ENHANCEMENT OF NOISY REVERBERANT SPEECH BY LINEAR FILTERING FOLLOWED BY NONLINEAR NOISE SUPPRESSION

Takuya Yoshioka, Tomohiro Nakatani, and Masato Miyoshi

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
email: takuya@cslab.kecl.ntt.co.jp

ABSTRACT

This paper proposes a novel method for enhancing noisy reverberant speech. One conventional approach is first suppressing noise by nonlinear filtering techniques such as spectral subtraction and then dereverberating the noise-suppressed reverberant speech by linear filtering. However, this approach sometimes suffers from poor performance because the nonlinearly filtered signals no longer have any linear relationship with the clean speech signals. Unlike this approach, the proposed method obtains an enhanced speech signal by using a linear dereverberation filter followed by a nonlinear noise suppression filter. Moreover, the linear filters are optimized directly from the observed signals by considering the presence of noise. The proposed method is derived based on the maximum likelihood (ML) estimation method. Experimental results showed the superiority of the proposed method to the conventional approach.

Index Terms— Speech enhancement, dereverberation, noise suppression

1. INTRODUCTION

Speech signals captured by microphones in rooms are often distorted by both reverberation and background noise. Fig. 1 shows an acoustic system that generates this distortion. As shown in Fig. 1, it is assumed that there is one speaker and one or more microphones and that the noise is stationary. The recovery of an original clean speech signal from observed noisy reververant speech signals will be indispensable for many audio applications.

If the noise is negligible, the speech enhancement task of interest reduces to a speech dereverberation task. Conventionally, the dereverberation has often been realized by linearly filtering reverberant signals both in the time domain [1] and in the frequency (or subband) domain [2, 3]. This is because reverberation is mathematically represented as the linear filtering of a clean speech signal with a room transfer function (RTF). As long as the speaker and microphones do not move during the observation, linear filtering based dereverberation yields enhanced speech of better quality than spectral subtraction (SS) based dereverberation (and noise suppression) methods [4, 5].

As regards the enhancement of noisy reverberant speech, the method proposed in [6] first suppresses the noise by the SS-type nonlinear filtering of the observed signals to estimate the noise-free reverberant speech signals. Then, it processes the estimated noise-free reverberant speech through a linear dereverberation filter to estimate the reverberation components contained in the noise-free reverberant speech estimates. Our recently developed speech enhancement method [7] suppresses the noise while taking account of the presence of reverberation. At the same time, it optimizes the dereverberation filter while taking the estimation errors of the noise-free reverberant speech signals into consideration. However, this method also obtains an estimate of the clean speech signal by nonlinear noise suppression filtering followed by linear dereverberation filtering.

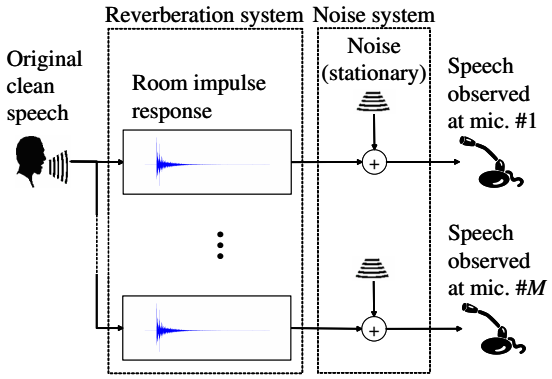

Fig. 1. Acoustic system of interest.

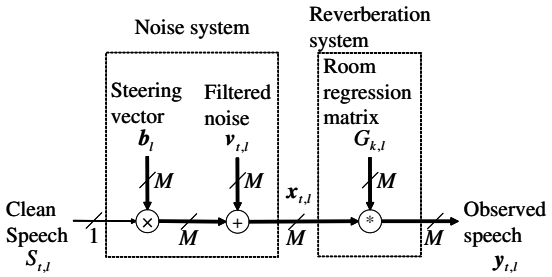

Fig. 2. Equivalent acoustic system.

We found that this nonlinear-then-linear filtering approach performed poorly especially when the reverberant signal to noise ratios were low. There are two reasons for this performance degradation.

1. Because of the nonlinear filtering, the estimated noise-free reverberant speech signals no longer have any linear relationship with the clean speech signal.
2. The linear filter amplifies the residual noise contained in the estimated noise-free reverberant speech.

The direct application of a dereverberation filter to noisy reverberant speech signals was proposed in [8]. However, the dereverberation filter is optimized based on correlation matrices obtained by

correlation subtraction. Therefore, the method described in [8] may encounter the same problem.

In this paper, we propose a novel speech enhancement method for overcoming these drawbacks of the conventional approach. The proposed method calculates an enhanced speech signal by linearly filtering the observed signals followed by nonlinear noise suppression filtering. Importantly, the linear filter is optimized directly from the observed signals by considering the presence of noise. The basic idea behind the proposed method is as follows. First, we transform the acoustic system in Fig. 1 into the equivalent acoustic system shown in Fig. 2. Although a detailed description of this system is deferred to Sect. 2, we emphasize here that the reverberation and noise systems are arranged in reverse order in Fig. 2. Then, we set up a statistical model of this acoustic system and estimate the model parameters with the maximum likelihood (ML) estimation method. Using the estimated model parameters, we obtain an enhanced speech signal by processing the observed signals through the inverse of the acoustic system shown in Fig. 2, which consists of a linear dereverberation filter followed by a nonlinear noise suppression filter.

2. TASK FORMULATION

2.1. Noisy reverberant speech enhancement

Suppose that there is one speaker and $M \geq 1$ microphones. Let $s(n)$ be a clean speech signal. A vector of M signals at microphone positions, $\boldsymbol{y}(n) = [y_1(n), \cdots, y_M(n)]^T$, is generated as

$$\boldsymbol{y}(n) = \sum_{k=0}^{\infty} \boldsymbol{h}(k) s(n-k) + \boldsymbol{d}(n), \tag{1}$$

where $\boldsymbol{h}(k) = [h_1(k), \cdots, h_M(k)]^T$ is the vector of the k-th coefficients of the RTFs, $\boldsymbol{d}(n) = [d_1(n), \cdots, d_M(n)]^T$ is the noise signal vector, and superscript T stands for the non-conjugate transposition operator.

Now, suppose that $\boldsymbol{y}(n)$ is observed at times $n = 0, \cdots, N-1$. Then, the noisy reverberant speech enhancement task is defined as estimating $s(n)$ for $n = 0, \cdots, N-1$.

2.2. Models and assumptions

The proposed method is derived based on the short-time Fourier transform (STFT) domain signal representation. Thanks to the STFT domain signal representation, we can employ an effective reverberation model [3], which is proven to be suitable for dereverberation.

Let $s_{t,l}$, $d_{t,l}^{(m)}$ and $y_{t,l}^{(m)}$ be the short-time spectral components of $s(n)$, $d_m(n)$, and $y_m(n)$, respectively, at the t-th frame and the l-th frequency band. We put the noise spectral components into a vector as $\boldsymbol{d}_{t,l} = [d_{t,l}^{(1)}, \cdots, d_{t,l}^{(M)}]^T$. In the same way, we collectively represent the observed spectral components as $\boldsymbol{y}_{t,l} = [y_{t,l}^{(1)}, \cdots, y_{t,l}^{(M)}]^T$. Now, the noisy reverberant speech enhancement task is redefined in the frequency domain as follows. Let L and T denote the number of frequency bands and the number of short-time frames corresponding to time-domain sample number N. Then, the task is to estimate all clean speech spectral components, $\mathcal{S} = \{s_{t,l}\}_{0 \leq t \leq T-1, 0 \leq l \leq L-1}$, from the observed spectral components, $\mathcal{Y} = \{\boldsymbol{y}_{t,l}\}_{0 \leq t \leq T-1, 0 \leq l \leq L-1}$.

In [3], it was shown that observed spectral component vector $\boldsymbol{y}_{t,l}$ is approximately related to clean speech spectral component $s_{t,l}$ by

$$\boldsymbol{y}_{t,l} = \sum_{k=1}^{K_l} G_{k,l}^H (\boldsymbol{y}_{t-k,l} - \boldsymbol{d}_{t-k,l}) + \boldsymbol{b}_l s_{t,l} + \boldsymbol{d}_{t,l}, \tag{2}$$

where $G_{k,l}$ and $\boldsymbol{b}_l$ are an M-dimensional square matrix and a column vector, respectively. (2) indicates that noise-free reverberant speech spectral component vector $(\boldsymbol{y}_{t,l} - \boldsymbol{d}_{t,l})$ is the output of a multi-channel auto-regressive (AR) system driven by $\boldsymbol{b}_l s_{t,l}$. Note that $\boldsymbol{b}_l$ corresponds to the so-called steering vector in the reverberation free case. Hereafter, we call $G_{k,l}$ and $\boldsymbol{b}_l$ a room regression matrix and a steering vector, respectively.

(2) is equivalentlly rewritten as

$$\boldsymbol{y}_{t,l} = \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l} + \boldsymbol{x}_{t,l} \tag{3}$$

$$\boldsymbol{x}_{t,l} = \boldsymbol{b}_l s_{t,l} + \boldsymbol{v}_{t,l}, \tag{4}$$

where $\boldsymbol{v}_{t,l} = \boldsymbol{d}_{t,l} - \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{d}_{t-k,l}$. The set of (3) and (4) means that the observed speech, $\boldsymbol{y}_{t,l}$, is generated as follows. The clean speech, $s_{t,l}$, is first scaled by the steering vector, $\boldsymbol{b}_l$. Then, it is contaminated by filtered noise $\boldsymbol{v}_{t,l}$ to yield noisy speech $\boldsymbol{x}_{t,l}$. Finally, the noisy speech is reverberated via the multi-channel AR system, given by $G_l(z) = (I_M - \sum_{k=1}^{K_l} G_{k,l}^H z^{-k})^{-1}$, where I_M is the M-dimensional identity matrix. This equivalent generative system is depicted in Fig. 2. Note that if the original noise, $\boldsymbol{d}_{t,l}$, is stationary, the filtered noise, $\boldsymbol{v}_{t,l}$, also becomes stationary. Hereafter, we refer to the filtered noise simply as noise.

To make the estimation task solvable, we assume the following conditions, which have been widely accepted in the literature. Below, Θ denotes the set of all model parameters, which will be defined later (see (10)).

1. The short-time power spectral density (PSD) of a clean speech signal has an all-pole form of order P. Therefore, if we let ω be an angular frequency, the PSD at the t-th frame, denoted by ${}_s\lambda_t(\omega)$, is written as

$${}_s\lambda_t(\omega) = \frac{{}_s\sigma_t^2}{|A_t(e^{j\omega})|^2} \tag{5}$$

$$A_t(z) = 1 - \sum_{k=1}^{P} a_{t,k} z^{-k}, \tag{6}$$

   where $a_{t,k}$ and ${}_s\sigma_t^2$ are called a linear predictor coefficient (LPC) and a prediction residual power, respectively. We also collectively refer to $a_{t,k}$ and ${}_s\sigma_t^2$ as all-pole parameters.

2. The short-time PSDs and cross spectral densities (CSDs) of noise signals are time-invariant, or independent of frame index t. We represent the PSDs and CSDs together in a matrix as

$${}_v\Lambda(\omega) = \begin{bmatrix} {}_v\lambda^{(1,1)}(\omega) & \cdots & {}_v\lambda^{(1,M)}(\omega) \\ \vdots & \ddots & \vdots \\ {}_v\lambda^{(M,1)}(\omega) & \cdots & {}_v\lambda^{(M,M)}(\omega) \end{bmatrix}, \tag{7}$$

   where ${}_v\lambda^{(m,m)}(\omega)$ is the PSD of the m-th noise signal and ${}_v\lambda^{(m_1,m_2)}(\omega)$ is the CSD of the m_1-th and m_2-th noise signals.

3. $\boldsymbol{v}_{t_1,l_1}$ and $\boldsymbol{v}_{t_2,l_2}$ are statistically independent unless $(t_1, l_1) = (t_2, l_2)$.

4. Similarly, s_{t_1,l_1} and s_{t_2,l_2} are also independent unless $(t_1, l_1) = (t_2, l_2)$.

5. For any (t_1, l_1, t_2, l_2), s_{t_1,l_1} and $\boldsymbol{v}_{t_2,l_2}$ are independent.

6. Clean speech spectral component $s_{t,l}$ has a complex Gaussian distribution with mean 0 and variance ${}_s\lambda_t(2\pi l/L)$:

$$p(s_{t,l};\Theta) = \mathcal{N}_{\mathbb{C}}\{s_{t,l}; 0,\ {}_s\lambda_t(2\pi l/L)\}. \tag{8}$$

7. Noise spectral component vector $\boldsymbol{v}_{t,l}$ has a multivariate complex Gaussian distribution with mean $\mathbf{0} = [0, \cdots, 0]^T$ and covariance matrix ${}_v\Lambda(2\pi l/L)$. Hence, by letting ${}_v\Lambda_l = {}_v\Lambda(2\pi l/L)$, we have

$$p(\boldsymbol{v}_{t,l};\Theta) = \mathcal{N}_{\mathbb{C}}\{\boldsymbol{v}_{t,l}; \mathbf{0},\ {}_v\Lambda_l\}. \tag{9}$$

Now, parameter set Θ is specifically given by

$$\Theta = \{\ {}_g\Theta,\ {}_b\Theta,\ {}_s\Theta,\ {}_v\Theta\} \tag{10}$$
$${}_g\Theta = \{\{G_{k,l}\}_{1\le k\le K_l}\}_{0\le l\le L-1} \tag{11}$$
$${}_b\Theta = \{\boldsymbol{b}_l\}_{0\le l\le L-1} \tag{12}$$
$${}_s\Theta = \{a_{t,1}, \cdots, a_{t,P},\ {}_s\sigma_t^2\}_{0\le t\le T-1} \tag{13}$$
$${}_v\Theta = \{\ {}_v\Lambda_l\}_{0\le l\le L-1}. \tag{14}$$

${}_g\Theta$, ${}_b\Theta$, ${}_s\Theta$, and ${}_v\Theta$ are the sets of room regression matrices, steering vectors, speech all-pole parameters, and noise covariance matrices, respectively. For later use, we also denote the set consisting of the parameters other than the room regression matrices by ${}_{-g}\Theta$:

$${}_{-g}\Theta = \{\ {}_b\Theta,\ {}_s\Theta,\ {}_v\Theta\}. \tag{15}$$

3. PROPOSED ALGORITHM

3.1. MMSE estimation of clean speech

Now, we assume that model parameter set Θ is known in advance. In this case, noisy speech spectral component vector $\boldsymbol{x}_{t,l}$ is available by using (3). Then, the minimum mean square error (MMSE) estimate of clean speech spectral component $s_{t,l}$ is obtained by using the well-known multi-channel Wiener filtering. Indeed, the posterior probability density function (PDF) of the clean speech is represented as

$$p(s_{t,l}|\boldsymbol{x}_{t,l};\Theta) = \mathcal{N}_{\mathbb{C}}\{s_{t,l}; \mu_{t,l}(\boldsymbol{x}_{t,l};\Theta), \gamma_{t,l}(\Theta)\} \tag{16}$$
$$\mu_{t,l}(\boldsymbol{x}_{t,l};\Theta) = \frac{\boldsymbol{b}_l^T\ {}_v\Lambda_l^{-1}}{{}_s\lambda_{t,l}^{-1} + \boldsymbol{b}_l^T\ {}_v\Lambda_l^{-1}\ \boldsymbol{b}_l}\boldsymbol{x}_{t,l} \tag{17}$$
$$\gamma_{t,l}(\Theta) = (\ {}_s\lambda_{t,l}^{-1} + \boldsymbol{b}_l^T\ {}_v\Lambda_l^{-1}\ \boldsymbol{b}_l)^{-1}. \tag{18}$$

$\mu_{t,l}(\boldsymbol{x}_{t,l};\Theta)$ and $\gamma_{t,l}(\Theta)$ correspond to the MMSE estimate of $s_{t,l}$ and the associated mean squared error, respectively.

However, the model parameter set, Θ, is unseen in reality. Therefore, Θ must be estimated from observed data $\mathcal{Y} = \{\boldsymbol{y}_{t,l}\}_{0\le t\le T-1, 0\le l\le L-1}$. We describe the estimator of Θ in Sect. 3.2. Thus, the overall structure of the proposed speech enhancement method consists of a time-frequency analyzer, a model parameter estimator, an MMSE clean speech estimator, and a time-domain signal synthesizer.

3.2. ML parameter estimation

We use the maximum likelihood (ML) estimation method to estimate parameter set Θ. Based on the assumptions described in Sect. 2.2, we can derive the PDF of observed data $\mathcal{Y}$ as

$$p(\mathcal{Y};\Theta) \propto \prod_{l=0}^{L-1}\prod_{t=0}^{T-1} |\ {}_x\Lambda_{t,l}\ |^{-1} \exp\Big\{-\Big(\boldsymbol{y}_{t,l} - \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l}\Big)^H \times\ {}_x\Lambda_{t,l}^{-1}\Big(\boldsymbol{y}_{t,l} - \sum_{k=1}^{K_l} G_{k,l}^H \boldsymbol{y}_{t-k,l}\Big)\Big\}, \tag{19}$$

where ${}_x\Lambda_{t,l}$ is the covariance matrix of $\boldsymbol{x}_{t,l}$, which is given by

$${}_x\Lambda_{t,l} = {}_s\lambda_{t,l}\,\boldsymbol{b}_l\boldsymbol{b}_l^H + {}_v\Lambda_l\,. \tag{20}$$

The ML estimate of the parameter set is obtained as $\hat{\Theta}$ that maximizes the log likelihood function, defined as $\log p(\mathcal{Y};\Theta)$.

Since it is impossible to calculate the ML estimate $\hat{\Theta}$ analytically, we use the following expectation maximization (EM)-like iterative algorithm. Below, $\hat{\Theta}^{(i)}$ denotes the tentative estimate of Θ after the i-th iteration.

E-step: Calculate the following clean speech posterior PDF given tentative parameter estimate $\hat{\Theta}^{(i)}$:

$$q^{(i)}(\mathcal{S}) = p(\mathcal{S}|\mathcal{Y};\hat{\Theta}^{(i)}). \tag{21}$$

Now, let us define an auxiliary function, $q^{(i)}(\Theta)$, as

$$q^{(i)}(\Theta) = \int q^{(i)}(\mathcal{S}) \log p(\mathcal{Y},\mathcal{S}|\Theta) d\mathcal{S}. \tag{22}$$

CM-step1: Update the estimate of ${}_{-g}\Theta$ by maximizing *the auxiliary function* as

$${}_{-g}\hat{\Theta}^{(i+1)} = \underset{{}_{-g}\Theta}{\operatorname{argmax}}\, q^{(i)}(\ {}_g\hat{\Theta}^{(i)},\ {}_{-g}\Theta), \tag{23}$$

CM-step2: Update the estimate of ${}_g\Theta$ by maximizing *the log likelihood function* as

$${}_g\hat{\Theta}^{(i+1)} = \underset{{}_g\Theta}{\operatorname{argmax}} \log p(\mathcal{Y}|\ {}_g\Theta,\ {}_{-g}\hat{\Theta}^{(i+1)}). \tag{24}$$

We can readily prove that this algorithm ensures the monotonic increase and convergence of the log likelihood function.

All steps can be calculated analytically, although we omit the detailed formulas owing to the space limitation. E-step is performed by calculating $p(s_{t,l}|\boldsymbol{x}_{t,l};\hat{\Theta}^{(i)})$, given by (16), over all t and l. As regards CM-step1, the all-pole parameters in ${}_s\Theta$ are updated by using the Levinson-Durbin algorithm, and each steering vector in ${}_b\Theta$ is updated based on the cross correlation between the tentative estimates of $s_{t,l}$ and $\boldsymbol{x}_{t,l}$. Noise covariance matrices in ${}_v\Theta$ are updated from signals obtained during the first 0.3 seconds, where speech is assumed to be absent. The update formula for CM-step2 is described in [9].

4. EXPERIMENT AND CONCLUSION

We conducted an experiment for evaluating the performance of the proposed speech enhancement method. We selected Japanese utterances spoken by 10 speakers (five male and five female) from the

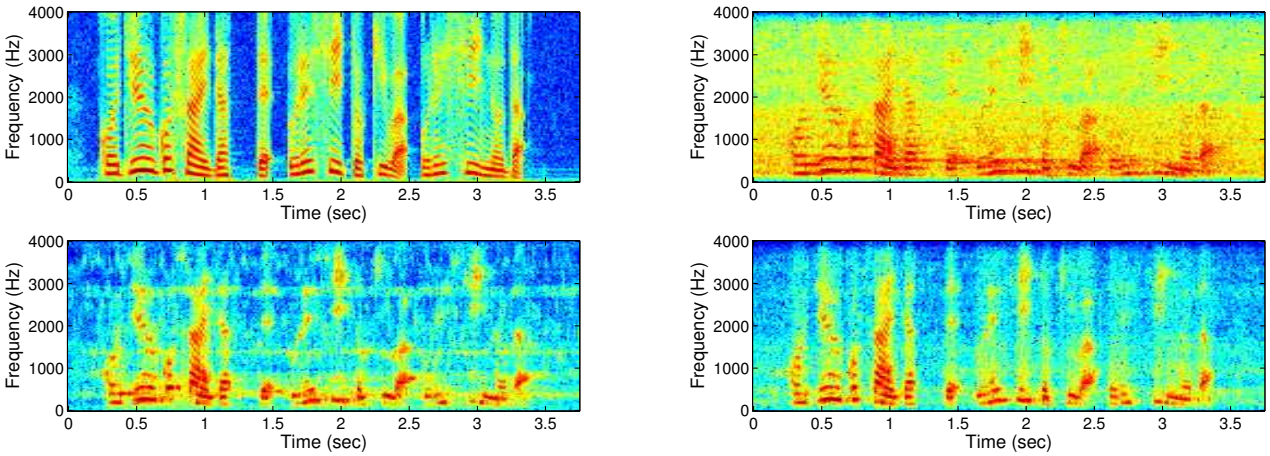

Fig. 3. Spectrograms of clean speech (top left), observed speech (top right), speech enhanced with the conventional method (bottom left), and speech enhanced with the proposed method (bottom right).

ASJ-JNAS database. To strike a balance between the controllability of the experimental conditions and the experimental reality, we played each utterance from a loudspeaker in a room and recorded the sound with two microphones. We also played uncorrelated pink noises simultaneously from four loudspeakers in the same room and recorded the sound with the same microphone setup. Then, the recorded noise was added to each recorded reverberant speech on a computer with a reverberant signal to noise ratio of 10 dB. The reverberation time of the room was around 0.6 seconds, and the distance between the loudspeaker and the microphone set was 1.8 meters. The waveforms were sampled at 8 kHz. The signal lengths ranged from 3.16 to 7.16 seconds.

The system parameters were as follows. The frame size and frame shift for the time-frequency analysis were 256 and 128 samples, respectively. The assumed number of speech poles, P, was set at 12. The room regression orders, K_l, were set at $K_l = 5$ for $f_l < 100$, $K_l = 10$ for $100 \leq f_l < 200$, $K_l = 30$ for $200 \leq f_l < 1000$, $K_l = 20$ for $1000 \leq f_l < 1500$, $K_l = 15$ for $1500 \leq f_l < 2000$, $K_l = 10$ for $2000 \leq f_l < 3000$, and $K_l = 5$ for $f_l \geq 3000$, where f_l is the center frequency in Hertz of the l-th frequency band. Note that setting the room regression orders at smaller values for higher frequency bands in this way helps us to save computing time with little performance degradation.

The average cepstral distances (CDs) for the observed speech, the speech enhanced with our recent method [7], and the speech enhanced with the proposed method were 7.39, 5.81, and 5.11, respectively. Fig. 3 shows example spectrograms, which indicate that the proposed method cancelled the effect of reverberation much better than the conventional method in several frequency bands. Indeed, the speech enhanced with the proposed method sounded less reverberant than the speech enhanced with the conventional method. These results show the superiority of the proposed method over the nonlinear-then-linear filtering approach.

Future work includes adaptive parameter estimation to cope with situations where speakers or microphones move around during observation. Another future research topic is the adaptive estimation of the noise PSDs and CSDs.

5. REFERENCES

[1] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[2] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2001, vol. VI, pp. 3701–3704.

[3] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Int'l Conf. Acoust. Speech, Signal Process.*, 2008, pp. 85–88.

[4] E. A. P. Habets, N. D. Gaubitch, and P. A. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4577–4580.

[5] A. Abramson, E. A. P. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation using GARCH modeling," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4565–4568.

[6] K. Kinoshita, T. Nakatani, M. Delcroix, and M. Miyoshi, "Multi-step linear prediction based speech dereverberation in noisy reverberant environment," in *Proc. Interspeech*, 2007, pp. 854–857.

[7] T. Yoshioka, T. Nakatani, T. Hikichi, and M. Miyoshi, "Maximum likelihood approach to speech enhancement for noisy reverberant signals," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4585–4588.

[8] M. Triki and D. T. M. Slock, "Robust delay-&-predict equalization for blind SIMO channel dereverberation," in *Proc. Joint Worksh. Hands-free Speech Com., Mic. Arrays*, 2008, pp. 248–251.

[9] T. Yoshioka, T. Nakatani, and M. Miyoshi, "An integrated method for blind separation and dereverberation of convolutive audio mixtures," in *Proc. Eur. Signal Process. Conf.*, 2008, accepted.