SUBBAND BASED BLIND SOURCE SEPARATION FOR CONVOLUTIVE MIXTURES OF SPEECH

Shoko Araki † *Shoji Makino* † *Robert Aichner* †‡ *Tsuyoki Nishikawa* ⋆ *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 Erlangen-Nuremberg, Cauerstrasse 7, 91058 Erlangen, Germany
⋆Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan

ABSTRACT

Subband processing is applied to 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 frame-shift is used to cover reverberation, the number of samples in each frequency decreases and the separation performance is degraded. In our proposed 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 handle long reverberation. Subband BSS achieves better performance than frequency-domain BSS. Moreover, we propose efficient separation procedures that take into consideration the frequency characteristics of room reverberation and speech signals. We achieve this (3) by using longer unmixing filters in low frequency bands, and (4) by adopting overlap-blockshift in BSS's batch adaptation in low frequency bands. Consequently, frequency-dependent subband processing is successfully realized in the proposed subband BSS.

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. This technique can be used for noise robust speech recognition and high-quality hearing aid systems.

We consider the BSS of speech signals in a real environment, *i.e.,* the BSS of convolutive mixtures of speech. Several methods have been proposed for achieving the BSS of convolutive mixtures [1, 2]. In a real environment, signals are mixed with their reverberation. In order to separate such complicated mixtures, we need to estimate unmixing filters of several thousands taps. Moreover, in a real environment, an impulse response does not remain unchanged even for several seconds. Therefore, we have to estimate unmixing filters with short mixed speech signals.

In this paper, we propose a method of BSS using subband processing. Hereafter, we call this method subband BSS. Our proposal is motivated by a problem related to frequency-domain BSS systems. We have shown that the performance becomes poor with frequency-domain BSS when we use a long frame to estimate a long unmixing filter that can cover realistic reverberation [3]. 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.

Motivated by this fact, we propose the use of subband processing for BSS. In this method, we can choose a moderate number of subbands. Therefore, 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. Therefore, we can obtain an unmixing filter long enough to cover reverberation.

Previous studies have used subband processing for BSS. [4] used subband BSS to reduce computational complexity. However, their subband framework suffered from large aliasing distortion, therefore, they failed to obtain a good result. We utilize a polyphase filterbank with oversampling, which is widely used in the echo-canceller area, and our aim is to maintain the number of samples in each subband. Some other authors [5, 6] utilized a scalar coefficient for the unmixing system in each subband. However, we use FIR filters as the unmixing system in each subband so as to estimate sufficiently long unmixing filters to cover the reverberation.

Furthermore, we propose an efficient separation procedure taking into consideration the frequency characteristics of room reverberation and speech signals, *i.e.,* using longer unmixing filters and the overlap-blockshift technique only in low frequency bands.

2. BSS OF CONVOLUTIVE MIXTURES

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

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

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

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

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

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

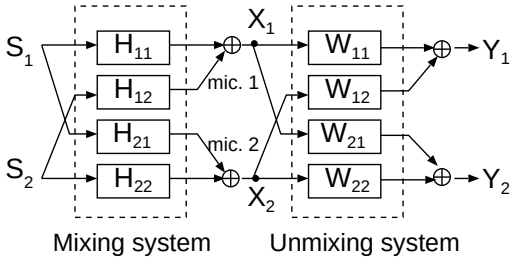

Figure 1: BSS system configuration.

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

3. SUBBAND BASED BSS

The subband BSS system is composed of three parts: a subband analysis stage, a BSS stage, and a subband synthesis stage (Fig. 2).

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 a polyphase filterbank [7] here. Because signals are band-limited in each subband, we can apply decimation at the down-sampling rate R. In the analysis/synthesis stage, we also utilized single sideband (SSB) modulation/demodulation [8]. We obtain 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 in the analysis stage, we can implement the time-domain BSS algorithm without expanding it into a complex value version. Since we employ down-sampling, short FIR filters 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, unmixed signals $y_i(n)$ are obtained by synthesizing each unmixed signal $Y_i^{SSB}(k,m)$.

3.1. 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, $X_j^{SSB}(k,m)$ and $Y_i^{SSB}(k,m)$ are written as $x_j(n)$ and $y_i(n)$, respectively.

In this paper, we used an algorithm based on time-delayed decorrelation for non-stationary signals [9]. The adaptation rule of *i-th* iteration used to obtain the optimal unmixing filters $\boldsymbol{w}(k) = \{w_{ij}(k)\}$ is

$$\Delta \boldsymbol{w}^i(k) = \frac{\alpha}{B} \sum_{b=1}^{B} \{(\mathrm{diag}\mathbf{R}_y^b(0))^{-1}(\mathrm{diag}\mathbf{R}_y^b(-k)) - (\mathrm{diag}\mathbf{R}_y^b(0))^{-1}\mathbf{R}_y^b(-k)\} * \boldsymbol{w}^i(k) \quad (3)$$

$$= -\alpha \begin{bmatrix} \frac{R_{21}(k)}{R_{11}(0)} * w_{21}^i(k) & \frac{R_{21}(k)}{R_{11}(0)} * w_{22}^i(k) \\ \frac{R_{12}(k)}{R_{22}(0)} * w_{11}^i(k) & \frac{R_{12}(k)}{R_{22}(0)} * w_{12}^i(k) \end{bmatrix}, \quad (4)$$

where $\mathbf{R}_y^b(k) = \{R_{ij}(k)\}$ represents the covariance matrix of outputs $\boldsymbol{y}(n) = [y_1(n), y_2(n)]^T$ in the *b-th* analysis block with time delay k, and α is a step-size parameter. 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. Initial value of unmixing filters

We have shown that the solution of BSS behaves as a set of adaptive beamformers, which make a spatial null towards a jammer direction [10]. Based on this fact, we can use constraint null beamformers as the initial value of the unmixing system $\boldsymbol{w}$, which can make a sharp null towards a given jammer direction and maintain the gain and phase of a given target direction. Without such an initial value, the time-domain BSS algorithm does not converge at all. Moreover, we can mitigate the permutation problem with this initial value.

To design the initial value, 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. In the frequency domain, $\boldsymbol{H}(\omega)$ is modeled as $\boldsymbol{H}(\omega) = \{h_{ji}(\omega)\} = \{\exp(j\omega\tau_{ji})\}$, 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, and c is the speed of sound. Then we calculate the inverse of $\boldsymbol{H}$ at each frequency, $\boldsymbol{W}(\omega) = \boldsymbol{H}^{-1}(\omega)$. Next, we convert this $\boldsymbol{W}(\omega) = \{W_{ij}(\omega)\}$ into the time domain, $w_{ij}(k) = \mathrm{IFFT}(W_{ij}(\omega))$, and then obtain the initial value in each subband by applying subband analysis to these $w_{ij}(k)$. Here, we gave $\theta_i = \pm 60^\circ$ as initial values.

3.3. Solving the permutation and scaling problem

Thanks of the initial value mentioned in Sec. 3.2, the permutation ambiguity was not observed. However, the scaling problem occurred, *i.e.*, the estimated source signal components had a different gain in the different subbands.

To solve this problem, we use the directivity pattern obtained by $\boldsymbol{w}$ [11]. First, we estimate the source directions from the directivity patterns in each frequency bin. In order to scale the signals of each frequency bin, we normalize the rows of $\boldsymbol{W}(\omega)$ so that the gains of the target directions become 0 dB in each frequency bin. After we convert these rescaled unmixing filters to the time domain, we execute a 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.

4. EXPERIMENTAL CONDITIONS

The impulse responses were recorded in a real room. The room size was 5.73 m × 3.12 m × 2.70 m and the distance between the loudspeakers and microphones was 1.15 m. The reverberant time was $T_R = 300$ ms. We used a two-element array with an inter-element spacing of 4 cm. The speech signals arrived from two directions, -30° and 40°. 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 was three seconds for adaptation and about eight seconds for separation. In order to evaluate the performance, we used the *signal to interference ratio* (SIR), defined as

$$\mathrm{SIR}_i = \mathrm{SIR}_{\mathrm{O}i} - \mathrm{SIR}_{\mathrm{I}i}$$

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

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

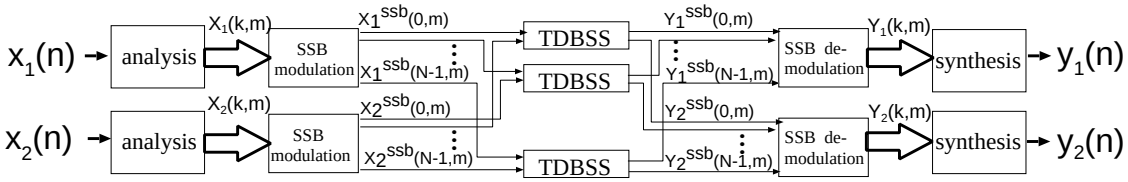

Figure 2: System configuration of subband BSS. TDBSS: time-domain BSS.

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.

4.1. Subband System

In the analysis stage, 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 used in the analysis was $f(n) = \mathrm{sinc}(\frac{n\pi}{N/2})$ of length $6N$ and in the synthesis was $g(n) = \mathrm{sinc}(\frac{n\pi}{R/2})$ of length $6R$. Here, the number of subbands $N = 64$ and the down-sampling rate $R = 16$.

For the time-domain BSS, we estimated the unmixing filters w_{ij} of 64 and 128-taps in each subband. The step-size for adaptation α was $1.0{\times}10^{-4}$ and the number of blocks B was fixed at 20 for three seconds of speech.

4.2. Conventional frequency-domain BSS

The frequency-domain BSS iteration algorithm was

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

where $\boldsymbol{Y}{=}\boldsymbol{Y}(\omega, m)$, superscript H denotes the conjugate transpose and $\langle\cdot\rangle$ denotes the time average. As the nonlinear function $\boldsymbol{\Phi}(\cdot)$, we used $\boldsymbol{\Phi}(\boldsymbol{Y}) = \tanh(g \cdot \mathrm{abs}(\boldsymbol{Y}))e^{j\arg(\boldsymbol{Y})}$, where g is a parameter to control the nonlinearity. We fixed the frame shift at a half of the DFT frame size T, so that the number of samples in the time-frequency domain were equal.

5. EXPERIMENTS AND DISCUSSIONS

5.1. Separation performance of subband BSS

Table 1 shows the separation result and the value of the average correlation coefficient CC between source signals averaged over all frequency bins and subbands. We used unmixing filters $\boldsymbol{w}$ of 64 and 128-taps in each subband; this corresponds to 1024 and 2048-taps in full-band, respectively. $N = 64$ subbands with decimation $R = 16$ corresponds to $T = 32$ in frequency-domain BSS with regard to down-sampling rate.

In frequency-domain BSS, CC becomes large and the independent assumption seems to collapses as frame size T becomes large. This is 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 $T = 2048$).

By contrast, better separation performance was achieved 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. Another possible reason for the superior performance of subband BSS is that the permutation problem does not arise in the subbands. If it occurs between subbands, it can be solved more easily than in frequency-domain BSS, simply because there are fewer problems in subband BSS than in frequency-domain BSS.

Table 1: Separation performance of frequency-domain BSS and subband BSS. $T_R = 300$ ms

	Frequency-domain BSS							Subband BSS	
T	32	64	128	256	512	1024	2048	1024	2048
SIR[dB]	5.01	4.82	5.77	7.21	8.19	8.27	7.40	9.19	9.63
CC	0.015	0.023	0.031	0.047	0.068	0.106	0.172	0.022	0.022

CC: Average correlation coefficient

5.2. Further improvement for low frequency subbands

In subband BSS, we can vary the method of estimating the unmixing filter in each subband. In this subsection, we propose a technique to improve separation performance by concentrating on low frequency bands.

Generally speaking, the SIR is worse in low frequency bands as shown in Fig. 3, in which the SIR values for each subband for three combinations of speakers are plotted. One of the reasons for poor performance at low frequencies is that an impulse response is usually longer and therefore it is more difficult to separate signals in low frequency bands than in high frequency bands. Since speech signals have high power in low frequency bands, it is important to improve the separation performance in low frequency bands to obtain better separation performance.

5.2.1. Longer unmixing filters in low frequency bands

One possible way to improve the SIR in low frequency bands is to estimate longer unmixing filters in low frequency bands. From this, we propose to using longer unmixing filters for low frequency bands (bands 0-5). The row labeled "no-overlap" in Table 2 shows the separation performance for each unmixing filter length condition. Here, we used unmixing filters of 32, 64 and 128 taps in each subband. It is conceivable that the 32-taps long unmixing filter cannot cover reverberation in low frequency bands. In this case, even when we used long unmixing filters only in low frequency bands, the separation performance was greatly improved. However, when we used 128-taps in low frequency bands, the separation performance degraded. This may be because the number of samples in each subband is too small to allow us to estimate 128-taps unmixing filter precisely. The proposal in Sec. 5.2.2 will overcome this problem.

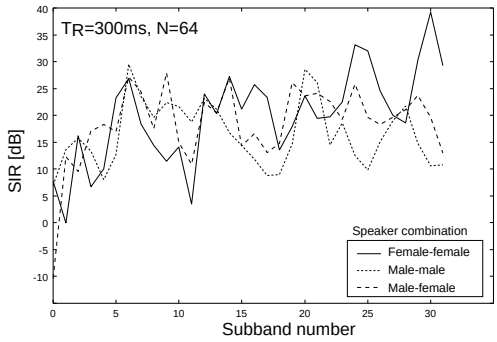

Figure 3: SIR in each subband.

5.2.2. Overlap-blockshift in low frequency bands

Another possible way to improve the SIR in low frequency bands is to utilize the overlap-blockshift in the time-domain BSS stage for low frequency bands. Using the overlap-blockshift, we can increase outwardly the number of samples in each subband, and can estimate unmixing filters more precisely. This is because we can estimate statistics correctly using sufficiently long data. Since our time-domain BSS algorithm (4) divides signals into B blocks to utilize the non-stationality of signals, we can divide signals into blocks with an overlap as long as the non-stationality is expressed among blocks. Note that this overlap-blockshift is executed in the BSS stage, *i.e.,* after the decimation for subband analysis.

In Table 2, the columns show the SIR obtained by the overlap-blockshift only for low frequency bands (bands 0-5). Overlap($\times$2) and overlap($\times$4) means that the block-shift rate was 1/2 and 1/4 of block size, respectively. When we used the overlap-blockshift only for low frequency bands, we obtained better separation performance. With four times overlap-blockshift, we can estimate the unmixing filters of 128-taps in low frequency bands, and we obtained the best separation performance. Even if we used 128-taps for all frequency bands, the performance was not increase compared to the case when we used 64-taps in bands 6-32. The use of 128-taps in all subband is the wasted effort, and the overlap-blockshift only in low frequencies is sufficient to obtain the improved performance for three seconds of speech. Furthremore, when the overlap-blockshift was used in all subband, the increase of SIR was at most 0.5 dB compared to the SIR in Table 2. It should be noted that the SIR values in this section are for 500 iterations, and the unmixing filters w_{ij} are roughly rescaled so that they have the same power as the power of the initial value in each subband.

By using long unmixing filters and the overlap-blockshift technique only in low frequency bands, we can efficiently separate the convolutive mixtures of speech. Such frequency-dependent processing is impossible in time- and frequency-domain BSS.

6. CONCLUSIONS

We proposed subband BSS: a BSS method with subband processing. This proposal was motivated by the fact that the separation performance is degraded when a long frame size is used for several seconds of speech in frequency-domain BSS. Our proposed 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. Furthermore, we showed that (3) we can improve the separation performance with long unmixing filters and (4) the overlap-blockshift technique only in low frequency bands.

Table 2: Separation performance of subband BSS. Overlap-blockshift was executed only for bands 0-5

band 0-5	band 6-32	no-overlap	overlap (x2)	overlap (x4)
32	32	5.82		
64	32	8.86	9.61	
128	32	8.71	9.75	10.09
64	64	10.31	10.80	10.79
128	64	10.28	11.21	12.01
128	128	10.28	11.22	12.00

(SIR [dB])

ACKNOWLEDGEMENTS

We would like to thank Dr. Y. Haneda, Mr. A. Nakagawa, and Mr. S. Sakauchi for their help with 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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, pp. 7–12.

[10] 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.

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