

MULTICHANNEL ONLINE SPEECH DEREVERBERATION UNDER NOISY ENVIRONMENTS

Masahito Togami

Central Research Laboratory, Hitachi Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan

ABSTRACT

In this paper, we propose a novel online speech dereverberation with multichannel microphone input signals for noisy environments. Unlike conventional dereverberation methods which optimizes the dereverberation filter by noisy microphone input signals, the proposed method optimizes the dereverberation filter by noiseless microphone input signals so as to achieve a good dereverberation filter under noisy environments. Noiseless microphone input signals are estimated by multichannel Wiener filtering which can be interpreted as combination of multichannel beamforming and time-varying single-channel Wiener filtering. In multichannel Wiener filtering, residual reverberation which cannot be reduced by the time-invariant dereverberation filter is also reduced. Optimization of the parameters are updated by using the expectation-maximization algorithm in an online manner. Experimental results show that the proposed method can reduce reverberation and background noise effectively in an online manner even when microphone input signals are observed under noisy environments.

Index Terms— Dereverberation, noise robustness, local Gaussian modeling, multichannel Wiener filtering, EM algorithm

1. INTRODUCTION

Speech signal captured in real environments is contaminated with background noise and reverberation. In the teleconferencing systems, noise reduction and speech dereverberation techniques are highly required for smooth conversation. Automatic speech recognition systems also require noiseless and dereverberated signals so as to achieve good recognition performance. In this context, speech dereverberation techniques with auto-regressive models based on multiple-input/output inverse-filtering (MINT) theorem [1] have been actively studied [1][2][3][4]. The speech dereverberation techniques with auto-regressive models can reduce reverberation under noiseless environments. However, when microphone input signals are contaminated by background noise, dereverberation performance degrades, because the dereverberation filters are poorly optimized under noisy environments.

In the previous work by authors [5], we propose a noise robust speech dereverberation technique, which is an extension of the speech dereverberation technique with auto-regressive models [4]. In the conventional methods, noisy microphone input signals are utilized so as to update the dereverberation filters, which causes poorly optimization of the dereverberation filters. On contrary, our proposed technique optimizes the dereverberation filters from noiseless microphone input signals which is estimated by using Kalman smoother [6]. The proposed method can reduce reverberation and background noise more effectively than the conventional methods.

However, this method is an offline method which optimizes all parameters after that all of the microphone input signals are obtained.

In this paper, we propose a novel online speech dereverberation technique which is robust against background noise signal. Yoshioka et al. [7] proposed an online speech dereverberation technique with single microphone input signal, which is based on an auto-regressive model of noisy microphone input signals. On contrary to the Yoshioka's method, the proposed method utilizes multichannel microphone input signal, and the dereverberation filter is optimized with an auto-regressive model of noiseless microphone input signals. To estimate the noiseless microphone input signals, the proposed method performs multichannel Wiener filtering like the previously proposed offline Kalman smoother based dereverberation technique. Multichannel Wiener filtering can be interpreted as combination of a single-channel noise reduction and a multichannel spatial beamformer. In the multichannel Wiener filtering, non-stationary characteristics of speech sources and stationary characteristics of stationary background noise signals are utilized by using the local Gaussian model [8] so as to extract speech sources effectively. In addition to noise reduction, residual reverberation is also reduced by multichannel Wiener filtering. The time-varying covariance matrix which reflects uncertainty of the dereverberation filter is considered in the local Gaussian model. Experimental results under noisy environments show that the proposed method can reduce background noise and reverberation effectively.

2. PROBLEM STATEMENT

2.1. Input signal model

The proposed method performs speech dereverberation and noise reduction in the time-frequency domain. The microphone input signal, $\mathbf{x}_{l,k}$ (l is frame index, k is frequency index), is modeled as $\mathbf{x}_{l,k} = [x_{l,k,1} \ \dots \ x_{l,k,N_m}]^T$, where N_m is the number of the microphones and T is the transpose operator of a matrix/vector. Under the assumption that there are one speech source and background noise signal, the microphone input signal is modeled as follows:

$$\mathbf{x}_{l,k} = \sum_{t=0}^{L_{\text{imp}}-1} \mathbf{h}_{k,t} s_{l-t,k} + \mathbf{w}_{l,k}, \quad (1)$$

where L_{imp} is the length of the impulse response, $\mathbf{h}_{k,t}$ is a vector which is composed of multichannel impulse responses which is defined as

$$\mathbf{h}_{k,t} = [h_{k,1,t} \ \dots \ h_{k,N_m,t}]^T, \quad (2)$$

$h_{k,m,t}$ is the t th tap of the impulse response between the speech source and the m th microphone, $s_{l,k}$ is the original signal, and $\mathbf{w}_{l,k}$ is the multichannel noise signal.

The proposed method performs dereverberation and noise reduction for each frequency independently. Therefore, the frequency index k is omitted from this. When there is no background noise, Eq. 1 can be expressed as follows:

$$\mathbf{c}_l = \sum_{t=0}^{L_{\text{imp}}-1} \mathbf{h}_t s_{l-t}, \quad (3)$$

where \mathbf{c}_l is multichannel noiseless microphone input signal. Eq. 3 can be converted into the following auto-regressive model [4]:

$$\mathbf{c}_l = \sum_{t=D}^{L_{\text{AR}}-1} \mathbf{G}_t \mathbf{c}_{l-t} + \mathbf{u}_l, \quad (4)$$

where L_{AR} is the length of the auto-regressive coefficients, \mathbf{G}_t is the t th tap of the auto-regressive coefficient, D is the tap-length of the early reflection, and \mathbf{u}_l is the summation of the direct path and the early reflection, which is defined as follows:

$$\mathbf{u}_l = \sum_{t=0}^{D-1} \mathbf{h}_t s_{l-t}. \quad (5)$$

The multichannel observed signal can be expressed as follows:

$$\mathbf{x}_l = \mathbf{c}_l + \mathbf{w}_l. \quad (6)$$

In the previously proposed Kalman smoother based speech dereverberation technique [5], (4) is regarded as a state-transition equation, and (6) is regarded as an observation equation. Sufficient statistics of latent variables \mathbf{c}_l and all the parameters are estimated so as to increase the likelihood function monotonically. However, the previously proposed technique is an offline method. The state vector is updated after that the microphone input signals of all frames are observed. In this paper, we propose an online extension of the previously proposed technique, which estimates the sufficient statistics of the latent variables and updates all the parameters in an online manner.

3. PROPOSED METHOD

3.1. Optimization of dereverberation filter

The multichannel speech dereverberation filter $\mathcal{G} = \{\mathbf{G}_t\}_{D \leq t \leq L_{\text{AR}}-1}$ is updated by using a Bayesian approach [7] based on maximum a posteriori estimation in an online manner as follows:

$$\begin{aligned} \hat{\mathcal{G}}_l &= \underset{\mathcal{G}}{\operatorname{argmax}} p(\mathcal{G}|\mathcal{C}_l), \\ &= \underset{\mathcal{G}}{\operatorname{argmax}} p(\mathbf{c}_l|\mathcal{G}, \mathcal{C}_{l-1})p(\mathcal{G}|\mathcal{C}_{l-1}), \end{aligned} \quad (7)$$

where $\hat{\mathcal{G}}_l$ is estimation of \mathcal{G} after that the l th observed signal is obtained and $\mathcal{C}_l = \{\mathbf{c}_{l'}\}_{1 \leq l' \leq l}$. The probabilistic density function (PDF) of the reverberant noiseless speech signal \mathbf{c}_l is modeled as a time-varying Gaussian distribution [4] as follows:

$$p(\mathbf{c}_l|\mathcal{G}, \mathcal{C}_{l-1}) = \mathcal{N}\left(\sum_{t=D}^{L_{\text{AR}}-1} \mathbf{G}_t \mathbf{c}_{l-t}, v_l \mathbf{R}\right), \quad (8)$$

where v_l is the time-varying variance of the original speech signal, \mathbf{R} is the covariance matrix of the steering vector of summation of the direct path and the early reflection. Under the condition that $l - 1$ th reverberant and noiseless multichannel signal is given, the

PDF of the auto-regressive model is set to the following Gaussian distribution:

$$p(\mathcal{G}|\mathcal{C}_{l-1}) = \mathcal{N}(\hat{\mathcal{G}}_{l-1}, \hat{\mathbf{V}}_{l-1}). \quad (9)$$

Therefore,

$$p(\mathcal{G}|\mathcal{C}_l) = \mathcal{N}(\hat{\mathcal{G}}_l, \hat{\mathbf{V}}_l), \quad (10)$$

where

$$\begin{aligned} \hat{\mathcal{G}}_l &= \hat{\mathbf{V}}_l (\mathbf{D}_{l-1}^H (v_l \mathbf{R})^{-1} \mathbf{c}_l + \mathbf{V}_{l-1}^{-1} \hat{\mathcal{G}}_{l-1}), \\ \hat{\mathbf{V}}_l &= (\mathbf{D}_{l-1}^H (v_l \mathbf{R})^{-1} \mathbf{D}_{l-1} + \mathbf{V}_{l-1}^{-1})^{-1}, \end{aligned} \quad (11)$$

where \mathbf{D}_{l-1} can be calculated from \mathcal{C}_{l-1} . By using the forgetting factor α , $\hat{\mathcal{G}}_l$ and $\hat{\mathbf{V}}_l$ are updated in an online manner as follows:

$$\begin{aligned} \hat{\mathcal{G}}_l &= \hat{\mathbf{V}}_l (\mathbf{D}_{l-1}^H (v_l \mathbf{R})^{-1} \mathbf{c}_l + \alpha \mathbf{V}_{l-1}^{-1} \hat{\mathcal{G}}_{l-1}), \\ \hat{\mathbf{V}}_l &= (\mathbf{D}_{l-1}^H (v_l \mathbf{R})^{-1} \mathbf{D}_{l-1} + \alpha \mathbf{V}_{l-1}^{-1})^{-1}. \end{aligned} \quad (12)$$

$p(\mathcal{G}|\mathcal{C}_l)$ assumes that \mathcal{C}_l is obtained, but \mathcal{C}_l is the noiseless microphone input signal, which is not observed in advance. In the conventional method, the original microphone input signals are set to \mathcal{C}_l as $\mathcal{C}_l = \mathcal{X}_l$. However, the noisy microphone input signal is harmful for optimization of dereverberation filters [5]. In the proposed method, we utilize an estimated noiseless microphone input signals as follows:

$$\mathcal{C}_l = E[\mathcal{C}_l | \mathbf{x}_l, \mathcal{C}_{l-1}, \hat{\mathcal{G}}_{l-1}, \hat{\mathbf{V}}_{l-1}], \quad (13)$$

where $E[\cdot]$ is an operator which calculates mathematical expectation.

3.2. Estimation of noiseless microphone input signal

The noiseless microphone input signal of the l th frame is estimated as follows:

$$\begin{aligned} \mathbf{c}_l &= \sum_{t=D}^{L_{\text{AR}}-1} E[\mathbf{G}_t \mathbf{c}_{l-t} | \mathbf{x}_l, \mathcal{C}_{l-1}, \hat{\mathcal{G}}_{l-1}, \hat{\mathbf{V}}_{l-1}] \\ &+ E[\mathbf{u}_l | \mathbf{x}_l, \mathcal{C}_{l-1}, \hat{\mathcal{G}}_{l-1}, \hat{\mathbf{V}}_{l-1}], \end{aligned} \quad (14)$$

where

$$E[\mathbf{G}_t \mathbf{c}_{l-t} | \mathbf{x}_l, \mathcal{C}_{l-1}, \hat{\mathcal{G}}_{l-1}, \hat{\mathbf{V}}_{l-1}] = \hat{\mathbf{G}}_{l-1,t} \mathbf{c}_{l-t} + \hat{\mathbf{r}}_l, \quad (15)$$

where $\hat{\mathbf{r}}_l$ is estimation of the residual reverberation. The expected values of the early reflection and the residual reverberation are $E[\mathbf{u}_l | \mathbf{x}_l, \mathcal{C}_{l-1}, \hat{\mathcal{G}}_{l-1}, \hat{\mathbf{V}}_{l-1}] = E[\mathbf{u}_l | \mathbf{y}_l, \hat{\mathbf{V}}_{l-1}]$ and $E[\mathbf{r}_l | \mathbf{x}_l, \mathcal{C}_{l-1}, \hat{\mathcal{G}}_{l-1}, \hat{\mathbf{V}}_{l-1}] = E[\mathbf{r}_l | \mathbf{y}_l, \hat{\mathbf{V}}_{l-1}]$, where $\mathbf{y}_l = \mathbf{x}_l - \hat{\mathcal{G}}_{l-1} \mathcal{C}_{l-1}$. $\hat{\mathbf{u}}_l = E[\mathbf{u}_l | \mathbf{y}_l, \hat{\mathbf{V}}_{l-1}]$ can be obtained by using multichannel Wiener filtering as $\hat{\mathbf{u}}_l = v_l \mathbf{R} (\mathbf{R}_{y,l})^{-1} \mathbf{y}_l$. The residual reverberation is also estimated as $\hat{\mathbf{r}}_l = \mathbf{D}_{l-1} \hat{\mathbf{V}}_{l-1} \mathbf{D}_{l-1}^H \mathbf{R} (\mathbf{R}_{y,l})^{-1} \mathbf{y}_l$, where $\mathbf{R}_{y,l}$ is the covariance matrix of \mathbf{y}_l , which is calculated as follows:

$$\mathbf{R}_{y,l} = v_l \mathbf{R} + \mathbf{R}_w + \mathbf{D}_{l-1} \hat{\mathbf{V}}_{l-1} \mathbf{D}_{l-1}^H, \quad (16)$$

$\mathbf{D}_{l-1} \hat{\mathbf{V}}_{l-1} \mathbf{D}_{l-1}^H$ is the covariance matrix of the residual reverberation. Therefore, the noiseless reverberant microphone input signals is obtained as follows:

$$\mathbf{c}_l = \sum_{t=D}^{L_{\text{AR}}-1} \hat{\mathbf{G}}_{l-1,t} \mathbf{c}_{l-t} + (v_l \mathbf{R} + \mathbf{D}_{l-1} \hat{\mathbf{V}}_{l-1} \mathbf{D}_{l-1}^H) (\mathbf{R}_{y,l})^{-1} \mathbf{y}_l. \quad (17)$$

3.3. Parameter optimization at each frame

In each frame, the parameters $\theta = \{v_l, \mathbf{R}, \mathbf{R}_w\}$ are updated so as to increase the likelihood function $p(\theta|\mathbf{x}_l, \mathcal{C}_{l-1}, \hat{\mathbf{G}}_{l-1}, \hat{\mathbf{V}}_{l-1})$ by using online local Gaussian modeling [9]. The E-step and the M-step are performed in an iterative manner.

E step at t th iteration:

$$\mathbf{R}_{y,l} = v_l^{(t-1)} \mathbf{R}^{(t-1)} + \mathbf{R}_w^{(t-1)} + \mathbf{D}_{l-1} \hat{\mathbf{V}}_{l-1} \mathbf{D}_{l-1}^H, \quad (18)$$

$$\mathbf{W}_u = v_l^{(t-1)} \mathbf{R}^{(t-1)} \mathbf{R}_{y,l}^{-1}, \quad (19)$$

$$\mathbf{W}_w = \mathbf{R}_w^{(t-1)} \mathbf{R}_{y,l}^{-1}, \quad (20)$$

$$\hat{\mathbf{c}}_l = \mathbf{W}_u \mathbf{y}_l, \quad (21)$$

$$\hat{\mathbf{c}}_{w,l} = \mathbf{W}_w \mathbf{y}_l, \quad (22)$$

$$\mathbf{R}_{c,n,l} = \hat{\mathbf{c}}_l \hat{\mathbf{c}}_l^H + (\mathbf{I} - \mathbf{W}_u) v_l^{(t-1)} \mathbf{R}^{(t-1)}, \quad (23)$$

$$\mathbf{R}_{c,w,l} = \hat{\mathbf{c}}_{w,l} \hat{\mathbf{c}}_{w,l}^H + (\mathbf{I} - \mathbf{W}_w) \mathbf{R}_w^{(t-1)}. \quad (24)$$

M step at t th iteration:

$$v_l^{(t)} = \frac{1}{N_m} \text{tr}(\mathbf{R}^{(t-1)} \mathbf{R}_{c,n,l}), \quad (25)$$

$$\mathbf{R}^{(t)} = \beta \mathbf{R} + (1 - \beta) \frac{1}{v_l^{(t)}} \mathbf{R}_{c,n,l}, \quad (26)$$

$$\mathbf{R}_w^{(t)} = \beta \mathbf{R}_w + (1 - \beta) \mathbf{R}_{c,w,l}, \quad (27)$$

where β is the forgetting factor for the parameters. At last, the covariance matrix is updated as $\mathbf{R} = \mathbf{R}^{(N_i)}$, $\mathbf{R}_w = \mathbf{R}_w^{(N_i)}$, where the N_i is the number of iterations.

3.4. Estimation of output signal after dereverberation and noise reduction

Finally, the output signal is obtained by using the multichannel Wiener filter as $\hat{\mathbf{c}}_l = \mathbf{W}_u \mathbf{y}_l$, where the residual reverberation and the background noise signal are reduced by the multichannel Wiener filter.

4. EXPERIMENT

4.1. Experimental conditions

Experimental environment and microphone array alignment are shown in Fig. 1. The reverberation time was about 700 ms. The impulse responses were recorded at Location 1, 2, 3 by using TSP (Time Stretched Pulse) method [10]. The speech source is convolved with the recorded impulse response. The original source signals were extracted from TIMIT database [11]. Two utterances of each speaker were merged into single source signal. The number of the speech sources was 34. The parameters are shown in Tab. 1.

S/N of the microphone input signal was set to 0, 10 dB, 20 dB.

In this experiments, the following 4 methods were compared.

- M1: No noise reduction and no residual noise reduction with time-invariant assumption of a speech source, optimization of dereverberation filter from noisy observed signal ($\mathbf{R}_w = \mathbf{0}$, $v_l = 1$, $\mathbf{c}_l = \mathbf{x}_l$, $\mathcal{C}_l = \mathcal{X}_l$, and $\mathbf{R}_{y,l} = v_l \mathbf{R}$)
- M2: M1 + time-varying assumption of a speech source
- M3: M2+ noise reduction
- M4: M3+ optimization of dereverberation filter from noiseless observed signal

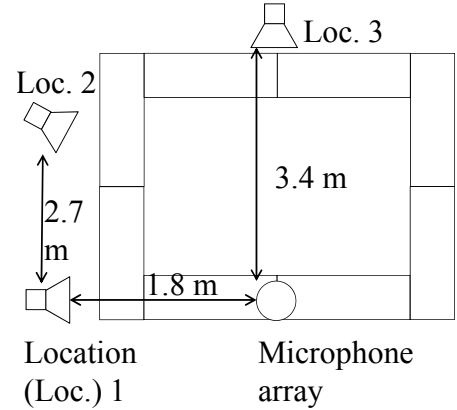


Fig. 1. Experimental environment and microphone array alignment

Table 1. Evaluation conditions

Sampling rate (Hz)	16,000
Frame size (pt)	1024
Frame shift (pt)	256
Number of microphones N_m	3
Number of EM iterations	20
α	0.96
β	0.99
D	2
L_w	10

- PROPOSED: M4+ residual noise reduction (the proposed method)

4.2. Evaluation results

Evaluation measure was SDR (Signal To Distortion Ratio) improvement, which is the averaged value of 34 utterances. The evaluation results when the S/N of the microphone input signal is 0 dB, 10 dB, 20 dB are shown in Tab. 2. The experimental results for the first utterance and the second utterance are shown separately for each result. In each result, the experimental result show that the proposed method can reduce background noise and reverberation more effectively than the other methods. The output waveforms and spectrograms are shown in Fig. 2. It is shown that the proposed method can reduce more background noise and reverberation than the other methods.

5. CONCLUSION

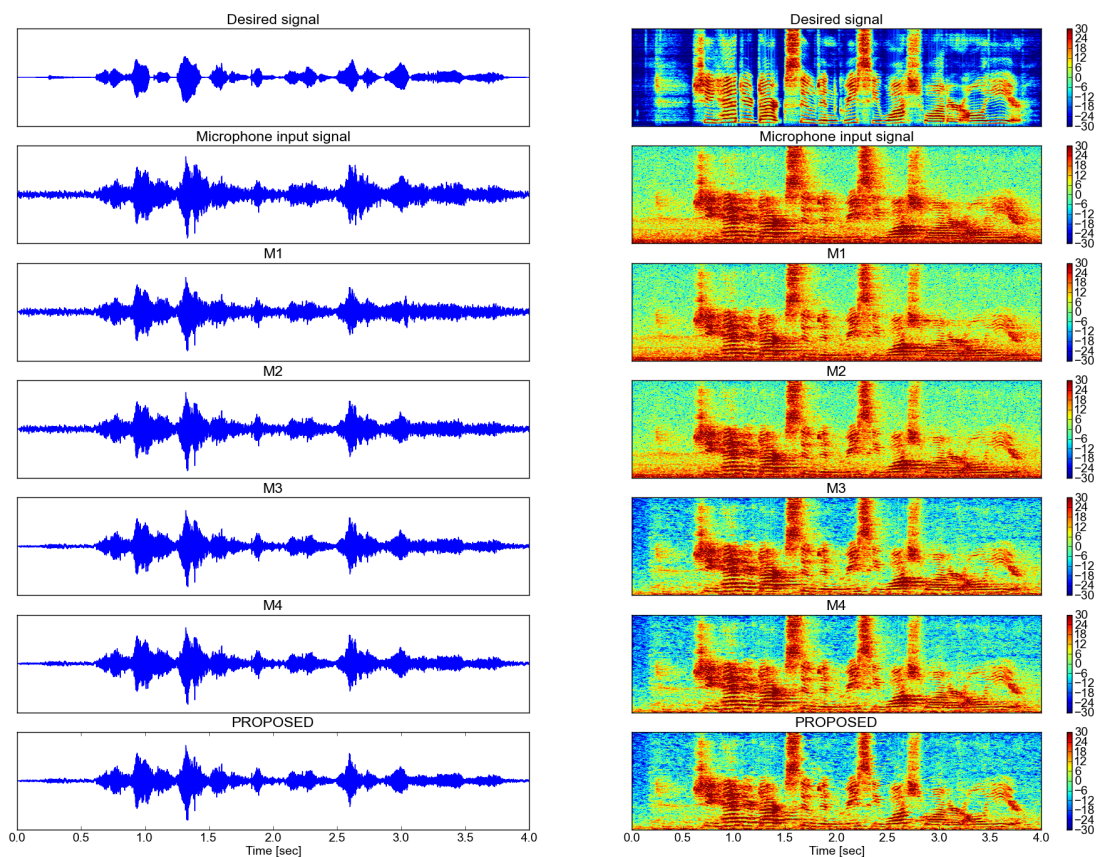
In this paper, we propose a novel online speech dereverberation with multichannel microphone input signals for noisy environments, in which noise reduction and residual reverberation reduction are integrated effectively. Experimental results show that the proposed method can reduce reverberation and background noise effectively.

REFERENCES

- [1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, vol. 30, no. 2, pp. 145-152, Feb. 1988.

Table 2. Evaluation results (SDR improvement): “Bold” means highest SDR improvement at each row.

			PROPOSED	M1	M2	M3	M4
SNR 0 dB case	LOCATION 1	FIRST	4.11	0.83	0.77	3.61	3.48
		SECOND	6.80	0.16	1.27	4.09	4.77
	LOCATION 2	FIRST	3.81	0.92	0.75	3.36	3.24
		SECOND	6.82	0.54	1.34	4.10	4.69
	LOCATION 3	FIRST	3.70	0.86	0.71	3.29	3.16
		SECOND	6.68	0.65	1.34	4.02	4.57
SNR 10 dB case	LOCATION 1	FIRST	2.27	1.19	1.13	1.77	1.71
		SECOND	6.59	-0.54	3.58	4.74	5.40
	LOCATION 2	FIRST	1.99	1.33	1.00	1.54	1.49
		SECOND	6.33	0.55	3.54	4.55	5.02
	LOCATION 3	FIRST	1.84	1.17	0.91	1.44	1.38
		SECOND	6.17	0.88	3.48	4.42	4.82
SNR 20 dB case	LOCATION 1	FIRST	1.99	1.44	1.45	1.53	1.50
		SECOND	7.07	-0.75	5.95	6.19	6.71
	LOCATION 2	FIRST	1.86	1.56	1.23	1.30	1.27
		SECOND	6.75	0.69	5.55	5.69	6.01
	LOCATION 3	FIRST	1.62	1.38	1.10	1.17	1.14
		SECOND	6.43	1.04	5.37	5.50	5.74

**Fig. 2.** Examples of output waveforms and spectrograms

- [2] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction,” *IEEE*

Trans. Audio, Speech, and Language Process., vol. 17, no. 4, pp. 534–545, May 2009.

- [3] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno,

- “Blind separation and dereverberation of speech mixtures by joint optimization,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.
- [4] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, “Optimized Speech Dereverberation From Probabilistic Perspective for Time Varying Acoustic Transfer Function,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 7, pp. 1369–1380, Jul. 2013.
- [5] M. Togami and Y. Kawaguchi, “Noise Robust Speech Dereverberation with Kalman Smoother,” *Proc. ICASSP2013*, pp. 7447–7451, 2013/5.
- [6] A.H. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic Press, 1970.
- [7] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, “Adaptive dereverberation of speech signals with speaker-position change detection,” in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, 2009, pp. 3733–3736.
- [8] N.Q.K. Duong, E. Vincent, R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Trans. Speech Audio Process.*, vol. 18, no. 7, pp. 1830–1840, 2010/9.
- [9] M. Togami, “Online speech source separation based on maximum likelihood of Local Gaussian modeling,” *IEEE ICASSP2011*, pp. 213–216, 2011.
- [10] 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,” *J. Acoust. Soc. Amer.* vol. 97, no. 2, pp. 1119–1123, Feb. 1995.
- [11] TIMIT corpus [Online]. Available: <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1>.