BLIND SPEECH SEPARATION IN A MEETING SITUATION WITH MAXIMUM SNR BEAMFORMERS

Shoko Araki[†‡] *Hiroshi Sawada*[†] *Shoji Makino*[†‡]

† NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
‡ Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo-shi, Hokkaido 060-0814, Japan
Email: shoko@cslab.kecl.ntt.co.jp

ABSTRACT

We propose a speech separation method for a meeting situation, where each speaker sometimes speaks and the number of speakers changes every moment. Many source separation methods have already been proposed, however, they consider a case where all the speakers keep speaking: this is not always true in a real meeting. In such cases, in addition to separation, speech detection and the classification of the detected speech according to speaker become important issues. For that purpose, we propose a method that employs a maximum signal-to-noise (MaxSNR) beamformer combined with a voice activity detector and online clustering. We also discuss the scaling ambiguity problem as regards the MaxSNR beamformer, and provide their solutions. We report some encouraging results for a real meeting in a room with a reverberation time of about 350 ms.

Index Terms— Speech separation, maximum SNR beamformer, scaling ambiguity, voice activity detector, online clustering.

1. INTRODUCTION

This paper considers speech separation in a meeting situation, where sources are sometimes active but silent for most of the observation period. Recently, many blind source separation methods have been proposed (e.g.,[1–3]). However, most of them assume that all sources keep speaking during the observation. However, this is not always the case in a real situation such as a conversation or a meeting. In such a scenario, the *meeting recognition* has recently been studied and it has been pointed out that the voice activity detection (VAD) for each speaker is one of the important topics (e.g., [4]). Beyond the VAD, this paper tries to recover each speech stream with a linear filtering: beamforming.

Let us formulate the task. Suppose that $N \geq 2$ speech sources $\mathrm{s}_1, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at M sensors,

$$\mathrm{x}_j(t) = \textstyle\sum_{k=1}^{N} \sum_l \mathrm{h}_{jk}(l)\,\mathrm{s}_k(t-l) + \mathrm{n}_j(t),\ j=1,\ldots,M, \quad (1)$$

where $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j, and $\mathrm{n}_j(t)$ is the observed stationary background noise at sensor j. The speech $\mathrm{s}_k(t)$ are intermittent signals. Our goal is to obtain estimates y_k of each source signal s_k separately from the sensor observations x without information about the number of sources N, the speech sources s_k or the mixing process h_{jk}.

Because we have $N \geq 2$ speeches, the other speakers act as the non-stationary noises for a target signal. Therefore, it is difficult to apply a single channel noise reduction approach (e.g.,[5]). Such an approach cannot enhance each source separately, either. When N is given, we can separate signals with, for instance, independent component analysis (e.g.,[2]) or a binary mask approach with k-means clustering (e.g.,[6]). If h_{jk} or the steering vector of the target signal is available, widely used beamformers such as the minimum variance beamformer (MVBF) [7] can enhance the target signal.

However, in this paper, we consider a scenario where none of these previous knowledge is available. In such a case, speech detection and the classification as well as separation become important issues. To realize such a system, we propose a signal extraction method with a maximum signal-to-noise (MaxSNR) beamformer [7, 8] combined with a VAD and online clustering. The MaxSNR beamformer maximizes the ratio between the signal power and interference-and-noise power, which are estimated in "target active" and "target silent (interferences-and-noise)" periods, respectively. These periods are estimated by the VAD and online clustering. With the online clustering, we can classify speech signals even if the number of speeches changes every moment.

The MaxSNR beamformer is attractive because it does not need information on the target location, such as the steering vector or h_{jk}, required in widely used beamformers e.g., the MVBF [7]. However, the MaxSNR beamformer has the scaling ambiguity problem: since the MaxSNR beamformer does not have any constraint for its gain to the target direction, the beamformer gains at different frequencies may differ. This characteristic is problematic when we apply the MaxSNR beamformer to wide-band signals such as speech mixtures [8]. The authors of [8] proposed methods for solving the scaling problem, however, their proposals still need (rough) target location information. It weakens the merit of the MaxSNR beamformer. We propose a method that uses a simple linear estimation for the scaling ambiguity problem.

We report some encouraging results of experiments conducted during a meeting in a room with reverberation time of about 350 ms.

2. PROPOSED METHOD

This paper employs a time-frequency domain approach. With a T-point short-time Fourier transform (STFT), (1) is converted into:

$$x_j(f,\tau) = \textstyle\sum_{k=1}^{N} h_{jk}(f) s_k(f,\tau) + n_j(f,\tau), \quad (2)$$

or in vector notation,

$$\mathbf{x}(f,\tau) = \textstyle\sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,\tau) + \mathbf{n}(f,\tau) \quad (3)$$

where $h_{jk}(f)$ is the frequency response from source k to sensor j, $s_k(f,\tau)$ and $n_j(f,\tau)$ are the STFTs of a source s_k and noise n_j respectively. $f \in \{0, \frac{1}{T}f_s, \ldots, \frac{T-1}{T}f_s\}$ is a frequency (f_s is the sampling frequency) and $\tau(=1,\cdots,K)$ is a time-frame index. The vectors are $\mathbf{x} = [x_1,\ldots,x_M]^T$, $\mathbf{h}_k = [h_{1k},\ldots,h_{Mk}]^T$ and $\mathbf{n} = [n_1,\ldots,n_M]^T$.

1-4244-0728-1/07/$20.00 ©2007 IEEE

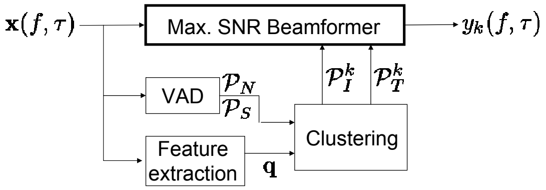

Fig. 1. Flow for proposed method.

Figure 1 shows the system flow of our proposed method. As the body of our system, we employ a maximum signal-to-noise (MaxSNR) beamformer [7]. It maximizes the ratio between the output powers for the target-active and the target-silent periods. When such a beamformer $\mathbf{w}_k(f)$ is obtained for source k, the k-th output signal can be obtained by

$$y_k(f,\tau) = \mathbf{w}_k^H(f)\mathbf{x}(f,\tau)\,. \tag{4}$$

Let $\mathcal{P} = \{1,\ldots,K\}$ be the whole period of K observations $\mathbf{x}(f,1),\ldots,\mathbf{x}(f,K)$ at each frequency, $\mathcal{P}_T^k \subset \mathcal{P}$ be the target-active period when the target source s_k is active, and $\mathcal{P}_I^k \subset \mathcal{P}$ be the target-silent period when the target s_k is NOT active but interferences and noise may active. In this paper we assume $\mathcal{P}_T^k \cup \mathcal{P}_I^k = \mathcal{P}$.

In this section, first we describe the MaxSNR beamformer, and then how to estimate $\mathcal{P}_T^k$ and $\mathcal{P}_I^k$ for the MaxSNR beamformer.

2.1. Maximum SNR beamformer

The design criterion for the beamformer $\mathbf{w}_k(f)$ is to maximize the ratio $\lambda(f)$ of the output power between the target-only period $\mathcal{P}_T^k$ and the interference-and-noise-only period $\mathcal{P}_I^k$:

$$\lambda(f) = \frac{\mathcal{E}\{|y_k(f,\tau)|^2\}_{\mathcal{P}_T^k}}{\mathcal{E}\{|y_k(f,\tau)|^2\}_{\mathcal{P}_I^k}} = \frac{\mathbf{w}_k^H(f)\mathbf{R}_\mathbf{T}^k(f)\mathbf{w}_k(f)}{\mathbf{w}_k^H(f)\mathbf{R}_\mathbf{I}^k(f)\mathbf{w}_k(f)} \tag{5}$$

where $\mathbf{R}_\mathbf{T}^k(f)$ and $\mathbf{R}_\mathbf{I}^k(f)$ are the correlation matrices of observations

$$\begin{aligned}\mathbf{R}_\mathbf{T}^k &= \mathcal{E}\{\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)\}_{\mathcal{P}_T^k} = \frac{1}{|\mathcal{P}_T^k|}\sum_{\tau\in\mathcal{P}_T^k}\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau),\\ \mathbf{R}_\mathbf{I}^k &= \mathcal{E}\{\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)\}_{\mathcal{P}_I^k} = \frac{1}{|\mathcal{P}_I^k|}\sum_{\tau\in\mathcal{P}_I^k}\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)\end{aligned}$$

where $|\mathcal{P}|$ denotes the number of elements of the set $|\mathcal{P}|$.

By differentiating $\lambda(f)$ with $\mathbf{w}_k(f)$ and setting it at 0, we have

$$\mathbf{R}_\mathbf{T}^k(f)\mathbf{w}_k(f) = \lambda(f)\mathbf{R}_\mathbf{I}^k(f)\mathbf{w}_k(f). \tag{6}$$

Obtaining the maximum $\lambda(f)$ corresponds to calculating the largest eigenvalue of the generalized eigenvalue problem (6), and the corresponding eigenvector $\mathbf{e}(f)$ gives the solution for the MaxSNR beamformer

$$\mathbf{w}_k(f) = \mathbf{e}(f). \tag{7}$$

(6) is simplified to an eigenvalue problem by multiplying both sides by $[\mathbf{R}_\mathbf{I}^k(f)]^{-1}$.

2.2. Scaling determination

The MaxSNR beamformer does not have any constraint for its gain, and so the beamformer gain provided by (7) has a scaling ambiguity. This characteristic should be compensated for if the MaxSNR beamformer is applied to wide-band signals such as speech.

Inspired by the deflation based blind source separation algorithm [9], we propose to compensate $\mathbf{w}_k(f)$ so that the output $y_k(f,\tau)$ becomes as close as observations:

$$\mathbf{x}(f,\tau) \approx \mathbf{a}(f)y_k(f,\tau) = \mathbf{a}(f)\mathbf{w}_k^H(f)\mathbf{x}(f,\tau).$$

That is, we calculate $\mathbf{a}(f)$, which minimizes the following cost function:

$$\mathcal{G}(\mathbf{a}(f)) = \mathcal{E}\{||\mathbf{x}(f,\tau) - \mathbf{a}(f)y_k(f,\tau)||^2\}. \tag{8}$$

This is a linear least-mean-squares estimation problem [10]. Therefore, an optimal $\mathbf{a}(f)$ can be obtained by setting the differentiation $\frac{\partial\mathcal{G}(\mathbf{a}(f))}{\partial\mathbf{a}(f)}$ at zero:

$$\mathbf{a}(f) = \frac{\mathcal{E}\{y_k^*(f,\tau)\mathbf{x}(f,\tau)\}}{\mathcal{E}\{|y_k(f,\tau)|^2\}} = \frac{\mathbf{R}_\mathbf{x}(f)\mathbf{w}_k(f)}{\mathbf{w}_k^H(f)\mathbf{R}_\mathbf{x}(f)\mathbf{w}_k(f)}, \tag{9}$$

where $\mathbf{R}_\mathbf{x}(f) = \mathcal{E}\{\mathbf{x}(f,\tau)\mathbf{x}^H(f,\tau)\}$ is the observation correlation matrix. The scale compensated beamformer is given by a selecting J-th component a_J,

$$\mathbf{w}_k(f) \leftarrow \mathrm{a}_J\mathbf{w}_k(f). \tag{10}$$

2.3. Target and noise period estimation

For the MaxSNR beamformer, the $\mathcal{P}_T^k$ and $\mathcal{P}_I^k$ estimations are very important. These estimations correspond to the target and interference classification. However, the well-known k-means clustering is difficult to use in a meeting situation where the number of speakers may change. Moreover, the noise-only periods when no speaker is active degrade the speaker classification performance.

In order to handle such issues, this paper proposes a method that employs a voice activity detector (VAD) and online clustering of the time-difference of arrival (TDOA) information. First, time frames τ, where no speaker is active (and noise may exist), are removed with a VAD. Then the remaining time frames τ, where speakers are active, are classified into the target period and the interference period with online clustering.

Step1: Voice activity detection (VAD)

First, we detect the noise-only period $\mathcal{P}_N \subset \mathcal{P}_I^k$ by using the VAD proposed by Sohn et al [11]. The noise-only period $\mathcal{P}_N$ and the speech-exist period $\mathcal{P}_S$ are determined by the ML based decision rule:

$$\begin{array}{llll}\text{if} & \Lambda(\tau) > \eta & \text{then} & \tau\in\mathcal{P}_S\\ \text{else} & & & \tau\in\mathcal{P}_N\end{array} \tag{11}$$

where

$$\Lambda(\tau) = \sum_{f=0}^{(T-1)f_s/T}\{\gamma(f,\tau) - \log\gamma(f,\tau) - 1\},$$

$\gamma(f,\tau) = ||\mathbf{x}(f,\tau)||^2/\sigma(f,\tau)$ is a posteriori SNR, $\sigma(f,\tau)$ is an estimated noise variance, and η is a threshold [11].

Step2: Feature extraction

As the speech-exist period $\mathcal{P}_S = \mathcal{P} - \mathcal{P}_N$ includes the target source and other interferences, it should be classified into target period $\mathcal{P}_T^k$ and interference period. These periods are determined in steps 2 and 3.

In order to classify the target (k-th signal) and interferences, we utilize the time differences of arrival (TDOA) at sensors j and j'. TDOA $q'_{jj'}(\tau)$ can be estimated by using the generalized cross correlation method with the phase transform (GCC-PHAT) [12]

$$q'_{jj'}(\tau) = \mathrm{argmax}_{q'}\sum_f \frac{x_j(f,\tau)x_{j'}^*(f,\tau)}{|x_j(f,\tau)x_{j'}^*(f,\tau)|}e^{\jmath 2\pi f q'}. \tag{12}$$

We can use a TDOA (column) vector $\mathbf{q}'(\tau)$, which consists of the $q'_{jj'}(\tau)$ of all the sensor pairs for the clustering. Instead, we utilized the direction of arrival (DOA) vector $\mathbf{q}(\tau)$ [13]. When the source azimuth is $\theta(\tau)$ and the elevation is $\phi(\tau)$, the DOA vector can be written as $\mathbf{q}(\tau) = [\cos\theta(\tau)\cos\phi(\tau),\ \sin\theta(\tau)\cos\phi(\tau), \sin\phi(\tau)]^T$. The DOA vector $\mathbf{q}(\tau)$ is calculated by the TDOA information $\mathbf{q}'(\tau)$

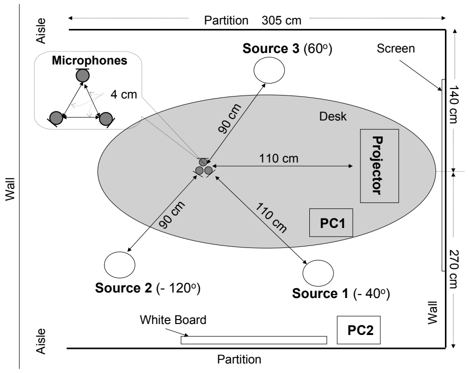

Fig. 2. Room setup. The reverberation time was around 350 ms..

and the given sensor coordinate information $\mathbf{D}$ [13]:

$$\mathbf{q}(\tau) = c\mathbf{D}^{-}\mathbf{q}'(\tau). \tag{13}$$

where c is the propagation velocity of the signals and $^{-}$ denotes the Moore-Penrose pseudo-inverse.

Step3: Clustering

To divide $\mathcal{P}_S$ into the target period $\mathcal{P}_T^k$ and interference period, we then perform clustering for the extracted features $\mathbf{q}(\tau)$ of all time-frame $\tau \in \mathcal{P}_S$. As we do not know the number of sources N to be separated, we employ the online clustering algorithm (leader-follower clustering) [14].

Let us select the clusters which have enough elements, and name one of them C_k. Now we set the target period $\mathcal{P}_T^k$ as

$$\tau \in \mathcal{P}_T^k \quad \text{if} \quad \mathbf{q}(\tau) \in C_k. \tag{14}$$

Finally the interference-and-noise-only period $\mathcal{P}_I^k$ is determined as

$$\mathcal{P}_I^k = \mathcal{P} - \mathcal{P}_T^k. \tag{15}$$

Method summary

1. Estimate noise-only period $\mathcal{P}_N$ and speech-exist period $\mathcal{P}_S$ with VAD.
2. Calculate the feature $\boldsymbol{q}(\tau)$ for each frame τ.
3. For each k ($k = 1, \cdots, N$),
   1) Determine the target period $\mathcal{P}_T^k$ from $\mathcal{P}_S$ with (14).
   2) Decide the interference-and-noise-only period $\mathcal{P}_I^k$ with (15).
   3) Calculate a MaxSNR beamformer $\mathbf{w}_k$ (7) with $\mathcal{P}_I^k$ and $\mathcal{P}_T^k$.
   4) Compensate the beamformer gain with (10).
   5) Calculate the output with (4).

There may be many options with respect to the decision for $\mathcal{P}_T^k$ and $\mathcal{P}_I^k$. For example, in this paper, we defined $\mathcal{P}_T^k$ and $\mathcal{P}_I^k$ for every time frame τ. We can also determine these periods for every time-frequency (f, τ) by using the VAD (11) and DOA vector (13) at each time-frequency point. Such an option may work well when speech signals often overlap.

3. EXPERIMENTS

3.1. Setup

Experiments were performed in the room shown in Fig. 2 whose reverberation time was around 350 ms. DOAs θ for sources 1, 2 and 3 were -40°, -120° and 60°, respectively. **Simulated meeting observations** were made by following (1) with the impulse responses $\mathrm{h}_{jk}(l)$ and the noise $\mathrm{n}_j(t)$ measured in the room, and English speech sources $\mathrm{s}_k(t)$ sampled at 16 kHz. The active time for each source in the simulation is shown in Fig. 3. We utilized the recorded noise from the projector, the personal computers PC1 and PC2 (see Fig. 2). We also tried **a recorded meeting** in the room shown in Fig. 2. During the recording, source 1 was standing, and sources 2 and 3 were sitting. As it was a real meeting, the sources moved. The frame size T for STFT was 2048 (128 ms), and the frame shift was 256 (16 ms).

3.2. Evaluation measures

The signal-to-interference plus noise-ratio (SINR) was calculated by

$$\mathsf{SINR}_i = 10\log_{10}\frac{\sum_t |\mathrm{y}_{ii}(t)|^2}{\sum_t |\sum_{k\neq i}\mathrm{y}_{ik}(t) + \mathrm{y}_n(t)|^2} \quad \text{[dB]}, \tag{16}$$

where $\mathrm{y}_{ik}(t)$ is the s_k component that appears at output $\mathrm{y}_i(t)$: $\mathrm{y}_i(t) = \sum_{k=1}^{N}\mathrm{y}_{ik}(t)$, and $\mathrm{y}_n(t)$ the noise output: $\mathrm{y}_n = \text{IFFT}\left[\mathbf{w}_i^H(f)\mathbf{n}(f,t)\right]$.

The sound quality was evaluated by the signal to distortion ratio (SDR):

$$\mathsf{SDR}_i = 10\log_{10}\frac{\sum_t |\mathrm{x}_{Ji}(t)|^2}{\sum_t |\mathrm{x}_{Ji}(t) - \beta\mathrm{y}_{ii}(t-\Delta)|^2} \quad \text{[dB]}, \tag{17}$$

where $\mathrm{x}_{Jk}(t) = \sum_l \mathrm{h}_{Jk}(l)\,\mathrm{s}_k(t-l)$, β and Δ are parameters used to compensate for the amplitude and phase difference between x_{Ji} and y_{ii}. In the simulation, we tried six speaker combinations and averaged all the results.

3.3. Results

In the simulation, there were overlaps between speakers as shown in Fig. 3, and Fig. 4. Figure 5 illustrates the estimated DOA $\theta(\tau)$ for the speech-exist period $\mathcal{P}_S$ (black dots) and noise-only period $\mathcal{P}_N$ (gray dots). The estimation succeeded for most of the periods, however, the VAD sometimes failed the noise-only period $\mathcal{P}_N$ estimation. Moreover, since we employed GCC, we can estimate only one DOA at each frame τ even when speech signals overlapped. However, we can separate each signal successfully as shown in Fig. 5 and Table 1. We can say that, for designing the MaxSNR beamformer, the target period $\mathcal{P}_T^k$ and the interference-and-noise-only period $\mathcal{P}_I^k$ estimation does not have to be so accurate.

Figure 7 shows the effect of the scaling method (10). (a) is the signal spectrum for a source s_1 of one second and (b) is that for a separated signal y_1 without scaling. Without scaling, the spectrum (b) is severely distorted. This is why the SDR was poor without scaling (see SDR' in Table 1). With the scaling method, we can restore the spectrum to that of speech (Fig. 7(c)).

We also tried to separate the signals for a recorded meeting. The top of Fig. 9 shows one of the observations x_1. There was heavy noise and speeches sometimes overlapped. Figure 8 illustrates the estimated DOA $\theta(\tau)$ for the speech-exist period $\mathcal{P}_S$ (black dots) and noise-only period $\mathcal{P}_N$ (gray dots). Because this is a real recording, the source 3 ($\theta \approx -40^\circ$) was moving. As we cannot calculate SINR and SDR for a real recording, we checked the output waveforms (in Fig. 9). We can see that each speech was extracted successfully.

4. CONCLUSION

We proposed a method for extracting speech from a meeting recording with the MaxSNR beamformer. We also proposed a method for the target active/silence period estimation with a VAD and TDOA clustering, and the scaling compensation method for the MaxSNR beamformer. We confirmed that our proposed method works well for a meeting situation. In this paper, the method was applied to the whole 30-second long observations, that is, it worked in a batch mode. However, the results with our method, which includes the online clustering, encourage us to implement this approach in real-time configuration.

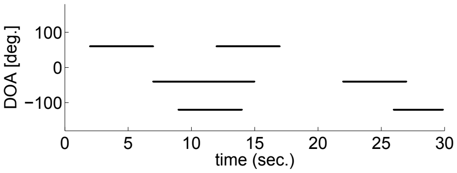

Fig. 3. True speech active periods for the simulated meeting.

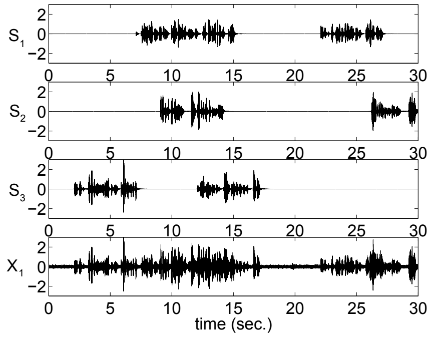

Fig. 4. Original signals and a mixture x_1 in the simulated meeting.

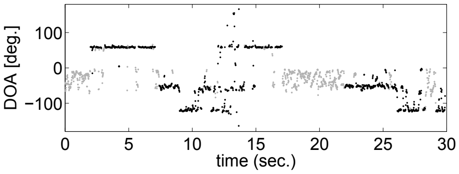

Fig. 5. Estimated DOA for speech-exist period (black) and noise-only period (gray) for the simulated meeting.

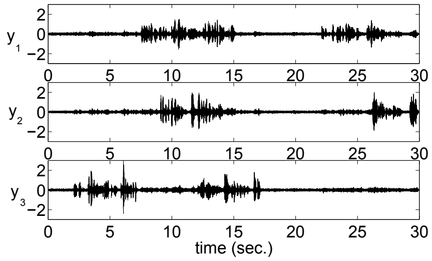

Fig. 6. Separated signals in the simulated meeting.

Table 1. Simulation results. SDR': without scaling solver (10).

SINR[dB]	SDR[dB]	SDR'[dB]
9.1	12.9	3.8

5. REFERENCES

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*, John Wiley & Sons, 2000.

[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., pp. 299–327. Springer, Mar. 2005.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[4] D. Ellis and J. Liu, "Speaker turn segmentation based on between-channel differences," in *Proc. NIST Meeting Recognition Workshop*, 2004, pp. 112–117.

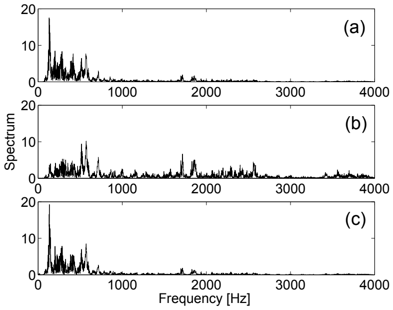

Fig. 7. Signal spectra for (a) source s_1, (b) separated signal y_1 without scaling, (c) separated signal y_1 with scaling.

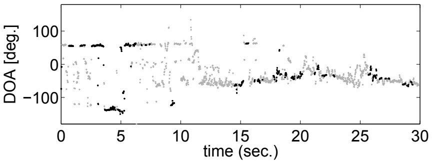

Fig. 8. Estimated DOA for speech-exist period (black) and noise-only period (gray) for the recorded meeting.

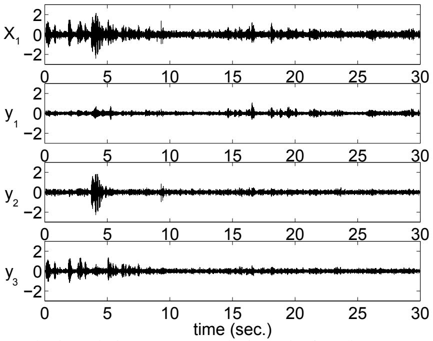

Fig. 9. A recorded and the separated signals for the recorded meeting.

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

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. IWAENC 2005*, Sept. 2005, pp. 117–120.

[7] H. L. Van Trees, Ed., *Optimum Array Processing*, John Wiley & Sons, 2002.

[8] E. Warsitz and R. Haeb-Unbach, "Controlling speech distortion in adaptive frequency-domain principal eigenvector beamforming," in *Proc. IWAENC 2006*, 2006.

[9] A. Cichocki, R. Thawonmas, and S. Amari, "Sequential blind signal extraction in order specified by stochastic properties," *Electronics Letters*, vol. 33, no. 1, pp. 64–65, 1997.

[10] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall, 2000.

[11] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[12] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

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

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley Interscience, 2nd edition, 2000.