimate unmixing filters which are long enough to cover reverberation using as short learning speech data as possible.

Based on these facts, we propose the use of subband processing for BSS. In this method, we can choose the number of subbands and, therefore, we can maintain a sufficient number of samples in each subband by selecting a moderate number of subbands. Subband analysis also allows us to estimate FIR filters as unmixing filters in each subband. Therefore, an unmixing filter long enough to cover reverberation should be attainable. Furthermore, as the unmixing filter length in each subband is shorter than the length of the time-domain BSS's filter, it is easier to estimate unmixing filters than in time-domain BSS.

4.1 Subband BSS

Figure 5 shows the configuration of subband BSS. The system is composed of a subband analysis stage, a BSS stage, and a subband synthesis stage.

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. We used polyphase filterbank [14] here. Because signals are band-limited in each subband, we can apply decimation at the down-sampling rate of R. In the analysis/synthesis stage, we also utilized single sideband (SSB) modulation/demodulation [15]. We get the SSB modulated signals $X_j^{SSB}(k,m)$ in each subband.

Then, time-domain BSS is executed on $X_j^{SSB}(k,m)$ in each subband. Because SSB modulation is performed on complex subbands to obtain real valued subbands, we can implement the time-domain BSS algorithm without expanding the algorithm into a complex value version. Since we apply down-sampling, short FIR filters are enough to separate the subband signals in each subband. Thus SSB modulated unmixed signals $Y_i^{SSB}(k,m)$ are obtained in each subband.

Finally, unmixed signals $y_i(n)$ are obtained by synthesizing each unmixed signal $Y_i^{SSB}(k,m)$.

4.1.1 Time-domain BSS

We can use any time-domain BSS algorithm for subband BSS. Here, we explain the algorithm we used in our experiment. To simplify the notation, $S_i^{SSB}(k,m)$, $X_j^{SSB}(k,m)$, and $Y_i^{SSB}(k,m)$ are written as $s_i(n)$, $x_j(n)$, and $y_i(n)$, respectively.

The assumption of independence causes the correlation matrix $\mathbf{R}_s$ of the sources $\mathbf{s}(n)$ to become a diagonal matrix. We assume that the source signals are non-stationary signals, *i.e.*, $\mathbf{R}_s$ changes with time n. As we are investigating the separation problem of mixed speech, this non-stationary assumption is generally true. Thus, if we force estimated outputs $\mathbf{y}(n)$ to be uncorrelated at every time point n, we obtain a much stronger condition than by simple decorrelation, and thus we can separate the sources [16].

We use the cost function $\mathcal{Q}$ proposed by Kawamoto [3] as a measure of uncorrelatedness:

$$\mathcal{Q}(\boldsymbol{W}(z)) = \frac{1}{2B}\sum_{b=1}^{B}\{\log(\det \operatorname{diag}\mathbf{R}_y^b(0)) - \log(\det\mathbf{R}_y^b)\}, \quad (9)$$

where $\mathbf{R}_y^b(0) = \langle \boldsymbol{y}(n)\boldsymbol{y}^T(n)\rangle_b$ is the correlation matrix of outputs of block B, $\boldsymbol{y}(n) = [y_1(n), y_2(n)]^T$ is the output signals, $< x >_b$ is the time-average for time interval b $(b = 1, \cdots, B)$, and B is the number of intervals. $\mathbf{W}(z)$ is the z-transform of the unmixing filter $\mathbf{W}(k)$ with $k = 0, ..., Q-1$:

$$\begin{aligned}\mathbf{W}(z) &= \textstyle\sum_{k=0}^{Q-1}\mathbf{W}(k)z^{-k} \\ &= \begin{bmatrix} \sum_{k=0}^{Q-1} w_{11}(k)z^{-k} & \sum_{k=0}^{Q-1} w_{12}(k)z^{-k} \\ \sum_{k=0}^{Q-1} w_{21}(k)z^{-k} & \sum_{k=0}^{Q-1} w_{22}(k)z^{-k} \end{bmatrix},\end{aligned} \quad (10)$$

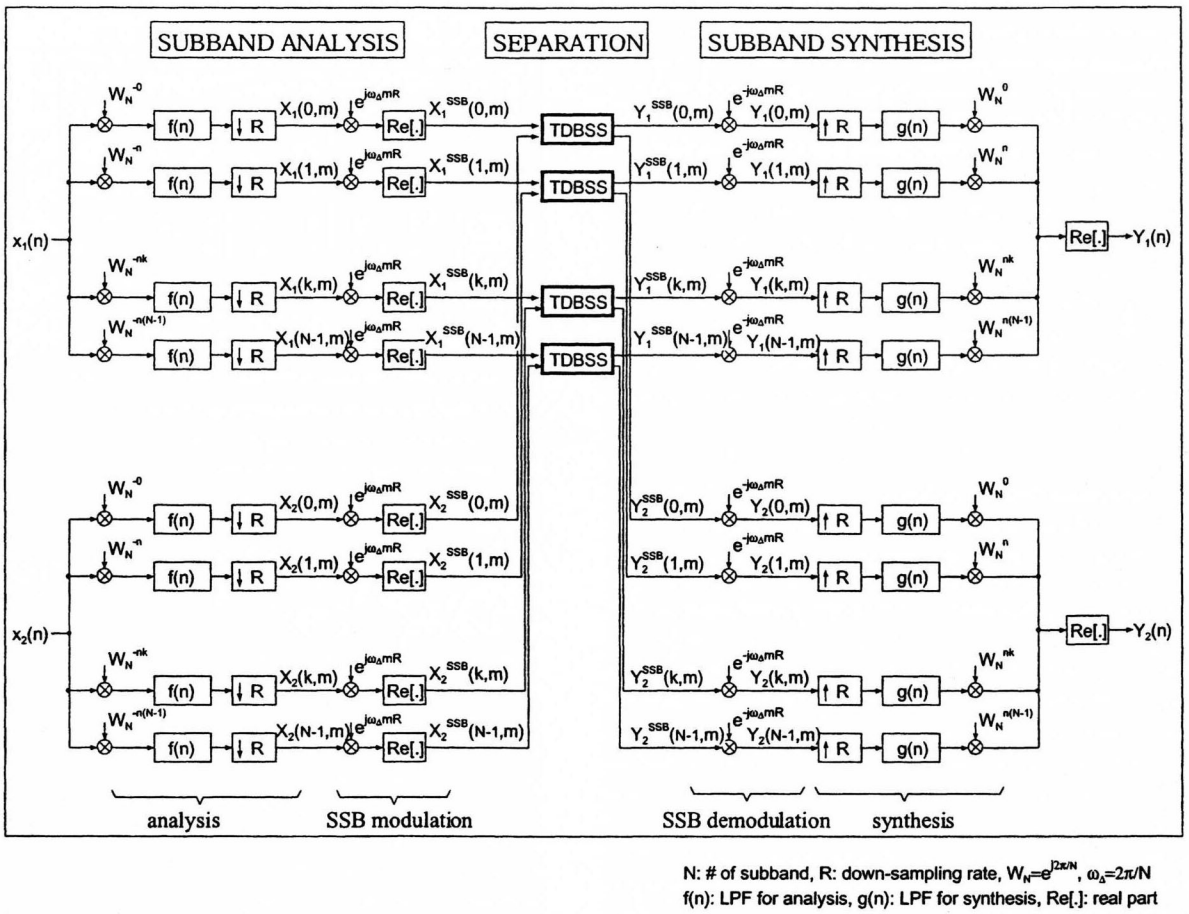

Figure 5: System configuration of subband BSS. TDBSS denotes time-domain BSS.

where Q 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, the unmixing signal $\boldsymbol{y}(n) = [y_1(n), y_2(n)]^T$ can be expressed as

$$\mathbf{y}(t) = \mathbf{W}(z)\,\mathbf{x}(t). \tag{11}$$

The error function (9) is always non-negative and takes the minimum value only when the cross-correlation of output signals becomes zero for all time intervals b.

By differentiating this $\mathcal{Q}$ with respect to $\boldsymbol{W}(z)$, considering the time-delayed components $\mathbf{R}_y(k)$, and using the natural gradient method, the iterative equation to obtain the optimal unmixing filters w_{ij} is [17]

$$\begin{aligned}\Delta \boldsymbol{w}_i(k) &= \frac{\alpha}{B}\sum_{b=1}^{B}\{(\mathrm{diag}\mathbf{R}_y^b(0))^{-1}(\mathrm{diag}\mathbf{R}_y^b(-k)) \\ &\quad -(\mathrm{diag}\mathbf{R}_y^b(0))^{-1}\mathbf{R}_y^b(-k)\}\boldsymbol{W}_i(z),\end{aligned} \tag{12}$$

where α is a step-size parameter.

5 Experiments

In order to confirm the superiority of our method, we compared the separation performance of subband BSS with that of frequency-domain BSS, using speech data convolved with impulse responses recorded in real environments.

5.1 Experimental Conditions

The impulse responses were recorded in real environments specified by different reverberation times: T_{R}=150 ms and 300 ms. Since the original sampling rate was 8 kHz, T_{R}=150 ms and 300 ms correspond to P = 1200 taps and P = 2400 taps, respectively. The layout of the room we used to measure the impulse responses is shown in Fig. 2. As the original speech, we used two sentences spoken by two male and two female speakers. We investigated three combinations of speakers: male-male, male-female, and female-female. The data length for adaptation was three seconds and for

separation it was about eight seconds. For evaluation of the separation performance, we used the signal to interference ratio (SIR) defined in section 3.2.

5.1.1 *Subband BSS*

For subband analysis and synthesis, a polyphase filterbank [14] with single sideband (SSB) modulation/demodulation was utilized. In order to avoid the aliasing influence, the SSB-modulated subband signals were not critically sampled, but two-times oversampled. That is, the down-sampling rate R was given by $R = \frac{N}{4}$, where N is the number of subbands (0-2π). The low-pass filter (LPF) used in the analysis was $f(n) = \mathrm{sinc}(\frac{n\pi}{N/2})$ of length $6 \times N$ and in the synthesis was $g(n) = \mathrm{sinc}(\frac{n\pi}{R/2})$ of length $6 \times R$. Here, the number of subbands $N = 64$ and down-sampling rate $R = 16$.

So as to evaluate the subband analysis-synthesis system, we measured the signal to distortion ratio (SDR) which is defined as

$$\mathrm{SDR} = 10 \log \frac{\sum_t^T b^2(t-D)}{\sum_t^T \{b(t-D) - a(t)\}^2} \ [\mathrm{dB}], \qquad (13)$$

where the system input $b(t) = \delta(t - \frac{T}{2})$, T is the length of the delta function, D is the delay caused by LPF in the analysis and synthesis stage, and $a(t)$ is the output (impulse response) of the subband analysis-synthesis system. The SDR was 59.2 dB. This distortion caused by subband analysis and synthesis can be ignored because the separation performance SIR (6) is at most 15 dB (see section 5.2).

For the time-domain BSS, we estimated the unmixing filters w_{ij} of 64 taps in each subband. The step-size for adaptation α was 1.0×10^{-4} and the block size B was fixed at 20 for three-second speech.

5.1.2 *Conventional frequency-domain BSS*

The frequency-domain BSS algorithm was (5) and the nonlinear function used here was $\Phi(\cdot) = \tanh(\boldsymbol{Y}^{(\mathrm{R})}) + j \tanh(\boldsymbol{Y}^{(\mathrm{I})})$.

We fixed the frame shift as half of the frame size T, so as to make the number of data samples in the time-frequency domain equal. This half shift is equivalent to $R = \frac{N}{4}$ in single sideband (SSB) filterbank. The analysis window was a Hamming window.

5.1.3 *Initial value of unmixing matrix*

We have shown that the solution of BSS behaves as adaptive beamformers, which make a spatial null towards a jammer direction [13]. Based on this fact, as the initial value of the unmixing system $\boldsymbol{w}$, we can use constraint null beamformers [18] which can make a sharp null towards a jammer direction and maintain the gain and phase of a target signal. By using this initial value, we can improve the performance of time-domain BSS

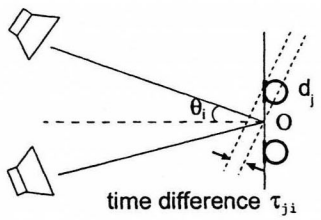

Figure 6: Setup of a null beamformer.

[18] and we can mitigate the permutation problem which occurs in frequency-domain BSS and in subband BSS.

First, we assume that the mixing system $\boldsymbol{H} = \{h_{ji}\}$ represents only the time difference of sound arrival τ_{ji} with respect to the midpoint between microphones (Fig. 6). This $\boldsymbol{H}$ is shown in the frequency domain as follows:

$$\boldsymbol{H}(\omega) = \begin{bmatrix} \exp(j\omega\tau_{11}) & \exp(j\omega\tau_{12}) \\ \exp(j\omega\tau_{21}) & \exp(j\omega\tau_{22}) \end{bmatrix}, \qquad (14)$$

where $\tau_{ji} = \frac{d_j}{c} \sin\theta_i$, d_j is the position of j-th microphone, θ_i is the direction of i-th source, and c is sound speed. Here, we gave $\theta_i = \pm 60°$ as initial values.

Then we calculate the inverse of $\boldsymbol{H}$ at each frequency, $\boldsymbol{W}(\omega) = \boldsymbol{H}^{-1}(\omega)$. For frequency-domain BSS, we used this $\boldsymbol{W}(\omega)$ as the initial value of an unmixing system. For subband BSS, we converted this $\boldsymbol{W}(\omega) = [W_{ij}(\omega)]$ into the time domain, $w_{ij}(k) = \mathrm{IFFT}(W_{ij}(\omega))$, and then got the initial value in each subband using subband analysis on these $w_{ij}(k)$.

5.1.4 *Scaling the signals and solving the permutation problem*

In frequency-domain BSS/subband BSS, the scaling and permutation problem occurs, *i.e.*, the estimated source signal components are recovered with a different order and gain in the different frequency bins. To solve these problem, we use the directivity pattern obtained by $\boldsymbol{w}$ [19]. First, we estimate the source directions from the directivity patterns in each frequency bin. Then, in order to solve the permutation problem, we reorder the row of $\boldsymbol{W}(\omega)$ so that the directivity pattern forms a null toward the same direction in all frequency bins. So as to scale the signals of each frequency bin, we normalize the row of $\boldsymbol{W}(\omega)$ so that the gains of the target directions become 0 dB in each frequency bin. Let $\hat{\boldsymbol{W}}(\omega)$ be this reordered and rescaled unmixing matrix.

For subband BSS, after we convert $\hat{\boldsymbol{W}}(\omega)$ into the time domain, we execute subband analysis. Then the unmixing filters w_{ij} are rescaled so that they have the same power as the subband analyzed rescaled unmixing filters in each subband. In subband BSS, we only rescaled the unmixing filters because the permutation ambiguity was not observed using our initial value of unmixing filters.

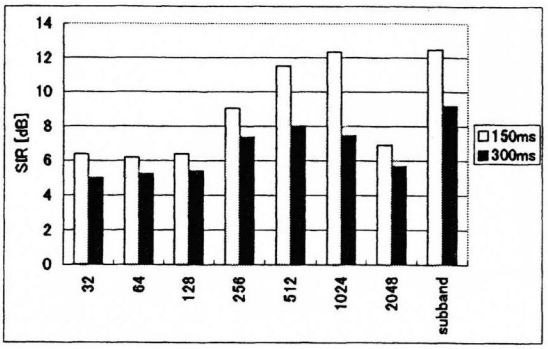

Figure 7: Separation performance of frequency-domain BSS and subband BSS. "32 $\cdots$ 2048" refers to the frame size T in frequency-domain BSS and "subband" means the performance obtained by subband BSS. The adaptation data length was 3 s and the separated data length was 8 s.

5.2 Experimental Results

Figure 7 shows the experimental result. $N = 64$ subbands with decimation $R = 16$ corresponds to $T = 32$ of frequency-domain BSS as to down-sampling rate. We used unmixing filter $\boldsymbol{w}$ of 64-taps in each subband; this corresponds to 1024-taps in full-band. Although in frequency-domain BSS, the performance degraded when we used the filter of 1024-taps (*i.e.*, frame length $T = 1024$), better separation performance was achieved in subband BSS. Moreover, the value of the averaged correlation coefficient $J(N) = \frac{1}{N}\sum_{\omega}^{N} |r_k|$, where N is the number of subbands, was 0.028 for male-male combination, 0.018 for male-female combination, and 0.020 for female-female combination. Comparing with Fig. 4, we were able to judge that the independence assumption was held well.

6 DISCUSSIONS

Using subband BSS, we can maintain the number of samples in each subband and obtain better separation performance. Using one second speech as adaptation data, we still obtained acceptable separation performance: SIR = 9.78 dB for $T_{\mathrm{R}} = 150$ ms and SIR = 7.47 dB for $T_{\mathrm{R}} = 300$ ms.

Moreover, using subband BSS, we obtained less distorted separated signals than using time-domain BSS. When we use the usual time-domain BSS algorithm, a flattened spectrum of output signals can be observed [4]. This is because we are removing the time-dependencies of speech signals. These whitened speech signals sound unnatural. On the other hand, because this whitening effect is limited in each subband, subband BSS can diminish the whitening effect. Figure 8 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 is held 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. This makes the time domain reconstruction of separated signals difficult. However, this problem is less serious in subband BSS than in frequency-domain BSS. This is because the separation procedure is executed in each subband, each of which have a wider bandwidth than frequency-domain BSS and, therefore, the permutation problem does not occur in each subband. In addition, because the number of problems is smaller in subband BSS, the permutation problem can be solved more easily in subband BSS than in frequency-domain BSS.

7 CONCLUSIONS

In frequency-domain BSS the problem of the collapse of the assumption of independence occurs when a long frame size T is used for several seconds of speech. In order to overcome this problem, we proposed subband BSS: a BSS method with subband processing. Subband BSS can (1) maintain a sufficient number of samples to estimate statistics in each subband and (2) estimate an unmixing filter long enough to cover the reverberation. We confirmed in experiments that subband BSS is effective.

8 ACKNOWLEDGEMENTS

We would like to thank Dr. Y. Haneda, Mr. A. Nakagawa, and Mr. S. Sakauchi for their help on the SSB subband. We also thank Dr. W. Kellermann for his collaboration and Dr. S. Katagiri for his continuous encouragement.

References

[1] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

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

[3] M. Kawamoto, A. K. Barros, A. Mansour, K. Matsuoka, and N. Ohnishi, "Real world blind separation of convolved non-stationary signals," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, Jan. 1999, pp. 347–352.

[4] X. Sun and S. Douglas, "A natural gradient convolutive blind source separation algorithm for speech mixtures," in *Proc. Conference Indep. Compon. Anal. Signal. Sep.*, Dec. 2001, pp. 59–64.

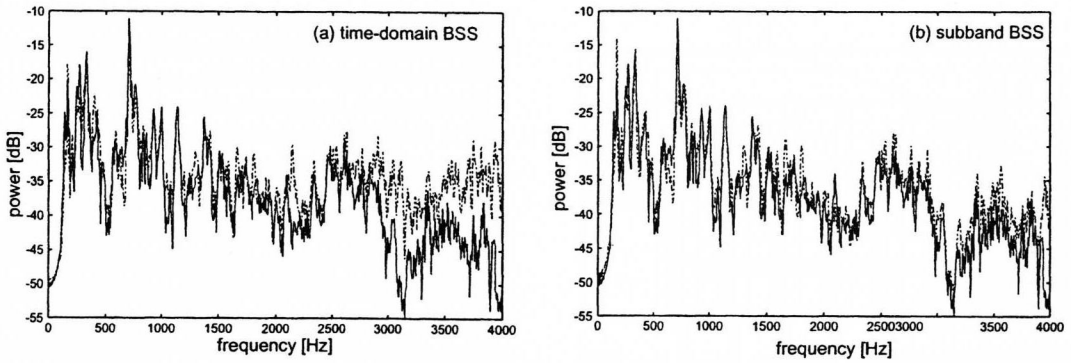

Figure 8: Example of a separated signal with (a) time-domain BSS and (b) subband BSS. Solid lines show the original speech.

[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[6] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, Jan. 1999, pp. 365–370.

[7] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[8] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," in *Proc. ICASSP2001*, May 2001, vol. 5, pp. 2737–2740.

[9] T. W. Lee, *Independent component analysis - Theory and applications*, Kluwer, 1998.

[10] N. Grbic, X-J. Tao, S. E. Nordholm, and I. Claesson, "Blind signal separation using overcomplete subband representation," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 524–533, July 2001.

[11] J. Huang, K-C. Yen, and Y. Zhao, "Subband-based adaptive decorrelation filtering for co-channel speech separation," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 4, pp. 402–406, July 2000.

[12] Y. Qi, P. S. Krishnaprasad, and S. Shamma, "The subband-based independent component analysis," in *Proc. Workshop Indep. Compon. Anal. Signal. Sep.*, June 2000, pp. 199–204.

[13] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," in *Proc. Eurospeech2001*, Sept. 2001, pp. 2595–2598.

[14] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform," *IEEE Trans. Speech Audio Processing*, vol. 24, no. 3, pp. 243–248, June 1976.

[15] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1983.

[16] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 405–413, Oct. 1993.

[17] 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, R-1, pp. 7–12.

[18] 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," in *NNSP2002*, 2002, (accepted).

[19] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," in *Proc. ICASSP2001*, May 2001, pp. 2733–2736.