Online Speech Source Separation in Meeting Scene with Time-Varying Weights of Noise Covariance Matrices

Masahito Togami[1] *and Koichi Hori*[1]

[1]Department of Aeronautics and Astoronautics, School of Engineering,
The University of Tokyo, Japan

mtogami@ailab.t.u-tokyo.ac.jp, hori@computer.org

Abstract

We propose an online speech source separation technique in a meeting situation. The purpose in this paper is online extraction of each speech source from multichannel microphone input signal which is contaminated by speech sources of the other persons (noise sources). The proposed method is one of adaptive beamformers. The proposed method estimates the noise covariance matrix of the multichannel microphone input signal as a weighting average value of a noise covariance matrix of each speech source that is estimated offline. Weighting is done by using estimated activity of each speech source. By using the proposed method, even when the noise covariance matrix of microphone input signal changes rapidly due to nodding, interruption, or turn taking, the speech sources can be separated. Experimental results indicate that the proposed method can track rapid change of the noise covariance matrix and the speech sources can be separated correctly.

Index Terms: speech source separation, online algorithm, beamforming

1. Introduction

Noise reduction techniques are greatly required for recording equipments which are used in a meeting situation. In the meeting situation, speech sources frequently overlap. Multiple speech sources are required to be separated. Conventional noise cancellers with a single microphone (e.g., [1]) is not suitable for separation of speech sources, because these noise cancellers can separate only stationary noise sources and cannot separate non-stationary noise sources such as human speech sources. Recently, multichannel separation techniques have been widely applied for meeting analysis (e.g., [2, 3, 4]). Many conventional approaches are offline approaches. There is some latency between recording and separation. When speech source separation is used for realtime applications such as realtime transcription, speech sources are required to be separated online. In this paper, we focus on online speech source separation. Adaptive null beamformers (e.g., [5, 6, 7, 8, 9, 10, 11, 12, 13]) or blind source separation techniques such as independent component analysis (ICA) [14] are famous approaches as speech source separation techniques. ICA is a offline speech source separation technique, and ICA is not suitable for online speech source separation. Conventionally, adaptive beamforming techniques are frequently used for online speech source separation [6, 7, 8, 9, 11]. For speech source separation, the noise covariance matrix is required to be estimated accurately. In the conventional methods, the noise covariance matrix is gradually updated by using an exponential decay coefficient. The separation filter also change gradually as a result (or the separation filter itself is updated gradually by LMS (Least Mean Square) like algorithms). However, the assumption that the noise covariance matrix change slowly is problematic. The noise sources suddenly appear such as nodding, interruption, or turn taking. Therefore, the exponential decay coefficient is required to be small to track change of the noise covariance matrix. However, the spatial covariance matrix with a small exponential decay coefficient becomes a singular matrix frequently, and adaptive filters diverge easily. Therefore, rapid tracking capability for the speech sources without divergence is required to separate the desired source online. Updating of the noise covariance matrix with a small exponential decay coefficient is equivalent to estimation of the noise covariance matrix with few samples. From a viewpoint of learning theory, divergence can be regarded as a result of over-fitting. To avoid over-fitting, the number of the parameters to be learned online is required to be reduced. In this paper, we divide the noise covariance matrix into a time-invariant component and a time-variant component. The location of each speech source can be assumed to be time-invariant. Therefore, the covariance matrix of each speech source can be assumed to be also time-invariant. On the other hand, activity of the speech sources change rapidly due to nodding, interruption, or turn taking. Focusing on the fact that the covariance matrix of each speech source is time-invariant and only activity of each speech source is time-variant, the proposed methods estimate only activity of each speech source online. The covariance matrix of each speech source is updated offline. By the proposed method, the number of the learning parameters can be reduced, and the covariance matrix of each speech source rarely diverge. The noise covariance matrix is estimated as a weighting average of the covariance matrix of each noise source with the estimated activity of the corresponding noise source.

2. Input signal model

Input signal in each microphone is converted into time-frequency domain by short-term Fourier transform (STFT). Multichannel input signal at each time-frequency point is depicted as follows:

$$\boldsymbol{x}(f,\tau) = [\ x_1(f,\tau) \quad x_m(f,\tau) \quad x_M(f,\tau)\ ]^T, \quad (1)$$

where $x_m(f,\tau)$ is the m-th microphone input signal, T is the operator for transpose of a matrix or a vector, M is the number of the microphones, f is the frequency index, and τ is the frame index. Multichannel input signal can be modelled as follows:

$$\boldsymbol{x}(f,\tau) = \sum_{n=1}^{N} s_n(f,\tau)\boldsymbol{a}_n(f), \quad (2)$$

where N is the number of the speech sources, $s_n(f,\tau)$ is the original signal of the n-th speech source, and $\boldsymbol{a}_n(f)$ is the steering vector of the n-th speech source. Goal of speech source separation is defined as extraction of each speech signal $s_n(f,\tau)\boldsymbol{a}_n(f)$ from microphone input signal $\boldsymbol{x}(f,\tau)$ in this paper.

3. Proposed method

3.1. Overview of proposed method

Block diagram of the proposed online speech source separation is shown in Fig. 1. The proposed method is composed of two

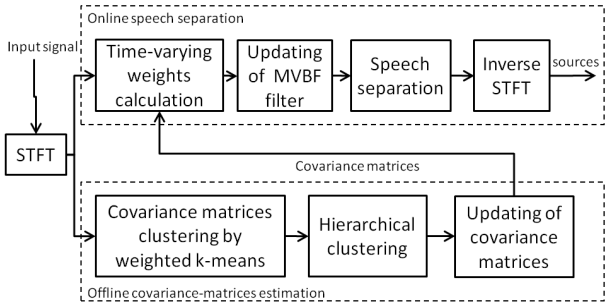

Figure 1: Block diagram of proposed method

blocks. The first block is "Online speech separation". The first block is an online block. This block is executed framewisely. In the first block, activity of each speech source is estimated at each frame. The noise covariance matrix is estimated based on the estimated activity of each speech source. Speech separation is performed by minimum variance beamformer (MVBF) [5]. The MVBF filter is updated by using the estimated noise covariance matrix framewisely. Extracted speech sources are converted into time domain by inverse STFT. The second block is an offline block. In this block, the noise covariance matrix of each noise source is estimated offline. The estimated covariance matrices are utilized in the first block.

3.2. Speech separation based on MVBF

Each speech source is separated as follows:

$$y_n(f,\tau) = \boldsymbol{w}_n(f,\tau)\boldsymbol{x}(f,\tau), \tag{3}$$

where $\boldsymbol{w}_n(f,\tau)$ is the time-varying multichannel separation filter that extracts the n-th speech source, and $y_n(f,\tau)$ is the extracted n-th source signal. $\boldsymbol{w}_n(f,\tau)$ is obtained based on MVBF [5] as follows:

$$\boldsymbol{w}_n(f,\tau) = \frac{\hat{\boldsymbol{a}}_n(f,\tau)^H \boldsymbol{R}_n(f,\tau)^{-1}}{\hat{\boldsymbol{a}}_n(f,\tau)^H \boldsymbol{R}_n(f,\tau)^{-1}\hat{\boldsymbol{a}}_n(f,\tau)}, \tag{4}$$

$$\boldsymbol{R}_n(f,\tau) = \sum_{\acute{n}\neq n} \|s_{\acute{n}}(f,\tau)\|^2 \boldsymbol{a}_{\acute{n}}(f)\boldsymbol{a}_{\acute{n}}(f)^*, \tag{5}$$

where $\hat{\boldsymbol{a}}_n(f,\tau)$ is estimation of the steering vector of the n-th speech source, $\boldsymbol{R}_n(f,\tau)$ is the noise covariance matrix that excludes the n-th source signal, and $*$ is the operator for the conjugate of a complex value. $\boldsymbol{R}_n(f,\tau)$ is drastically time-varying, because activity of each speech source change rapidly. Theoretically, from Eq. 2, the time-varying noise covariance matrix can be rewritten as follow:

$$\boldsymbol{R}_n(f,\tau) = \sum_{\acute{n}\neq n} \alpha_{\acute{n}}(f,\tau)\boldsymbol{R}_{f,\acute{n}}, \tag{6}$$

where

$$\alpha_{\acute{n}}(f,\tau) = \|s_{\acute{n}}(f,\tau)\|^2, \tag{7}$$

$$\boldsymbol{R}_{f,\acute{n}} = \boldsymbol{a}_{\acute{n}}(f)\boldsymbol{a}_{\acute{n}}(f)^*. \tag{8}$$

$\boldsymbol{R}_{f,\acute{n}}$ is the covariance matrix of the $\acute{n}$-th speech source. $\alpha_{\acute{n}}(f,\tau)$ can be regarded as activity of the $\acute{n}$-th speech source at each time-frequency point. Commonly, activity of each source synchronizes among frequencies. In this paper, $\alpha_{\acute{n}}(f,\tau)$ is approximated as a frequency-independent value $\alpha_{\acute{n}}(\tau)$. Therefore, $\alpha_{\acute{n}}(\tau)$ and $\boldsymbol{R}_{f,\acute{n}}$ are required to be obtained to achieve the time-varying multichannel separation filter $\boldsymbol{w}_n(f,\tau)$. In the proposed method, these two variables are obtained by different ways. $\boldsymbol{R}_{f,\acute{n}}$ is independent of the frame index, and $\boldsymbol{R}_{f,\acute{n}}$ is estimated offline. On the other hand, $\alpha_{\acute{n}}(\tau)$ is depending on the frame index, and $\alpha_{\acute{n}}(\tau)$ is estimated framewisely. Estimation of the steering vector of the n-th speech source is obtained as follows:

$$\hat{\boldsymbol{a}}_n(f,\tau) = \text{max\_eig}\boldsymbol{R}_{f,n}, \tag{9}$$

where max_eig returns the eigenvector whose eigenvalue is maximum among all eigenvalues.

3.3. Calculation of time-varying weights of noise covariance matrices

If the covariance matrix of the multichannel input signal can be obtained, $\alpha_n(\tau)$ can be naturally estimated by minimizing the following cost function:

$$D(\tau) = \sum_f \left\| \sum_{n=1}^{N} \alpha_n(\tau)\boldsymbol{R}_{f,n} - \boldsymbol{R}(f,\tau) \right\|_F, \tag{10}$$

$$\boldsymbol{R}(f,\tau) = \mathrm{E}[\boldsymbol{x}(f,\tau)\boldsymbol{x}(f,\tau)^H] \tag{11}$$

where E is the operator of the mathematical expectation, H is Hermite transpose of a matrix or a vector, and $\|\boldsymbol{x}\|_F$ is Frobenius norm of the matrix $\boldsymbol{x}$. However, $\boldsymbol{R}(f,\tau)$ cannot be obtained. In the proposed method, $\boldsymbol{R}(f,\tau)$ is approximated by $\hat{\boldsymbol{R}}(f,\tau)$ which is defined as follows:

$$\hat{\boldsymbol{R}}(f,\tau) = \sum_{d=1}^{d=\tau} \lambda^{\tau-d}\boldsymbol{x}(f,\tau-d)\boldsymbol{x}(f,\tau-d)^H, \tag{12}$$

where λ is an exponential decay coefficient, and λ is less than 1. By adjusting λ to a small value, $\hat{\boldsymbol{R}}(f,\tau)$ is expected to be close to the correct covariance matrix $\boldsymbol{R}(f,\tau)$. However, $\hat{\boldsymbol{R}}(f,\tau)$ with a small λ is a singular matrix, because $\hat{\boldsymbol{R}}(f,\tau)$ contains microphone input signals at only few frames. Therefore, $\hat{\boldsymbol{R}}(f,\tau)$ is not suitable as an alternative of $\boldsymbol{R}_n(f,\tau)$ in Eq. 4. In the proposed method, $\hat{\boldsymbol{R}}(f,\tau)$ is used indirectly for estimation of $\alpha_n(\tau)$. Even when $\hat{\boldsymbol{R}}(f,\tau)$ is a singular matrix, $\sum_{n=1}^{N}\alpha_n(\tau)\boldsymbol{R}_{f,n}$ is not a singular matrix under the condition that $\boldsymbol{R}_{f,n}$ is not a singular matrix. In the proposed method, $\boldsymbol{R}_{f,n}$ is estimated offline. When $\boldsymbol{R}_{f,n}$ is updated slowly, $\boldsymbol{R}_{f,n}$ is far from a singular matrix. The cost function for time-varying weights of noise covariance matrices is defined as follows:

$$D(\tau) = \sum_f \left\| \sum_{n=1}^{N} \alpha_n(\tau) \frac{\boldsymbol{R}_{f,n}}{\mathrm{trace}\{\boldsymbol{R}_{f,n}\}} - \frac{\hat{\boldsymbol{R}}(f,\tau)}{\mathrm{trace}\{\hat{\boldsymbol{R}}(f,\tau)\}} \right\|_F, \tag{13}$$

where trace$\{\boldsymbol{x}\}$ is the trace component of the matrix $\boldsymbol{x}$, the cost function is normalized among frequencies by dividing the covariance matrix by its trace component. $\alpha_n(\tau)$ which minimizes $D(\tau)$ can be obtained as follows:

$$\boldsymbol{\alpha}(\tau) = \boldsymbol{V}^{-1}\boldsymbol{C}, \tag{14}$$

$$\boldsymbol{\alpha}(\tau) = [\ \alpha_1(\tau) \quad \ldots \quad \alpha_N(\tau)\ ]^T, \tag{15}$$

where

$$[\boldsymbol{V}]_{n,\acute{n}} = \sum_{f,i,j} [\boldsymbol{R}_{f,n}]^*_{i,j} [\boldsymbol{R}_{f,\acute{n}}]_{i,j}, \tag{16}$$

$$[\boldsymbol{C}]_n = \sum_{f,i,j} [\boldsymbol{R}_{f,n}]^*_{i,j} [\hat{\boldsymbol{R}}(f,\tau)]_{i,j}, \tag{17}$$

where $[\boldsymbol{x}]_{i,j}$ is the i-th row and the j-th column element of $\boldsymbol{x}$, and $[\boldsymbol{x}]_i$ is the i-th row element. When $\alpha_n(\tau)$ is negative-valued, $\alpha_n(\tau)$ is replaced by 0. Finally, $\alpha_n(\tau)$ is normalized as follows:

$$\alpha_n(\tau) \leftarrow \frac{\alpha_n(\tau)}{\|\alpha_n(\tau)\|}. \tag{18}$$

3.4. Covariance matrix estimation by weighted k-means

A covariance matrix of each speech source is updated at every B frames interval by a semi-offline approach. Similarly to a conventional observed vector clustering approach proposed by Araki, et al., [15], the covariance matrix is obtained by clustering the input signal under the sparseness assumption that the number of the sources in each time-frequency point is assumed to be 1. Under the sparseness assumption, there is one active source in the time-frequency point, and the multichannel microphone input signal can be approximated as follows:

$$\boldsymbol{x}(f,\tau) \approx s_{\text{active}}(f,\tau)\boldsymbol{a}_{\text{active}}(f), \tag{19}$$

where active is the active source index. The covariance matrix of each speech source is obtained by minimizing the following cost function:

$$G(\boldsymbol{R}_{f,1}(b), \ldots, \boldsymbol{R}_{f,N}(b), I) = \sum_{\tau} \|\boldsymbol{x}(f,\tau)\|^P \left\| \boldsymbol{R}_{f,I(f,\tau)}(b) - \frac{\boldsymbol{x}(f,\tau)\boldsymbol{x}(f,\tau)^H}{\|\boldsymbol{x}(f,\tau)\|^2} \right\|_F \tag{20}$$

where b is the number of updates, P is the coefficient that controls the weight for the corresponding time-frequency point, and $I(f,\tau)$ is the index of the noise cluster in which $\boldsymbol{x}(f,\tau)$ is segregated. When P is close to 0, each time-frequency point is equally weighted. On the other hand, when P is large value, time-frequency points in which microphone input signal is small is small weighted. The cost function $C(\boldsymbol{R}_f, I)$ can be minimized by the weighted k-means algorithm. Even when $\boldsymbol{R}_f$ can be correctly estimated at each frequency bin separately, permutation problem is remained. The permutation problem is solved by the power-envelope correlation [16]. Only the power component of the active source is set to be 1, and the power components of the other sources are set to be 0. The effect of P is evaluated. The experimental result is shown in Fig. 2. The experimental condition is the same condition in the experiment of the latter section. In "No weight", P is 0. In "Weighting", P is set to be more than 0. The evaluation measure is noise reduction ratio (NRR) which is defined in the latter section. NRR can be improved by choosing $P > 0$ and $P < 1.5$.

3.5. Hierarchical clustering of noise covariance matrices

When the number of clusters in the weighted k-means algorithm is more than the correct number of the active sources in B frames, a same source is divided into multiple clusters. To overcome this problem, we adopt a hierarchical clustering of the covariance matrices based on a direction-of-arrival (DOA) estimate of each cluster. The DOA of each cluster is estimated by a similar way to SRP-PHAT [17] as follows:

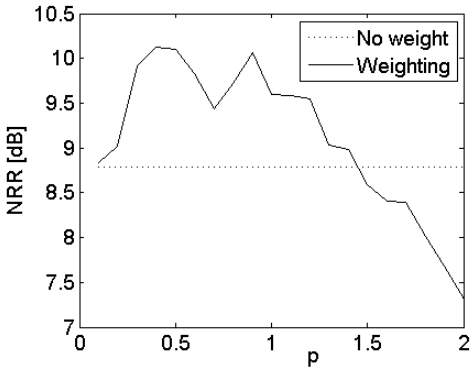

Figure 2: An evaluation result of covariance matrix estimation method by weighted k-means

$$\theta_n(b) = \underset{\theta}{\text{argmax}} \sum_f \frac{\boldsymbol{a}_\theta(f)^H \boldsymbol{R}_{f,n}(b) \boldsymbol{a}_\theta(f)}{\text{trace}\{\boldsymbol{R}_{f,n}(b)\}}, \tag{21}$$

where $\boldsymbol{a}_\theta(f)$ is the virtual steering vector at DOA=θ. In this paper, speech sources are located at a same horizontal plane. Therefore, θ is set to be the azimuth angle of a speech source. The closest pair of clusters is extracted as follows:

$$(n_1, n_2) = \underset{n_1 < n_2}{\text{argmin}} \max_u |\theta_{n_1}(b) - \theta_{n_2}(b) + 360u|, \tag{22}$$

where u is an arbitrary integer. When the difference between θ_{n_1} and θ_{n_2} is less than predefined threshold, these clusters are merged. Otherwise, the merging process stops. When two clusters are merged, the noise covariance matrix is also merged as follows:

$$\boldsymbol{R}_{f,n_1}(b) \leftarrow \boldsymbol{R}_{f,n_1}(b) + \boldsymbol{R}_{f,n_2}(b), \tag{23}$$

Final clusters are assigned to the nearest speech source as follows:

$$\acute{n} = \underset{\acute{n}}{\text{argmin}} |\theta_{\acute{n}} - \theta_n(b) + 360u|, \tag{24}$$

where

$$\theta_{\acute{n}} = -180 + \frac{360(\acute{n})}{N}. \tag{25}$$

Therefore, the horizontal plane is divided into N regions. each speech source is assigned into one region of the N regions. When the number of the clusters which is assigned to $\acute{n}$ is more than 0,

$$\boldsymbol{R}_{f,\acute{n}} \leftarrow \beta \boldsymbol{R}_{f,\acute{n}} + (1-\beta) \sum_{n \in \Omega_{\acute{n}}} \boldsymbol{R}_{f,n}(b), \tag{26}$$

where $\Omega_{\acute{n}}$ is a set of the clusters that are assigned to the $\acute{n}$-th source.

3.6. Summary of proposed method

In the proposed method, offline covariance matrices are estimated parallel to online speech separation.

Offline covariance matrices estimation

1. Microphone input signals are clustered by weighted k-means (minimization of G defined in Eq. 20).
2. Covariance matrices are merged hierarchically.

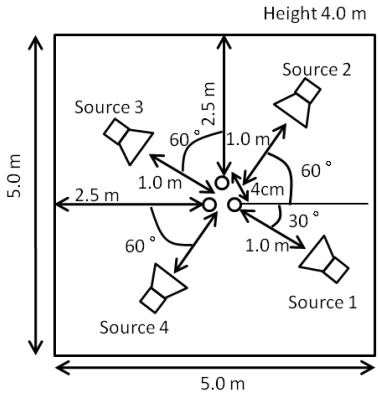

Figure 3: Simulated environment

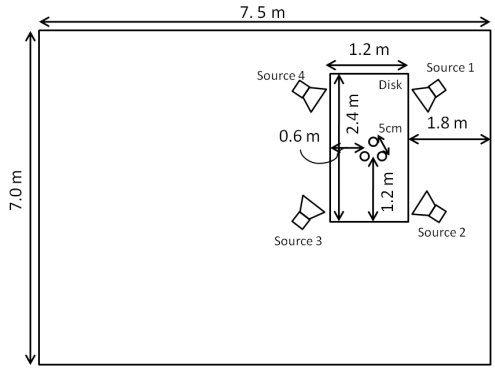

Figure 4: A real meeting room for EXP 2

3. The covariance matrix of each speech source is updated by Eq. 26.

Online speech separation

1. The time-varying weights are estimated by Eq. 14.
2. Covariance matrix is obtained by $\hat{\boldsymbol{R}}_n(f,\tau) = \sum_{\acute{n}\neq n} \alpha_{\acute{n}}(\tau)\boldsymbol{R}_{f,\acute{n}}$.
3. Steering vectors are estimated by Eq. 9.
4. The separation filter is obtained by Eq. 4.
5. Each speech source is separated by Eq. 3.

4. Experiment

The proposed method was evaluated by using simulated impulse responses (EXP 1) and by using measured impulse responses in a real meeting room (EXP 2). The simulated impulse responses were made by room impulse response generator [18]. The number of the speech sources was set to be 4 in each experimental environment. The number of the microphones is 3. Therefore, from acoustical point of view, the experimental condition is an underdetermined condition. The simulated experimental environment is shown in Fig. 3. An equilateral triangle one side of which is 4 cm was used. In EXP 1, reverberation time was changed from 0.2 [sec] to 0.6 [sec]. The experimental environment in a real meeting room is shown in Fig. 4. The impulse responses were measured by using a TSP (time-stretched pulse) signal [19]. An equilateral triangle one side of which is 5 cm was used. Sampling rate was 8 kHz. The evaluation data was made by convolution of dry sources with the impulse responses. The dry sources were picked up from RWCP-SP01 database [20]. This database contains recorded sound at several meeting situations. Speech of each speaker was recorded by a close-talking microphone, and this signal was regarded as a dry source. From RWCP-SP01, we utilized 3 meetings (meeting IDs=M01,M02,M04). Language was Japanese. In each meeting, there were 4 participants. Time-length of each meeting was about 19-22 min. Other conditions are shown in Table 1. The

Table 1: *Experimental conditions.*

type	value
frame size	512 pt
frame shift	256 pt
B	100
λ	0.4
β	0.9
P	1.0

evaluation measure is NRR (noise reduction ratio). NRR for extraction of the i-th speaker is defined as follows:

$$\text{NRR} = 10\log_{10}\frac{\sum_t \|\sum_{j\neq i} s_j(t)\|^2}{\sum_t \|s_i(t)-\hat{s}_i(t)\|^2} \tag{27}$$

where $\hat{s}_i(t)$ is a separated speech signal of the i-th speaker. When speech sources are separated effectively with low distortion, NRR takes high value. The proposed method with time-varying weights which are obtained by Eq. 14 ("Time-varying α") was compared with two methods. The first method uses constant weights in place of time-varying weights obtained by Eq. 14 ("Constant α"). In "Constant α", α_n was set to be $\frac{1}{N}$. The second method uses a time-invariant MVBF filter which was made by using the accurate noise statistics from the whole period in the meeting ("Batch"). The second method is the upper-limit of the batch algorithm. The evaluation result of EXP 1 is shown. The reverberation time was set to be 0.2 sec. The experimental result is shown in Table 2 for each meeting and each speaker. The proposed method ("Time-varying α")

Table 2: *Experimental result of EXP 1: reverberation time was set to be 0.2 sec.*

Meeting ID	Speaker ID	Constant α	Batch	Time-varying α
M01	M01	7.44	8.09	**12.18**
	F01	4.88	**13.10**	10.41
	M02	6.44	12.34	**12.86**
	M03	8.42	12.65	**13.55**
M02	M03	8.14	10.63	**15.06**
	M04	8.69	9.80	**16.62**
	F02	7.19	**14.53**	13.80
	F03	8.71	13.23	**15.51**
M04	M03	8.34	9.71	**13.04**
	M04	6.84	9.23	**11.32**
	F05	5.77	9.54	**10.03**
	F07	7.52	9.20	**11.48**

outperformed "Constant α" and "Batch" in many cases. The experimental results with various reverberation time (RT_{60}) are shown in Fig. 5. In Fig. 5(a), the experimental result for the whole meeting period is shown. In (b), the experimental result

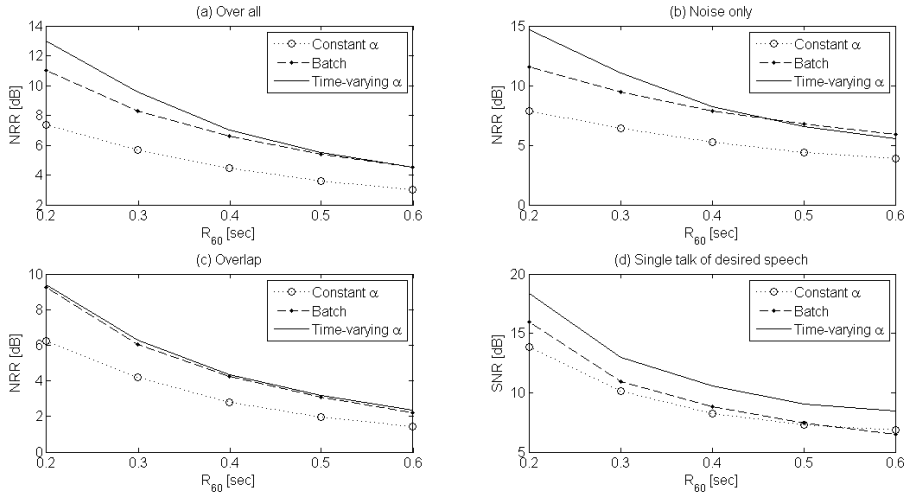

Figure 5: Experimental results with various reverberation time: Experimental environment is EXP 1

for the noise only period is shown. In (c), the experimental result for the period when the noise sources and the desired source overlap is shown. In (d), the experimental result for the period when there is only the desired source is shown. In only (d), the evaluation measure is SNR (signal to noise ratio), because SNR of the microphone input signal is infinite in this case and NRR is meaningless in this case. In all cases, the proposed method is superior to "Constant α". The proposed time-varying weights is shown to be effective. When RT_{60} is high, "Batch" and the proposed method is close to each other. However, the proposed method is superior to "Batch" at low RT_{60}. Furthermore, the proposed method is always superior to "Constant α". Next, the experimental result of EXP 2 is shown in Table. 3. The proposed method is slightly inferior to "Batch". However, the proposed method is superior to "Constant α". The proposed time-varying weights is shown to be effective. The proposed method is an online algorithm, but the proposed method is comparable to the batch algorithm.

Table 3: *Experimental result of EXP 2*

Meeting ID	Speaker ID	Constant α	Batch	Time-varying α
M01	M01	2.13	1.81	**3.64**
	F01	2.92	**8.50**	5.09
	M02	2.35	**7.78**	4.72
	M03	4.75	**7.64**	6.47
M02	M03	3.15	3.97	**6.20**
	M04	4.62	6.64	**7.69**
	F02	4.21	**8.77**	6.65
	F03	4.01	**7.62**	6.73
M04	M03	4.29	**5.70**	5.56
	M04	2.80	3.87	**4.66**
	F05	2.78	**6.69**	4.62
	F07	3.79	**5.40**	5.24

An example of time-varying weights is shown in Fig. 6. The meeting ID is M01. Even when speech period is short, it is shown that activity of each source is appropriately estimated. Finally, a sample of output signal is shown in Fig. 7. each speech source is shown to be separated effectively.

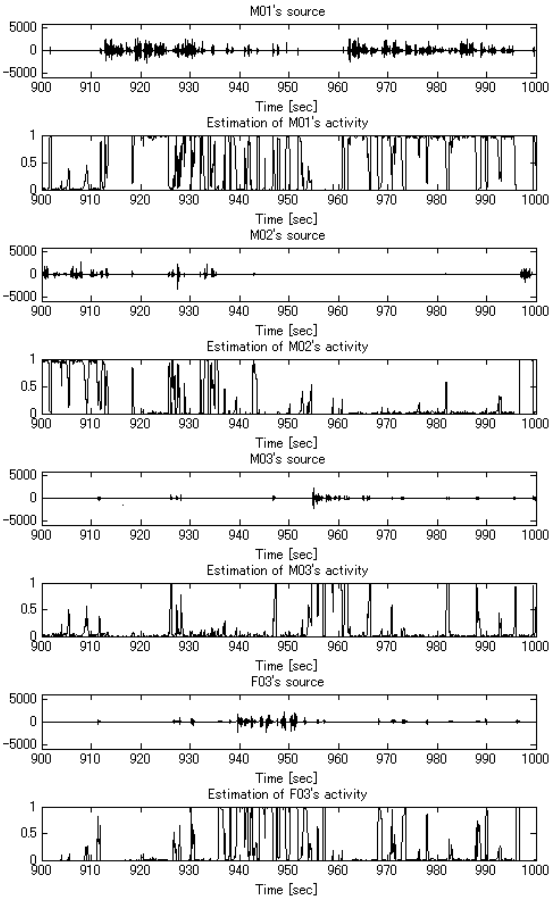

Figure 6: An example of time-varying weights

5. Discussions

The proposed method is composed of the online speech separation and the offline covariance-matrices estimation. In the offline covariance-matrices estimation, the proposed method

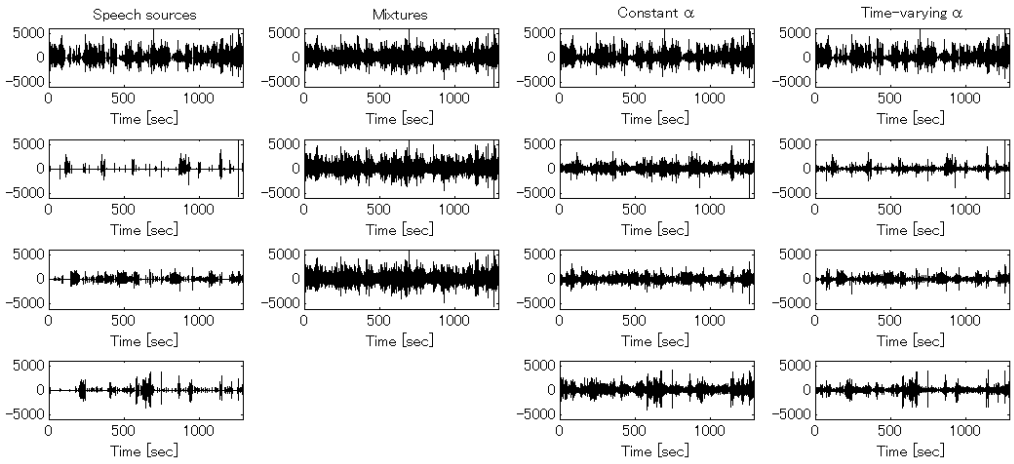

Figure 7: A sample of output signal

adopts a clustering method based on the weighted k-means under the assumption that speech sources rarely overlap at a same time-frequency point. When the number of speech sources is more than the number of microphones, the clustering method is suitable for covariance-matrices estimation. However, even when the number of participants in a meeting is more than the number of microphones, the number of speech sources is usually less than the number of microphones in a short-time period. Therefore, independent component analysis (e.g., [14]) is also suitable for offline covariance matrices estimation. Study of offline covariance matrices estimation based on independent component analysis is regarded as a future work in this paper.

6. Conclusions

In this paper, we proposed an online speech source separation technique in a meeting situation. Instead of the noise covariance matrix with an exponential decay coefficient, the proposed method estimates the noise covariance matrix as the weighting average value of the noise covariance matrix of each speech source that is estimated offline. Weighting is done by using estimated activity of each speech source. From the experimental results with the simulated impulse responses and the impulse responses with the real meeting room, the proposed method is shown to be effective.

7. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. 32, pp. 1109–1121, 1984.

[2] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "Speaker indexing and speech ehnancement in real meetings / conversations," *ICASSP2008,* pp. 93–96, 2008.

[3] F. Asano, K. Yamamoto, J. Ogata, M. Yamada, and M. Nakamura, "Detection and separation of speech events in meeting recordings using a microphone array," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 2007, article ID 27616, 2007.

[4] Augmented multi-party interaction, http://www.amiproject.org/.

[5] O.L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, 1972.

[6] L.J. Griffith and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. AP*, vol. 30, i.1, pp. 27–32, 1982.

[7] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. SP*, vol. 47, no. 10, pp. 2677–2684, 1999.

[8] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationary with applications to speech," *IEEE Trans. SAP,* vol. 5, no. 5, pp. 425–437, Sep. 1997.

[9] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environmentt with multiple interfering speech signals," *IEEE Trans. ASLP,* vol. 17, no. 6, Aug. 2009.

[10] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Proccessing,* vol. 84, no. 12, pp. 2367–2387, 2004.

[11] A. Spriet, M. Moonen, and J. Wouters, "Stochasticc gradient implementation of spatially pre-processed multi-channel Wiener filtering for noise reduction in hearing aids," *IEEE ICASSP2004,* vol. 4, pp. 57-60, 2004.

[12] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," *IEEE ICASSP2007*, vol. 1, pp. 41–44, 2007.

[13] M. Togami, Y. Obuchi, and A. Amano, "Automatic Speech Recognition of Human-Symbiotic Robot EMIEW," *Human-Robot Interaction*, pp. 395–404, I-tech Education and Publishing, 2007.

[14] A. Hyvärinen, H. Karhunen, and E. Oja, *Independent Component Analysis,* John Wiley & Sons, 2000.

[15] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," *IEEE ICASSP2006*, vol. 5, pp. 33–36, 2006.

[16] N. Murata, S. Ikeda, A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[17] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, *Microphone arrays: signal processing techniques and applications,* chapter 8. Robust localization in reverberant rooms, pp. 157–180, Eds: Michael Brandstein and Darren Ward, Springer-Verlag, 2001.

[18] Room Impulse Response Generator for MATLAB, http://home.tiscali.nl/ehabets/rir_generator.html.

[19] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *Journal of the Acoustic Society of America*, vol. 97, no. 2, pp. 1119–1123, 1995.

[20] Speech Resources Consortium, http://research.nii.ac.jp/src/eng/list/detail.html.