REAL-TIME INTEGRATION OF STATISTICAL MODEL-BASED SPEECH ENHANCEMENT WITH UNSUPERVISED NOISE PSD ESTIMATION USING MICROPHONE ARRAY

T. Kawase[⋆] *K. Niwa*[⋆] *M. Fujimoto*[†] *N. Kamado*[⋆] *K. Kobayashi*[⋆] *S. Araki*[†] *T. Nakatani*[†]

[⋆] NTT Media Intelligence Laboratories, Japan
[†] NTT Communication Science Laboratories, Japan

ABSTRACT

We propose a technique of multi-channel speech enhancement based on integration of beamforming and statistical model-based speech enhancement to clearly extract the target speech, even in very noisy environments. Conventional microphone array-based techniques estimate speech and noise power spectral densities (PSDs) from the spatial cues of the sound sources; however, their estimation errors dramatically increase when there are many noise sources. We integrated clean speech models trained in advance and the noise PSDs estimated in beamspace to compose observation models and designed a precise Wiener filter. Experiments under adverse noise conditions showed that the proposed technique significantly improved the signal-to-noise ratios (SNRs) compared with the conventional microphone array processing technique.

Index Terms— Microphone array, beamforming, power spectral density estimation, statistical model, Wiener filter.

1. INTRODUCTION

Hands-free communication can be used anywhere on various audio devices, such as smartphones, wireless headsets, and car microphones. We can use multiple microphones mounted on such audio devices. However, the quality of communication deteriorates when they are in noisy environments such as crowded places, factories, and cars running at high speed. It is important to develop techniques to clearly extract speech even in such environments by reducing noise without any signal distortion. For this purpose, we developed a microphone array-based technique. It is important for processing applied to communication in actual noisy environments to robustly adapt to various environments in real time.

Various speech enhancement techniques with microphone arrays have been reported [1]. Applying a Wiener filter [2–6] to the beamforming output [7, 8] is an effective way to reduce noise with several microphones. We have already proposed a *PSD-estimation-in-beamspace* method for estimating the power spectral densities (PSDs) of the target and other sounds on the basis of the phase and amplitude differences between microphones, referred to as spatial cues [9–11]. The target/noise PSD estimation has been demonstrated to be robust in many circumstances, but the target PSD estimation errors sometimes drastically increase when the sound sources are not sparse, especially in very noisy environments. These errors can cause musical noise or distort signals.

Aside from this, statistical model-based speech enhancement techniques have been extensively studied, mainly for the purpose of developing robust automatic speech recognition in adverse noise environments. These techniques incorporate statistical models of speech (speech models), which are trained using certain speech corpora, as prior knowledge about a speech. They can accurately preserve the characteristics of the speech spectra even after noise reduction [12–16]. However, they are designed for single microphone signals, and their estimation accuracy rapidly deteriorates as the signal-to-noise ratio (SNR) of the captured signal decreases. To overcome this limitation, a few techniques have been developed to attempt to integrate statistical model-based speech enhancement with microphone-array-based beamforming [17–19]. However, one of them merely applies a cascade connected beamforming and a statistical model-based approach [17], while others require iterative optimization based on batch processing to adapt the model parameters to the environments [18,19]; thus, these techniques cannot adapt to every environment in real time.

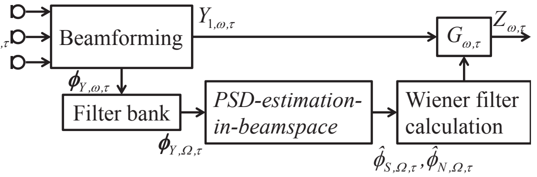

Fig. 1. Noise reduction based on *PSD-estimation-in-beamspace* and Wiener filtering

Inspired by these studies, we propose a technique for integrating beamforming [9–11] with model-based speech enhancement [16] in a way that can adapt to every environment in real time. This technique uses fixed beamforming filters (BFs) and estimates the PSD of the residual noise by using the *PSD-estimation-in-beamspace* method in real time. Pre-trained clean speech models are used to preserve the general characteristics of the speech spectra. This prevents serious errors in speech PSD estimation due to many noise sources and improves the Wiener filter. There can be synergy between the beamforming and the model-based technique because using BFs provides more accurate noise PSD estimation than conventional single channel model-based techniques.

The rest of the paper is organized as follows. Sec. 2 explains the *PSD-estimation-in-beamspace* method and Sec. 3 presents how to integrate it with the model-based technique. After verifying the effectiveness of the proposed technique experimentally in Sec. 4, the paper concludes in Sec. 5.

2. PSD-ESTIMATION-IN-BEAMSPACE FOR NOISE REDUCTION

This section briefly reviews our previous work, i.e., noise reduction based on *PSD-estimation-in-beamspace* and Wiener filtering. Fig. 1 is the flowchart of the noise-reduction process.

Let us assume that the microphone array is composed of M microphones and the arrival direction of the target speech θ_1 is known. The observed signal in frequency bin ω and time-frame τ is denoted

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

as $\boldsymbol{X}_{\omega,\tau} \in \mathbb{C}^M$. The $\boldsymbol{X}_{\omega,\tau}$, target speech $S_{\omega,\tau}$, coherent interference noise $\boldsymbol{N}_{\text{interf},\omega,\tau} \in \mathbb{C}^Q$, and incoherent background noise $\boldsymbol{N}_{\text{backg},\omega,\tau} \in \mathbb{C}^M$ obey Eq. (1), where $\boldsymbol{A}_\omega : \mathbb{C}^{Q+1} \to \mathbb{C}^M$ denotes the transfer function between the sound sources and microphones.

$$\boldsymbol{X}_{\omega,\tau} = \boldsymbol{A}_\omega \begin{bmatrix} S_{\omega,\tau} \\ \boldsymbol{N}_{\text{interf},\omega,\tau} \end{bmatrix} + \boldsymbol{N}_{\text{backg},\omega,\tau} \tag{1}$$

Fig. 1 gives an overview of noise reduction based on *PSD-estimation-in-beamspace*. In this framework, multiple BFs are used to analyze sounds arriving from not only the target arrival direction but also different L (≥ 2) directions θ_l. The BF output $\boldsymbol{Y}_{\omega,\tau} \in \mathbb{C}^L$ is derived as Eq. (2), where $\boldsymbol{W}_\omega : \mathbb{C}^L \to \mathbb{C}^M$ denotes the BF coefficients designed by minimum variance distortionless response (MVDR) [7,20].

$$\boldsymbol{Y}_{\omega,\tau} = \boldsymbol{W}_\omega^{\mathrm{H}} \boldsymbol{X}_{\omega,\tau} \tag{2}$$

The superscript $^{\mathrm{H}}$ denotes the Hermitian conjugate.

In the PSD estimation, the frequency bin of $\boldsymbol{Y}_{\omega,\tau}$ is compressed through a filter bank. We denote the filter bank channel and compressed PSD (CPSD) of the BF output as Ω and $\boldsymbol{\phi}_{Y,\Omega,\tau} \in \mathbb{C}^L$, respectively. The frequency band corresponding to each Ω is set uniformly in the equivalent rectangular bandwidth (ERB) scales [21].

We describe the beamspace in terms of the angle range Θ_l centered on the direction θ_l. Given the sparseness and non-correlativity of the source signals, $\boldsymbol{\phi}_{Y,\Omega,\tau}$ can be modeled as Eq. (3) [9], where $\boldsymbol{\phi}_{\Theta,\Omega,\tau} \in \mathbb{C}^L$ and $\boldsymbol{D}_\Omega : \mathbb{C}^L \to \mathbb{C}^L$ denote the CPSD of the sound sources inside each Θ_l and the gains of the BFs to the beamspace.

$$\boldsymbol{\phi}_{Y,\Omega,\tau} = \boldsymbol{D}_\Omega \boldsymbol{\phi}_{\Theta,\Omega,\tau} \tag{3}$$

We assume Θ_1 is the target beamspace.

The relationship between the instantaneous powers calculated frame-by-frame also obeys Eq. (3) if the sparseness of the source signals is high. The $\boldsymbol{D}_\Omega$ can be calculated in advance by multiplying the BFs and array manifold vectors [7]. Therefore, the instantaneous powers of the sources in each beamspace $\hat{\boldsymbol{\phi}}_{\Theta,\Omega,\tau}$ can be estimated from the BF outputs in real time by sequentially solving the inverse problem of Eq. (3), as in Eq. (4).

$$\hat{\boldsymbol{\phi}}_{\Theta,\Omega,\tau} = \boldsymbol{D}_\Omega^{-1} \boldsymbol{\phi}_{Y,\Omega,\tau} \tag{4}$$

The speech CPSD $\hat{\phi}_{S,\Omega,\tau}$ and noise CPSD $\hat{\phi}_{N,\Omega,\tau}$ are derived from $\hat{\boldsymbol{\phi}}_{\Theta,\Omega,\tau}$ [10]. Since observations in actual fields contain spatially incoherent background noise as well as interference noise (Eq. (1)), $\hat{\phi}_{N,\Omega,\tau}$ can be expressed as Eq. (5), where $\phi_{\text{interf},\Omega,\tau}$ and $\phi_{\text{Tbackg},\Omega,\tau}$ denote the CPSDs of the interference sources and the background noise inside the target beamspace, respectively.

$$\hat{\phi}_{N,\Omega,\tau} = \hat{\phi}_{\text{interf},\Omega,\tau} + \hat{\phi}_{\text{Tbackg},\Omega,\tau} \tag{5}$$

The CPSDs $\hat{\phi}_{\text{interf},\Omega,\tau}$ are calculated as Eq. (6), where $\hat{\phi}_{\text{Nbackg},\Omega,\tau}$ denotes the PSD of the background noise outside the target beamspace.

$$\hat{\phi}_{\text{interf},\Omega,\tau} = \sum_{l=2}^{L} \hat{\phi}_{\Theta_l,\Omega,\tau} - \hat{\phi}_{\text{Nbackg},\Omega,\tau} \tag{6}$$

Assumed to be highly stationary, the background noise CPSDs $\hat{\phi}_{\text{Tbackg},\Omega,\tau}$ and $\hat{\phi}_{\text{Nbackg},\Omega,\tau}$ are estimated as the minimum value in a time interval [10]. The $\hat{\phi}_{S,\Omega,\tau}$ is calculated as Eq. (7).

$$\hat{\phi}_{S,\Omega,\tau} = \hat{\phi}_{\Theta_1,\Omega,\tau} - \hat{\phi}_{\text{Tbackg},\Omega,\tau} \tag{7}$$

The output $Z_{\omega,\tau}$ is obtained by applying a Wiener filter $G_{\omega,\tau}$ to the target directional signal $Y_{1,\omega,\tau}$ output by the BF, as in Eq. (8).

$$Z_{\omega,\tau} = G_{\omega,\tau} Y_{1,\omega,\tau} \tag{8}$$

The $G_{\omega,\tau}$ is obtained with the estimated speech and noise CPSDs, as in Eq. (9), where the frequency scale is mapped from Ω to ω.

$$G_{\omega,\tau} = \frac{\hat{\phi}_{S,\Omega,\tau}}{\hat{\phi}_{S,\Omega,\tau} + \hat{\phi}_{N,\Omega,\tau}} \tag{9}$$

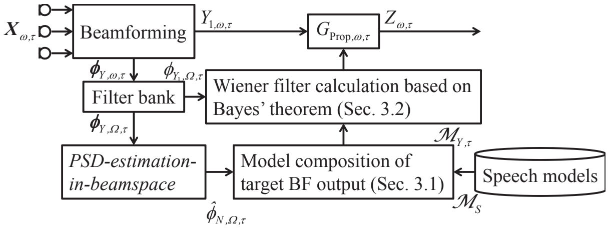
$\boldsymbol{X}_{\omega,\tau}$
Beamforming
$Y_{1,\omega,\tau}$
$G_{\text{Prop},\omega,\tau}$
$Z_{\omega,\tau}$
$\boldsymbol{\phi}_{Y,\omega,\tau}$
$\phi_{Y_1,\Omega,\tau}$
Filter bank
Wiener filter calculation based on Bayes' theorem (Sec. 3.2)
$\boldsymbol{\phi}_{Y,\Omega,\tau}$
$\mathcal{M}_{Y,\tau}$
PSD-estimation-in-beamspace
Model composition of target BF output (Sec. 3.1)
Speech models
$\hat{\phi}_{N,\Omega,\tau}$
$\mathcal{M}_S$

Fig. 2. Integration technique of *PSD-estimation-in-beamspace* and statistical model-based speech enhancement

3. PROPOSED TECHNIQUE

The aforementioned *PSD-estimation-in-beamspace* method is used to design a Wiener filter deterministically using only spatial cues and the given sparseness of the sound sources. However, the estimation errors of the target and noise CPSDs increase and result in musical noise or signal distortion when noise levels are high and the source sparseness is low. To overcome this problem, we investigated ways of integrating *PSD-estimation-in-beamspace* and statistical model-based speech enhancement by using statistical models as prior knowledge about the target speech. With these statistical models, we can accurately preserve the characteristics of speech spectra even in very noisy environments with many noise sources.

The overview of the proposed technique is shown in Fig. 2. The Wiener filter is improved despite there being a few errors in $\hat{\phi}_{N,\Omega,\tau}$ in noisy environments if precise statistical models of the target BF output, referred to as observation models, can be obtained by sequentially integrating statistical speech models and the estimated noise CPSD.

3.1. Model composition of the target BF output

The statistical clean speech model is an ergodic hidden Markov model (HMM) with two internal states, i.e., states of silence ($j = 1$) and speech ($j = 2$), where j denotes the state index. Each state is modeled in advance by a Gaussian mixture model (GMM) with K Gaussian components in a I-dimensional logarithmic CPSD (LCPSD) domain as Eq. (10).

$$\boldsymbol{\phi}_\tau^{\log} \triangleq \{\log(\phi_{\Omega,\tau})\}_{\Omega=0}^{I-1} \tag{10}$$

Each state model has model parameters $\mathcal{M}_S$, as in Eq. (11), where $\lambda_{j,k}$, $\boldsymbol{\mu}_{j,k}$, $\boldsymbol{\Sigma}_{j,k}$, and $\sigma_{\Omega,j,k}^2$ denote the mixture weight, mean vector, diagonal variance matrix, and variance, respectively.

$$\begin{aligned} \mathcal{M}_S &= \{\lambda_{S,j,k},\ \boldsymbol{\mu}_{S,j,k},\ \boldsymbol{\Sigma}_{S,j,k}\} \\ &\triangleq \left\{\lambda_{S,j,k},\ \{\mu_{S,\Omega,j,k}\}_{\Omega=0}^{I-1},\ \operatorname{diag}\{\sigma_{S,\Omega,j,k}\}_{\Omega=0}^{I-1}\right\} \end{aligned} \tag{11}$$

The k denotes the index of the Gaussian component.

The time varying parameters of the observation models in the LCPSD domain $\mathcal{M}_{Y,\tau}$ are expressed by Eq. (12).

$$\begin{aligned} \mathcal{M}_{Y,\tau} &= \{\lambda_{Y,j,k},\ \boldsymbol{\mu}_{Y,\tau,j,k},\ \boldsymbol{\Sigma}_{Y,j,k}\} \\ &\triangleq \left\{\lambda_{Y,j,k},\ \{\mu_{Y,\Omega,\tau,j,k}\}_{\Omega=0}^{I-1},\ \operatorname{diag}\{\sigma_{Y,\Omega,j,k}\}_{\Omega=0}^{I-1}\right\} \end{aligned} \tag{12}$$

With the model parameters of clean speech $\mathcal{M}_S$ and the estimated noise CPSD $\hat{\phi}_{N,\omega,\tau}$, the time varying parameters of the observation models are sequentially composed by using zeroth order vector Taylor series composition [12], as in Eqs. (13)–(15).

$$\lambda_{Y,j,k} = \lambda_{S,j,k} \tag{13}$$

$$\mu_{Y,\Omega,\tau,j,k} = \log\left\{\exp\left(\mu_{S,\Omega,j,k}\right) + \hat{\phi}_{N,\Omega,\tau}\right\} \tag{14}$$

$$\boldsymbol{\Sigma}_{Y,j,k} = \boldsymbol{\Sigma}_{S,j,k} \tag{15}$$

3.2. Wiener filter calculation based on Bayes' theorem

There have been a number of studies on designing Wiener filters on the basis of statistical models [14–16]. The proposed technique applies one of these techniques, which calculates the Wiener filter in the CPSD domain by simply following Bayes' theorem [14, 16].

After model composition of Sec. 3.1, the Wiener filter is designed by using the model parameters of the speech $\mathcal{M}_S$ and the target BF output $\mathcal{M}_{Y,\tau}$. Each model has J states, and each state consists of K Gaussian components. The Wiener filter is calculated using Eq. (16) if the I-dimensional LCPSD vector of the current target BF output $\boldsymbol{\phi}^{\log}_{Y_1,\tau}$ is deterministically known to belong to the j-th state and k-th Gaussian component.

$$\mathcal{G}_{\text{Prop},\Omega,\tau,j,k} = \frac{\exp\left(\mu_{S,\Omega,j,k}\right)}{\exp\left(\mu_{S,\Omega,j,k}\right) + \hat{\phi}_{N,\Omega,\tau}} \tag{16}$$

In contrast to Eq. (9), the estimated target speech CPSD $\hat{\phi}_{S,\Omega,\tau}$ is substituted with the exponential mean contained in the clean speech model $\exp\left(\mu_{S,\Omega,j,k}\right)$ in Eq. (16).

However, $\boldsymbol{\phi}^{\log}_{Y_1,\tau}$ would belong to every state and every component with a certain probability. Therefore, the Wiener filter is expressed as Eqs. (17) and (18) by weighted summing $\mathcal{G}_{\text{Prop},\Omega,\tau,j,k}$ w.r.t. each state and each Gaussian component depending on the posterior probability.

$$G_{\text{Prop},\omega,\tau} = \sum_{j=1}^{J}\sum_{k=1}^{K} P\left(j,k\,\middle|\,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right)\cdot\mathcal{G}_{\text{Prop},\Omega,\tau,j,k} \tag{17}$$

$$P\left(j,k\,\middle|\,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right) = P\left(j\,\middle|\,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right) P\left(k\,\middle|\,j,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right) \tag{18}$$

The $P\left(j,k\,\middle|\,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right)$ denotes the posterior probability w.r.t. the j-th state and k-th Gaussian component.

From Bayes' theorem, the $P\left(k\,\middle|\,j,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right)$ is expressed as Eq. (19).

$$P\left(k\,\middle|\,j,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right) = \frac{p\left(\boldsymbol{\phi}^{\log}_{Y_1,\tau}\,|j,k\right)P\left(k\,|j\right)}{\sum_{k=1}^{K} p\left(\boldsymbol{\phi}^{\log}_{Y_1,\tau}\,|j,k\right)P\left(k\,|j\right)} \tag{19}$$

To calculate $P\left(k\,\middle|\,j,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right)$, the likelihood of the corresponding Gaussian component $p\left(\boldsymbol{\phi}^{\log}_{Y_1,\tau}\,\middle|\,j,k\right)$ is calculated as Eq. (20), where $\mathcal{N}(\cdot|\cdot)$ denotes the probability density function of the multivariate Gaussian distribution.

$$p\left(\boldsymbol{\phi}^{\log}_{Y_1,\tau}\,\middle|\,j,k\right) = \mathcal{N}\left(\boldsymbol{\phi}^{\log}_{Y_1,\tau}\,\middle|\,\boldsymbol{\mu}_{Y,\tau,j,k},\boldsymbol{\Sigma}_{Y,j,k}\right) \tag{20}$$

The $P(k|j)$ is regarded as Eq. (21).

$$P(k|j) = \lambda_{Y,j,k} \tag{21}$$

The $P\left(j|\,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right)$ is derived as Eq. (22), unlike the general method that computes it sequentially with the HMM's state transition probability.

$$P\left(j\,\middle|\,\boldsymbol{\phi}^{\log}_{Y_1,\tau}\right) = \begin{cases} 1 - G_{\omega,\tau} & (j=1) \\ G_{\omega,\tau} & (j=2) \end{cases} \tag{22}$$

Therefore, it is not necessary to train the HMM's state transition probability in advance while its output probability is trained as the conventional technique [16].

Table 1. Type and angle of interference noise

Test	Type	Angle, θ_N
1	speech	90°
2	speech	45°
3	music	45°

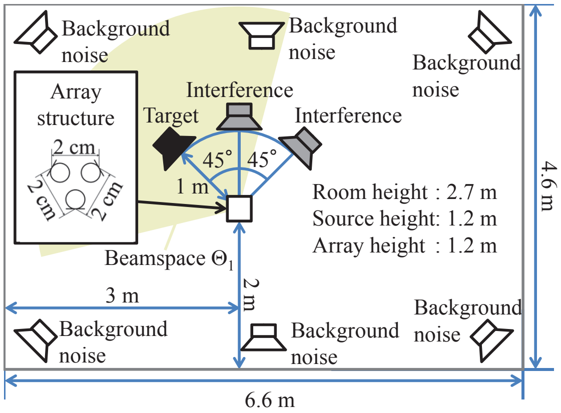

Fig. 3. Noise and impulse response measurement setup to create evaluation data simulating microphone array observation

Finally, the Wiener filter is obtained by substituting Eqs. (16), (19) and (22) into Eq. (17).

4. EXPERIMENT

The performance of the proposed technique was investigated by evaluating the output SNR and comparing it with a conventional technique [10] based on only spatial cues.

4.1. Setup

The microphone array consisted of three cardioid microphones. Each microphone was turned 120° from the others.

Evaluation data were obtained by simulating the microphone array observation as follows. We measured the impulse responses of the target and interference by using a set of microphone arrays in a reverberant chamber. The measurement setup is shown in Fig. 3. The measurements were carried out under two reverberation time conditions to confirm that the proposed technique is effective in various environments. The angle between the target and interference θ_N was set to 45° or 90°. Clean speech signals and interference signals were convolved with the recorded impulse responses to make the target sound and interference noise. Table 1 summarizes the interference source types and angles with the target. Three different types of background noise that had been recorded in offices, shopping centers, and exhibition halls were played from the loudspeakers against a wall. The impulse responses and background noise were measured at different times with the microphone array, and the target sound, interference noise, and background noise were added with different SNRs through simulation. The background noise level was varied from -10 to 10 dB compared to the target, while the interference noise level was the same as the target.

We trained the clean speech models by using a training set of the WSJ0 corpus [22]. The speech model had $J = 2$ states and each state had $K = 64$ Gaussian components. The feature parameters of the speech model were $I = 40$-dimensional LCPSDs.

Table 2. Experimental conditions

Sampling rate	16 kHz
# of microphones, M	3
# of interference noise, Q	1
Angle, θ_N	$45°, 90°$
Target source distance	1.0 m
Background noise level to speech level	$-5, 0, 5, 10$ dB
Reverberation time (1 kHz)	$230, 350$ ms
Frame length	32 ms
Frame shift	16 ms
# of HMM states, J	2
# of Gaussian mixtures, K	64
# of filter bank channels, I	40
Training data	WSJ0

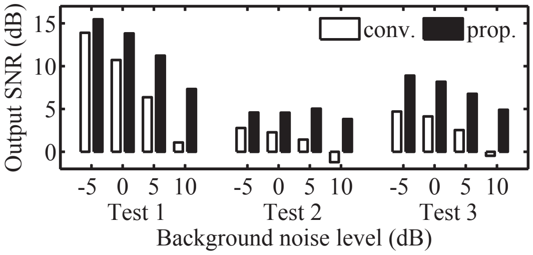

Fig. 4. Experimental evaluation results of output SNR

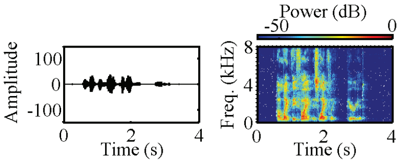

Fig. 5. Target signal

Clean data for evaluation was taken from the evaluation set of the WSJ0 corpus. Eight utterances were used for the target speech and 64 were used for the interference. They had been spoken by four males and four females. Sixteen utterances were evaluated under each noise condition. Table 2 summarizes the experimental conditions.

4.2. Results

Fig. 4 summarizes the results of the SNR evaluation, averaging the results corresponding to the two reverberation time conditions. Fig. 5 shows the waveforms and spectrograms of the target signal and Figs. 6–8 show those of captured and output signals, respectively.

The background noise levels indicated on the horizontal axis of Fig. 4 do not include the interference noise level; thus, the total noise levels were higher than these values. The output SNR includes both the interference noise and background noise. From Fig. 4, it was confirmed that the proposed technique successfully outperformed the conventional technique in terms of noise reduction performance under all experimental conditions. As seen in Fig. 4, the output SNR did not improve very much in Test 2, where the angle between the target and interference was $45°$. In Test 2, it was difficult to separate the interference by spatial cues and speech model since the source was speech. However, the interference was reduced in Test 3, where the source was music, even though the angle between the target and interference was $45°$. It should be noted from the preliminary listening test that musical noise occurred in the output of the conventional technique drastically reduced with the proposed technique.

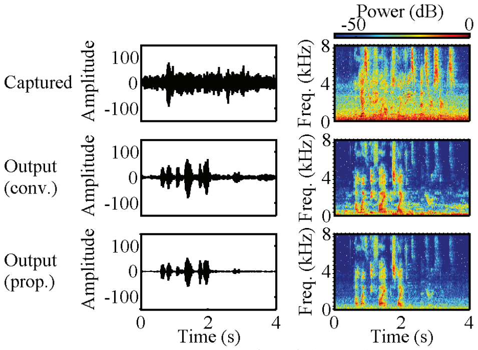

Fig. 6. Output signal (Test 1)

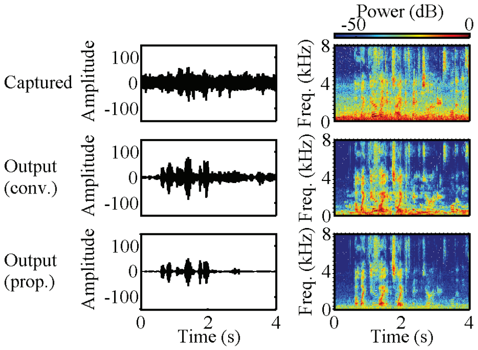

Fig. 7. Output signal (Test 2)

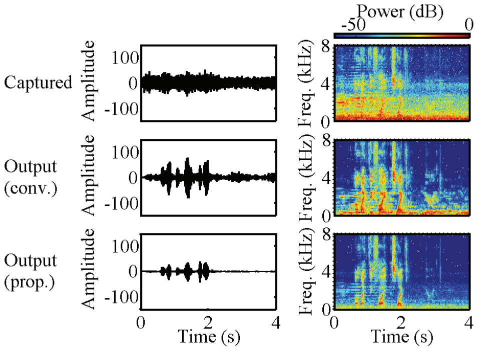

Fig. 8. Output signal (Test 3)

5. CONCLUSION

We proposed a technique that integrates the *PSD-estimation-in-beamspace* method and statistical model-based speech enhancement. The observation models were composed of speech models and noise PSDs estimated using the *PSD-estimation-in-beamspace* method. A Wiener filter was designed based on Bayes' theorem using the observation models and beamforming output. The experimental results in several different noise environments showed that SNR improved compared with the conventional technique under all experimental conditions. Future work should include sound quality evaluations by using formal listening tests.

6. REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, Springer, 2005.

[2] C. Marro, Y. Mahieux, and K. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, pp. 240–259, 1998.

[3] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 1988, vol. 5, pp. 2578–2581.

[4] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 709–716, 2003.

[5] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin Heidelberg, 2001.

[6] K. Kumatani, B. Raj, R. Singh, and J. W. McDonough, "Microphone array post-filter based on spatially-correlated noise measurements for distant speech recognition," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, 2012, pp. 298–301.

[7] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Simon & Schuster, Englewood Cliffs, NJ, 1993.

[8] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Speech Acoust., Speech, Signal Proc.*, vol. 35, pp. 1365–1376, 1987.

[9] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 1240–1250, 2013.

[10] K. Niwa, Y. Hioka, and K. Kobayashi, "Post-filter design for speech enhancement in various noisy environments," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, 2014, pp. 35–39.

[11] Y. Hioka and K. Niwa, "Psd estimation in beamspace for source separation in a diffuse noise field," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, 2014, pp. 85–88.

[12] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96., 1996 International Conference on*, 1996, vol. II, pp. 733–736.

[13] J. C. Segura, Á. de la Torre, M. C. Benítez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. experiments using the aurora II database and tasks," in *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event*, 2001, pp. 221–224.

[14] M. Fujimoto and S. Nakamura, "Particle filter based non-stationary noise tracking for robust speech recognition," in *Acoustics, Speech and Signal Processing, 2005. ICASSP 2005 Proceedings. 2005 IEEE International Conference on.*, 2005, vol. I, pp. 257–260.

[15] T. Arakawa, M. Tsujikawa, and R. Isotani, "Model-based wiener filter for noise robust speech recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on.*, 2006, vol. I.

[16] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A study of mutual front-end processing method based on statistical model for noise robust speech recognition," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 1235–1238.

[17] W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamura, "Hands-free speech recognition and communication on PDAs using microphone array technology," in *Automatic Speech Recognition and Understanding, 2005. ASRU 2005 Proceedings. 2006 IEEE International Workshop on.*, 2005, pp. 302–307.

[18] X. Zhao and Z. Ou, "Closely coupled array processing and model-based compensation for microphone array speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 1114–1122, 2007.

[19] T. Nakatani, M. Souden, S. Araki, T. Yoshioka, T. Hori, and A. Ogawa, "Coupling beamforming with spatial and spectral feature based spectral enhancement and its application to meeting recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7249–7253.

[20] M. Wolfel and J. McDonough, "Minimum variance distortionless response spectral estimation," *Signal Processing Magazine, IEEE*, vol. 22, pp. 117–126, 2005.

[21] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory- filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, pp. 750–753, 1983.

[22] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," https://catalog.ldc.upenn.edu/LDC93S6A.