Blind Source Separation for Speech Based on Fast-Convergence Algorithm with ICA and Beamforming

Hiroshi SARUWATARI, Toshiya KAWAMURA, and Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN
E-mail: sawatari@is.aist-nara.ac.jp

Abstract

We propose a new algorithm for blind source separation (BSS), in which independent component analysis (ICA) and beamforming are combined to resolve the low-convergence problem through optimization in ICA. The proposed method consists of the following three parts: (1) frequency-domain ICA with direction-of-arrival (DOA) estimation, (2) null beamforming based on the estimated DOA, and (3) integration of (1) and (2) based on the algorithm diversity in both iteration and frequency domain. The inverse of the mixing matrix obtained by ICA is temporally substituted by the matrix based on null beamforming through iterative optimization, and the temporal alternation between ICA and beamforming can realize fast- and high-convergence optimization. The results of the signal separation experiments reveal that the signal separation performance of the proposed algorithm is superior to that of the conventional ICA-based BSS method, even under reverberant conditions.

1. Introduction

Blind source separation (BSS) is the approach taken to estimate original source signals using only the information of the mixed signals observed in each input channel. This technique is applicable to the realization of noise-robust speech recognition and high-quality hands-free telecommunication systems. In the recent works for the BSS based on the independent component analysis (ICA) [1, 2], several methods, in which the inverse of the complex mixing matrices are calculated in the frequency domain, have been proposed to deal with the arrival lags among each of the elements of the microphone array system [3, 4, 5]. However, this ICA-based approach has the disadvantage that there is difficulty with the low convergence of nonlinear optimization [6].

In this paper, we describe a new algorithm for BSS in which ICA and beamforming are combined. The proposed method consists of the following three parts: (1) frequency-domain ICA with estimation of the direction of arrival (DOA) of the sound source, (2) null beamforming based on the estimated DOA, and (3) integration of (1) and (2) based on the algorithm diversity in both iteration and frequency domain. The temporal utilization of null beamforming through ICA iterations can realize fast- and high-convergence optimization. The following sections describe the proposed method in detail, and it is shown that the signal separation performance of the proposed algorithm is superior to that of the conventional ICA-based BSS method.

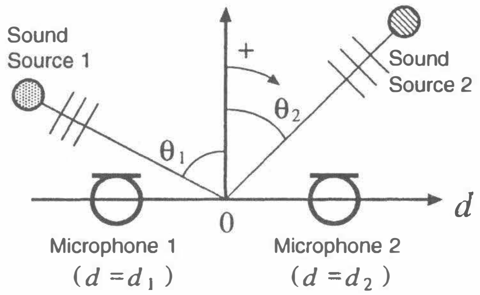

Figure 1: Configuration of a microphone array and signals.

2. Data model and conventional BSS method

In this study, a straight-line array is assumed. The coordinates of the elements are designated as d_k $(k = 1, \cdots, K)$, and the directions of arrival of multiple sound sources are designated as θ_l $(l = 1, \cdots, L)$ (see Fig. 1), where we deal with the case of $K = L = 2$.

In general, the observed signals in which multiple source signals are mixed linearly are given by the following equation in the frequency domain:

$$\boldsymbol{X}(f) = \boldsymbol{A}(f)\boldsymbol{S}(f), \tag{1}$$

where $\boldsymbol{X}(f)$ is the observed signal vector, $\boldsymbol{S}(f)$ is the source signal vector, and $\boldsymbol{A}(f)$ is the mixing matrix; these are given as

$$\boldsymbol{X}(f) = [X_1(f), \cdots, X_K(f)]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{S}(f) = [S_1(f), \cdots, S_L(f)]^{\mathrm{T}}, \tag{3}$$

$$\boldsymbol{A}(f) = \begin{bmatrix} A_{11}(f) & \cdots & A_{1L}(f) \\ \vdots & & \vdots \\ A_{K1}(f) & \cdots & A_{KL}(f) \end{bmatrix}. \tag{4}$$

$\boldsymbol{A}(f)$ is assumed to be complex-valued because we introduce a model to deal with the arrival lags among each of the elements of the microphone array and room reverberations.

In the frequency-domain ICA, first, the short-time analysis of observed signals is conducted by frame-by-frame discrete Fourier transform (DFT) (see Fig. 2). By plotting the spectral values in a frequency bin of each microphone input frame by frame, we consider them as s time series. Hereafter, we designate the time series as $\boldsymbol{X}(f,t) = [X_1(f,t), \cdots, X_K(f,t)]^{\mathrm{T}}$. Next, we perform signal separation using the complex-valued inverse of the mixing matrix, $\boldsymbol{W}(f)$, so that the L time-series output $\boldsymbol{Y}(f,t)$ becomes mutually independent; this procedure

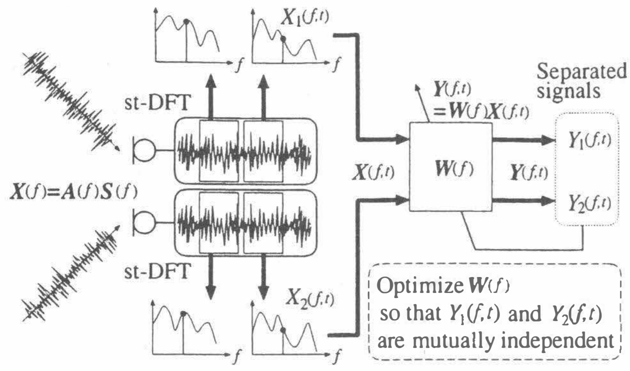

Figure 2: BSS procedure based on frequency-domain ICA.

can be given as

$$Y(f,t) = W(f)X(f,t), \tag{5}$$

where

$$Y(f,t) = [Y_1(f,t), \cdots, Y_L(f,t)]^{\mathrm{T}}, \tag{6}$$

$$W(f) = \begin{bmatrix} W_{11}(f) & \cdots & W_{1K}(f) \\ \vdots & & \vdots \\ W_{L1}(f) & \cdots & W_{LK}(f) \end{bmatrix}. \tag{7}$$

We perform this procedure with respect to all frequency bins. Finally, by applying the inverse DFT and the overlap-add technique to the separated time series $Y(f,t)$, we reconstruct the resultant source signals in the time domain.

In the conventional ICA-based BSS method, the optimal $W(f)$ is obtained by the following iterative equation [3, 7]:

$$W_{i+1}(f) = \eta\Big[\mathrm{diag}\Big(\big\langle\Phi(Y(f,t))Y^{\mathrm{H}}(f,t)\big\rangle_t\Big) - \big\langle\Phi(Y(f,t))Y^{\mathrm{H}}(f,t)\big\rangle_t\Big]W_i(f) + W_i(f), \tag{8}$$

where $\langle\cdot\rangle_t$ denotes the time-averaging operator, i is used to express the value of the i th step in the iterations, and η is the step-size parameter. Also, we define the nonlinear vector function $\Phi(\cdot)$ as

$$\Phi(Y(f,t)) \equiv [\Phi(Y_1(f,t)), \cdots, \Phi(Y_L(f,t))]^{\mathrm{T}}, \tag{9}$$

$$\Phi(Y_l(f,t)) \equiv \left[1+\exp(-Y_l^{(\mathrm{R})}(f,t))\right]^{-1} + j\cdot\left[1+\exp(-Y_l^{(\mathrm{I})}(f,t))\right]^{-1}, \tag{10}$$

where $Y_l^{(\mathrm{R})}(f,t)$ and $Y_l^{(\mathrm{I})}(f,t)$ are the real and imaginary parts of $Y_l(f,t)$, respectively.

3. Proposed algorithm

The conventional ICA method inherently has a significant disadvantage which is due to low convergence through nonlinear optimization in ICA. In order to resolve the problem, we propose an algorithm based on the temporal alternation of learning between ICA and beamforming; the inverse of the mixing matrix, $W(f)$, obtained through ICA is temporally substituted by the matrix based on null beamforming for a temporal initialization or acceleration of the iterative optimization. The proposed algorithm is conducted by the following steps with respect to all frequency bins in parallel (see Fig. 3).

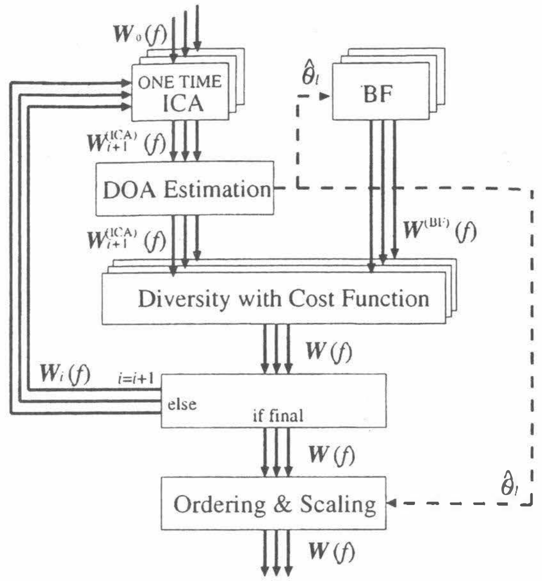

Figure 3: Proposed algorithm combining frequency-domain ICA and beamforming.

[Step 1: Initialization] Set the initial $W_i(f)$, i.e., $W_0(f)$, to an arbitrary value, where the subscripts i is set to be 0.

[Step 2: 1-time ICA iteration] Optimize $W_i(f)$ using the following 1-time ICA iteration:

$$W_{i+1}^{(\mathrm{ICA})}(f) = \eta\Big[\mathrm{diag}\Big(\big\langle\Phi(Y(f,t))Y^{\mathrm{H}}(f,t)\big\rangle_t\Big) - \big\langle\Phi(Y(f,t))Y^{\mathrm{H}}(f,t)\big\rangle_t\Big]W_i(f) + W_i(f), \tag{11}$$

where the superscript "(ICA)" is used to express that the inverse of the mixing matrix is obtained by ICA.

[Step 3: DOA estimation] Estimate DOAs of the sound sources by utilizing the directivity pattern of the array system, $F_l(f,\theta)$, which is given by

$$F_l(f,\theta) = \sum_{k=1}^{K} W_{lk}^{(\mathrm{ICA})}(f)\,\exp\left[j2\pi f d_k \sin\theta/c\right], \tag{12}$$

where $W_{lk}^{(\mathrm{ICA})}(f)$ is the element of $W_{i+1}^{(\mathrm{ICA})}(f)$. In the directivity patterns, directional nulls exist in only two particular directions. Accordingly, by obtaining statistics with respect to the directions of nulls at all frequency bins, we can estimate the DOAs of the sound sources. The DOA of the l th sound source, $\hat{\theta}_l$, can be estimated as $\hat{\theta}_l = 2\sum_{m=1}^{N/2}\theta_l(f_m)/N$, where N is a total point of DFT, and $\theta_l(f_m)$ represents the DOA of the l th sound source at the m th frequency bin. These are given by

$$\theta_1(f_m) = \min\Big[\operatorname*{argmin}_{\theta}|F_1(f_m,\theta)|, \operatorname*{argmin}_{\theta}|F_2(f_m,\theta)|\Big], \tag{13}$$

$$\theta_2(f_m) = \max\Big[\operatorname*{argmin}_{\theta}|F_1(f_m,\theta)|, \operatorname*{argmin}_{\theta}|F_2(f_m,\theta)|\Big], \tag{14}$$

where $\min[x,y]$ ($\max[x,y]$) is defined as a function in order to obtain the smaller (larger) value among x and y.

[Step 4: Beamforming] Construct an alternative matrix for signal separation, $W^{(\mathrm{BF})}(f)$, based on the null-beamforming technique where the DOA results obtained in the previous step is used. In the case that the look direction is $\hat{\theta}_1$ and the directional null is steered to $\hat{\theta}_2$, the elements of the matrix for signal separation are given as

$$\begin{aligned} W_{11}^{(\mathrm{BF})}(f_m) = & \exp\left[-j2\pi f_m d_1 \sin\hat{\theta}_1/c\right] \\ & \times\left\{\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_2-\sin\hat{\theta}_1)/c\right]\right. \\ & \left.-\exp\left[j2\pi f_m d_2(\sin\hat{\theta}_2-\sin\hat{\theta}_1)/c\right]\right\}^{-1}, \quad (15) \\ W_{12}^{(\mathrm{BF})}(f_m) = & -\exp\left[-j2\pi f_m d_2 \sin\hat{\theta}_1/c\right] \\ & \times\left\{\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_2-\sin\hat{\theta}_1)/c\right]\right. \\ & \left.-\exp\left[j2\pi f_m d_2(\sin\hat{\theta}_2-\sin\hat{\theta}_1)/c\right]\right\}^{-1}. \quad (16) \end{aligned}$$

Also, in the case that the look direction is $\hat{\theta}_2$ and the directional null is steered to $\hat{\theta}_1$, the elements of the matrix are given as

$$\begin{aligned} W_{21}^{(\mathrm{BF})}(f_m) = & -\exp\left[-j2\pi f_m d_1 \sin\hat{\theta}_2/c\right] \\ & \times\left\{-\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_1-\sin\hat{\theta}_2)/c\right]\right. \\ & \left.+\exp\left[j2\pi f_m d_2(\sin\hat{\theta}_1-\sin\hat{\theta}_2)/c\right]\right\}^{-1}, \quad (17) \\ W_{22}^{(\mathrm{BF})}(f_m) = & \exp\left[-j2\pi f_m d_2 \sin\hat{\theta}_2/c\right] \\ & \times\left\{-\exp\left[j2\pi f_m d_1(\sin\hat{\theta}_1-\sin\hat{\theta}_2)/c\right]\right. \\ & \left.+\exp\left[j2\pi f_m d_2(\sin\hat{\theta}_1-\sin\hat{\theta}_2)/c\right]\right\}^{-1}. \quad (18) \end{aligned}$$

[Step 5: Diversity with cost function] Select the most suitable unmixing matrix in each frequency bin and each iteration point, i.e., algorithm diversity in both iteration and frequency domain. As a cost function used to achieve the diversity, we calculate two kinds of cosine distances between the separated signals which are obtained by ICA and beamforming. These are given by

$$J^{(\mathrm{ICA})}(f)=\frac{\left|\left\langle Y_1^{(\mathrm{ICA})}(f,t)Y_2^{(\mathrm{ICA})}(f,t)^*\right\rangle_t\right|}{\left\langle\left|Y_1^{(\mathrm{ICA})}(f,t)\right|^2\right\rangle_t^{\frac{1}{2}}\left\langle\left|Y_2^{(\mathrm{ICA})}(f,t)\right|^2\right\rangle_t^{\frac{1}{2}}}, \quad (19)$$

$$J^{(\mathrm{BF})}(f)=\frac{\left|\left\langle Y_1^{(\mathrm{BF})}(f,t)Y_2^{(\mathrm{BF})}(f,t)^*\right\rangle_t\right|}{\left\langle\left|Y_1^{(\mathrm{BF})}(f,t)\right|^2\right\rangle_t^{\frac{1}{2}}\left\langle\left|Y_2^{(\mathrm{BF})}(f,t)\right|^2\right\rangle_t^{\frac{1}{2}}}, \quad (20)$$

where $Y_l^{(\mathrm{ICA})}(f,t)$ is the separated signal by ICA, and $Y_l^{(\mathrm{BF})}(f,t)$ is the separated signal by beamforming. If the separation performance of beamforming is superior to that of ICA, we obtain the condition, $J^{(\mathrm{ICA})}(f) > J^{(\mathrm{BF})}(f)$; otherwise $J^{(\mathrm{ICA})}(f) \le J^{(\mathrm{BF})}(f)$. Thus, an observation of the conditions yields the following algorithm:

$$\boldsymbol{W}(f) = \begin{cases} \boldsymbol{W}_{i+1}^{(\mathrm{ICA})}(f), & \left(J^{(\mathrm{ICA})}(f) \le J^{(\mathrm{BF})}(f)\right) \\ \boldsymbol{W}^{(\mathrm{BF})}(f), & \left(J^{(\mathrm{ICA})}(f) > J^{(\mathrm{BF})}(f)\right). \end{cases} \quad (21)$$

If the $(i+1)$th iteration was the final iteration, go to **step 6**; otherwise go beck to **step 2** and repeat the ICA iteration inserting the $\boldsymbol{W}(f)$ given by Eq. (21) into $\boldsymbol{W}_i(f)$ in Eq. (11) with an increment of i.

[Step 6: Ordering and scaling] Using the DOA information obtained in **step 3**, we detect and correct the source permutation and the gain inconsistency [8].

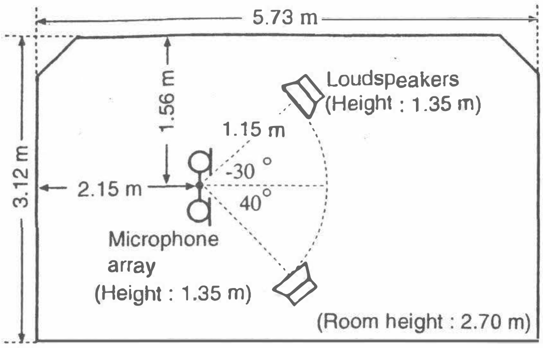

Figure 4: Layout of reverberant room used in experiments.

4. Experiments and results

4.1. Conditions for experiments

A two-element array with the interelement spacing of 4 cm is assumed. The speech signals are assumed to arrive from two directions, -30° and 40°. Two kinds of sentences, those spoken by two male and two female speakers selected from the ASJ continuous speech corpus for research [9], are used as the original speech samples. Using these sentences, we obtain 12 combinations with respect to speakers and source directions. In these experiments, we use the following signals as the source signals: the original speech convolved with the impulse responses specified by different reverberation times (RTs) of 150 msec and 300 msec. The impulse responses are recorded in a variable reverberation time room as shown in Fig. 4. The analytical conditions of these experiments are as follows: the sampling frequency is 8 kHz, the frame length is 128 msec, the frame shift is 2 msec, and the step-size parameter η is set to be 1.0×10^{-5}.

4.2. Objective evaluation of separated signals

In order to compare the performance of the proposed algorithm with that of the conventional BSS described in Sect. 2 for different iteration points in ICA, the *noise reduction rate* (NRR), defined as the output signal-to-noise ratio (SNR) in dB minus input SNR in dB, is shown in Fig. 5. These values were averages of all of the combinations with respect to speakers and source directions. As for the proposed algorithm, we also plot the NRR which is rescaled by the computational cost (see dotted lines) because the proposed algorithm has a computational complexity of about 1.9-fold compared with the conventional ICA.

In Fig. 5, it is evident that the separation performances of the proposed algorithm are superior to those of the conventional ICA-based BSS method at every iteration point, even considering the additional computational cost of the proposed algorithm. For example, compared with the conventional method, the proposed method can improve the NRR of about 4.6 dB at the 50-iteration point in the conventional ICA when the RT is 150 msec. Also, when the RT is 300 msec, the proposed method can improve the NRR of about 1.5 dB.

Figure 6 shows a result of alternation between ICA and null beamforming through iterative optimization by the proposed algorithm when the RT is 300 msec. In this figure, the symbol "▬" represents that the null beamforming is used in the iteration point and frequency bin. As shown in Fig. 6, the proposed algorithm can work automatically as follows: (1) null beam-

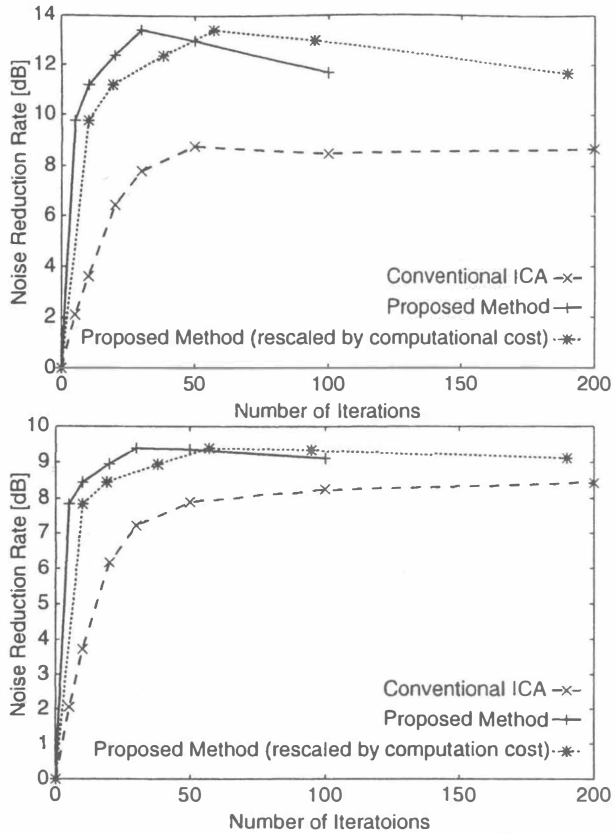

Figure 5: Noise reduction rates for different iteration in ICA. Reverberation time is 150 msec (top) and 300 msec (bottom).

forming is used for the acceleration of learning at early times in the iterations because $W^{(BF)}(f)$ is a rough approximation of the inverse of the mixing matrix $A(f)$, (2) ICA is used after the early part of the iterations because ICA can update the inverse of the mixing matrix more accurately, and (3) the inverse of the mixing matrix obtained by ICA is substituted by the matrix based on null beamforming through whole iteration points at particular frequency bins where the independence between the sources is low. From these results, although null beamforming is not suitable for signal separation under the condition that the direct sounds and their reflections exist, we can confirm that the temporal utilization of null beamforming for algorithm diversity through ICA iterations is effective for improving the separation performance and convergence.

5. Conclusion

In this paper, we described a fast- and high-convergence algorithm for BSS where null beamforming is used for temporal algorithm diversity through ICA iterations. The results of the signal separation experiments reveal that the signal separation performance of the proposed algorithm is superior to that of the conventional ICA-based BSS method, and the utilization of null beamforming in ICA is effective for improving the separation performance and convergence, even under reverberant conditions.

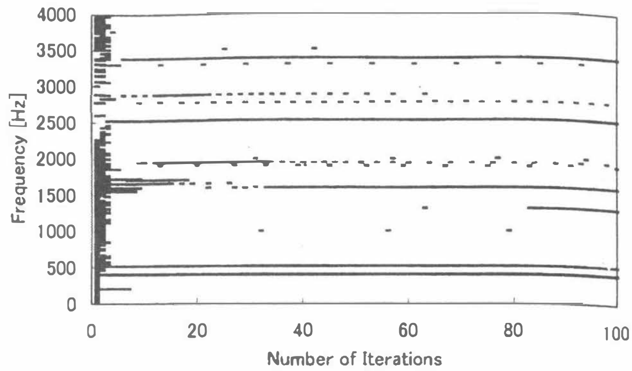

Figure 6: The result of alternation between ICA and null beamforming through iterative optimization by the proposed algorithm. The symbol "-" represents that the null beamforming is used in the iteration point and frequency bin. The RT is 300 msec.

6. Acknowledgement

The authors are grateful to Dr. Shoji Makino and Dr. Ryo Mukai of NTT. CO., LTD. for their suggestions and discussions on this work. This work was partly supported by CREST (Core Research for Evolutional Science and Technology) in Japan.

7. References

[1] P. Common, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287–314, 1994.

[2] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol.7, pp.1129–1159, 1995.

[3] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *Proceedings of 1998 International Symposium on Nonlinear Theory and Its Application (NOLTA '98)*, vol.3, pp.923–926, Sep. 1998.

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

[5] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech & Audio Process.*, vol.8, pp.320–327, 2000.

[6] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, and K. Shikano, "Blind source separation based on subband ICA and beamforming," *Proc. ICSLP2000*, vol.3, pp.94–97, Oct. 2000.

[7] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. Circuits and Systems I*, vol.43, no.11, pp.894–906, 1996.

[8] 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*, vol.5, pp.3140–3143, June 2000.

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