REAL-TIME SPEECH ENHANCEMENT IN NOISY REVERBERANT MULTI-TALKER ENVIRONMENTS BASED ON A LOCATION-INDEPENDENT ROOM ACOUSTICS MODEL

Tomohiro Nakatani[†] *Takuya Yoshioka*[†] *Keisuke Kinoshita*[†] *Masato Miyoshi*[†] *Biing-Hwang Juang*[†‡]

[†]NTT Communication Science Labs., NTT Corporation, Kyoto, Japan
[‡]School of ECE, Georgia Institute of Technology, GA, USA
{nak,takuya,kinoshita,miyo}@cslab.kecl.ntt.co.jp, juang@ece.gatech.edu

ABSTRACT

This paper describes a new real-time speech enhancement method that reduces signal distortion caused by stationary noise and late reflections of reverberation in speech signals captured by a single distant microphone under multi-talker conditions. A major problem here is how to estimate the energy of the late reflections in real time when the room impulse responses from individual talkers to the microphone are not given or fixed in advance. To solve this problem, we introduce a probabilistic room acoustics model, and provide a method for estimating the energy of late reflections based on this model. In this method, parameters of the model for a room can be fixed in advance only from a few seconds of observation. By incorporating the proposed approach into a conventional frequency domain noise reduction scheme, we realize an integrated real-time speech enhancement framework. The effectiveness of the proposed method is confirmed experimentally for a case where there are two talkers in a room.

Index Terms— Speech enhancement, Real time systems, Room acoustics, Reverberation, Autoregressive processes

1. INTRODUCTION

Speech signals captured by a distant microphone in an enclosed space will inevitably contain noise and reverberant components. These components have detrimental effects on the quality of the speech and seriously degrade many applications including hands-free telecommunication. Although numerous techniques have been proposed for suppressing these noise components, techniques for reducing reverberant components have not been well investigated particularly as regards real-time processing.

One approach to reducing reverberation is inverse filtering [1]. With this approach, the reverberation can be precisely canceled out by applying an inverse filter of the room impulse responses (RIRs) to the observed signal. However, because the RIRs depend strongly on such factors as the talker and microphone locations, we cannot adopt this approach when the talker location cannot be fixed precisely in advance. To overcome this problem, blind dereverberation has been proposed, in which the property of the room acoustics is estimated from the observed signal and used to reduce the reverberation [2, 3, 4, 5, 6]. In [4, 5, 6], the observation process is modeled by a long-term autoregressive (AR) process and dereverberation is accomplished by estimating the long-term AR coefficients from the observed signals. However, this technique requires a minimum of a few seconds of observation to obtain a reliable estimation of the long-term AR coefficients. In addition, when two or more talkers are in the room, the estimation will need to be controlled, for example, depending on "who speaks when", because the AR coefficients also depend strongly on the talker locations, and this will require a rather complex control mechanism. These problems preclude its use for real-time applications.

In contrast, the use of a prior knowledge of the room acoustics, represented by employing the prior probability density function (pdf) of the long-term AR coefficients, has recently been tested for incremental dereverberation using a single microphone and reported to be promising under single talker conditions [7]. In this paper, we extend this approach to real-time speech enhancement in more general environments, where an unknown number of talkers may speak alternately, or even simultaneously, under certain stationary background noise conditions. This means that the RIR convolved with the utterances may change discontinuously when the talkers alternate, and even that different RIRs may be convolved with simultaneous utterances and mixed into the observation. Because it is difficult to estimate the long-term AR coefficients for individual talkers precisely in this scenario, the proposed method uses the prior pdf of the coefficients as their location-independent estimates for real-time processing. The proposed method can be viewed as a variation of our previous incremental dereverberation method that can cope with multi-talker and stationary background noise conditions with a very little computing cost.

2. METHOD

Suppose there are one or more talkers in a room and their utterances are captured by a distant microphone. Let us denote a time-frequency point of the captured signal represented based on, for example, the short time Fourier transform (STFT) by $y_{t,k}$, where t and k are frame and frequency indices of the point, and denote that of the l-th unknown reverberant signal by $x^{(l)}_{t,k}$, which is derived from the l-th talker and included in the observed signal. Then, the observed signal can be modeled by

$$y_{t,k} = \sum_l x^{(l)}_{t,k} + n_{t,k}, \tag{1}$$

where $n_{t,k}$ denotes the stationary background noise. As discussed in [6], we assume that the reverberation process, by which each reverberant signal $x^{(l)}_{t,k}$ is generated from the l-th talker's utterance, can be modeled by a long-term AR process in each frequency bin as

$$x^{(l)}_{t,k} = (\mathbf{c}^{(l)}_k)^H \mathbf{x}^{(l)}_{t-d,k} + \tilde{s}^{(l)}_{t,k}, \tag{2}$$

where H denotes the conjugate transposition of a matrix, $\mathbf{c}^{(l)}_k$ and $\mathbf{x}^{(l)}_{t-d,k}$ are vectors of length T that contain the long-term AR coefficients and a past reverberant signal sequence preceding a time frame

$t-d$, respectively, defined as

$$\mathbf{c}_k^{(l)} = [c_{1,k}^{(l)}, c_{2,k}^{(l)}, \ldots, c_{T,k}^{(l)}]^{\mathcal{T}},$$
$$\mathbf{x}_{t-d,k}^{(l)} = [x_{t-d-1,k}^{(l)}, x_{t-d-2,k}^{(l)}, \ldots, x_{t-d-T,k}^{(l)}]^{\mathcal{T}},$$

where $\mathcal{T}$ denotes the non-conjugate transposition of a matrix. In (2), the reverberant signal $x_{t,k}^{(l)}$ at t is predicted from $\mathbf{x}_{t-d,k}^{(l)}$ by convolving it with $\mathbf{c}_k^{(l)}$, and the prediction residual, $\tilde{s}_{t,k}^{(l)}$, is taken as the speech signal to be enhanced. Because one goal of this paper is to reduce the late reflections of the reverberation, d is introduced as a time delay into the past reverberant signal, $\mathbf{x}_{t-d,k}^{(l)}$, in the first term of (2) [4]. This delay means that the early reflections of the reverberation cannot be predicted from the first term, and thus remain in the residual. To clarify this, the residual is denoted as $\tilde{s}_{t,k}^{(l)}$.

According to (2), $\mathbf{c}_k^{(l)}$ can be viewed as a linear filter[1] that predicts the late reflections, $r_{t,k}^{(l)}$, as

$$r_{t,k}^{(l)} = (\mathbf{c}_k^{(l)})^H \mathbf{x}_{t-d,k}^{(l)}. \quad (3)$$

2.1. Problem definition

From (2) and (3), the observation process (1) can be rewritten as

$$y_{t,k} = \tilde{s}_{t,k} + \tilde{n}_{t,k}, \quad (4)$$
$$\tilde{s}_{t,k} = \sum_l \tilde{s}_{t,k}^{(l)}, \quad (5)$$
$$\tilde{n}_{t,k} = \sum_l r_{t,k}^{(l)} + n_{t,k}, \quad (6)$$

where $\tilde{s}_{t,k}$ is the mixture of the speech signals that remains after subtracting the stationary background noise $n_{t,k}$ and the late reflections of the reverberation $r_{t,k}^{(l)}$ of all the speech signals from the observed signal $y_{t,k}$.

Throughout this paper, $\tilde{s}_{t,k}$ in (4) is taken as the desired signal to be obtained when the observed signal, $\psi_{t,k} = \{y_{t',k}\}_{t' \le t}$, is given. In contrast, $\tilde{n}_{t,k}$ in (4) is taken as nonstationary additive noise to be reduced. The proposed method estimates $\tilde{s}_{t,k}^{(l)}$ by reducing the energy of $\tilde{n}_{t,k}$ from that of $y_{t,k}$ based on certain noise reduction techniques, such as spectral subtraction and Wiener filtering. Therefore, the proposed method needs to estimate the power spectral density (psd) of the nonstationary noise. For this purpose, we introduce the following assumptions.

1. Of the nonstationary noise $\tilde{n}_{t,k}$ in (6), the stationary background noise $n_{t,k}$ and the late reflections $r_{t,k}^{(l)}$ from individual speech signals are mutually uncorrelated, and thus the psd of $\tilde{n}_{t,k}$ is equal to the sum of the psds of $n_{t,k}$ and $r_{t,k}^{(l)}$ for all l.
2. The psd of the stationary background noise, $n_{t,k}$, is given in advance, or can be estimated as N_k from the observed signal during non-speech periods.
3. The prior pdf of the long-term AR coefficients, $p(\mathbf{c}_k^{(l)})$, in a room is given or can be estimated in advance from the observed signal during certain speech periods.

In the following, we first discuss a way of preparing/estimating the prior pdf $p(\mathbf{c}_k^{(l)})$ from the observed signal in advance, and then explain how we can achieve the speech enhancement based on the above assumptions.

[1] Although in general the late reverberation is not precisely predicted by a single channel causal linear filter, it is empirically confirmed that this modeling error can be mitigated when using a time-frequency representation.

2.2. Prior pdf of long-term AR coefficients

We first assume that an RIR from a talker to a microphone is a random variable that depends on, for example, the talker location. Then, the long-term AR coefficients can be viewed as random variables of a room with uncertainty derived from that of the RIR and modeling errors in (2). Thus, the long-term AR coefficients $\mathbf{c}_k^{(l)}$ for the l-th talker location are taken as realizations of the random variables.

We adopt the following as the model for the prior pdf of the long-term AR coefficients.

$$p_{c_k}(\mathbf{c}_k^{(l)}) = \mathcal{N}(\mathbf{c}_k^{(l)}; \mu_{c_k}, \Sigma_{c_k}), \quad (7)$$

where $\mathcal{N}(\mathbf{a}; \mu, \Sigma)$ denotes the pdf of a multivariate complex Gaussian random variable $\mathbf{a}$ with a mean μ and a covariance matrix Σ.

In this paper, we test a way of defining the prior pdf, in which the parameters of the pdf are defined mainly on relatively location-independent features of the room acoustics. First, we assume that the mean of each AR coefficient is zero, namely $\mu_{c_k} = 0$, because the phase of an RIR varies greatly over different talker locations. Next, we assume that the long-term AR coefficients are mutually uncorrelated, and thus $\Sigma_{c_k} = E\{\mathbf{c}_k \mathbf{c}_k^H\}$ becomes diagonal, namely

$$\Sigma_{c_k} = \mathrm{diag}([\gamma_{1,k}^2, \ldots, \gamma_{T,k}^2]), \quad (8)$$

where $\gamma_{t,k}^2 = E\{|c_{t,k}|^2\}$, and $\mathrm{diag}(\mathbf{a})$ is a diagonal matrix that contains elements of a vector $\mathbf{a}$ as its diagonal components. In other words, the above prior pdf is characterized solely by the temporal power envelope of the long-term AR coefficients. Because our preliminary experiments showed that the temporal power envelope of the long-term AR coefficients in a room is relatively insensitive to differences in microphone and talker locations, we assume that the above pdf may be used as a location-independent pdf for the long-term AR coefficients in a room. In this paper, Σ_{c_k} is assumed to be determined by collecting a few seconds of observed signals in a room, by deriving long-term AR coefficients from the signals based on existing speech dereverberation algorithms such as [6], and by obtaining Σ_{c_k} based on (8).

It is important to note that there are several alternatives for determining μ_{c_k} and Σ_{c_k} in (7). For example, they can be determined from the posterior pdf of the long-term AR coefficients given a certain observed signals in a room, which can be obtained based on existing dereverberation algorithms [7]. Comparison of such alternatives should be included in future work.

2.3. Speech enhancement based on prior pdf of AR coefficients

Let us first rewrite the conditional expectation of the late reflection psd given the observed signal, $\psi_{t,k} = \{y_{t',k}\}_{t' \le t}$, as

$$E\{|r_{t,k}^{(l)}|^2 | \psi_{t,k}\} = |\bar{r}_{t,k}^{(l)}|^2 + R_{t,k}^{(l)}, \quad (9)$$

where $\bar{r}_{t,k}^{(l)} = E\{r_{t,k}^{(l)} | \psi_{t,k}\}$ and $R_{t,k}^{(l)} = E\{(r_{t,k}^{(l)} - \bar{r}_{t,k}^{(l)})(r_{t,k}^{(l)} - \bar{r}_{t,k}^{(l)})^H | \psi_{t,k}\}$, are the conditional mean and variance of $r_{t,k}^{(l)}$. According to (3), $\bar{r}_{t,k}^{(l)}$ and $R_{t,k}^{(l)}$ can further be rewritten using the unknown reverberant signal, $x_{t,k}^{(l)}$, as

$$\bar{r}_{t,k}^{(l)} = (\bar{\mathbf{c}}_k^{(l)})^H \mathbf{x}_{t-d,k}^{(l)}, \quad (10)$$
$$R_{t,k}^{(l)} = (\mathbf{x}_{t-d,k}^{(l)})^H C_k^{(l)} \mathbf{x}_{t-d,k}^{(l)}, \quad (11)$$

where $\bar{\mathbf{c}}_k^{(l)} = E\{\mathbf{c}_k^{(l)} | \psi_{t,k}\}$ and $C_k{}^{(l)} = E\{(\mathbf{c}_k^{(l)} - \bar{\mathbf{c}}_k^{(l)})(\mathbf{c}_k^{(l)} - \bar{\mathbf{c}}_k^{(l)})^H | \psi_{t,k}\}$ are, respectively, the mean and the covariance matrix of the posterior pdf of the long-term AR coefficients, $p(\mathbf{c}_k^{(l)} | \psi_{t,k})$.

With the proposed method, we do not track any states of the individual talkers in the observed signal for the sake of computational simplicity, and thus the posterior pdf of the long-term AR coefficients for the l-th talker is not specifically obtained in the estimation procedure. In addition, the past observed signal does not necessarily contain information on the long-term AR coefficients of the current observed signal because the talkers may alternate at any time. Instead, we consider the prior pdf $p(\mathbf{c}_k^{(l)})$ to be a location-independent estimate of the posterior pdf $p(\mathbf{c}_k^{(l)}|\psi_{t,k})$, and we substitute the mean and covariance matrix of $p(\mathbf{c}_k^{(l)})$ for those of $p(\mathbf{c}_k^{(l)}|\psi_{t,k})$. Then, (10) and (11) can be rewritten using (7) and (8) as

$$\begin{aligned} \bar{r}_{t,k}^{(l)} &= 0, \\ R_{t,k}^{(l)} &= \sum_{t'=1}^{T} \gamma_{t',k}^2 |x_{t-t'-d,k}^{(l)}|^2. \end{aligned}$$

Finally, according to (6), the psd of the nonstationary noise $\tilde{n}_{t,k}$ in (4) can be estimated as

$$E\{|\tilde{n}_{t,k}|^2|\psi_{t,k}\} = \sum_{t'=1}^{T} \gamma_{t',k}^2 X_{t-t'-d,k} + N_k, \qquad (12)$$

where $X_{t,k} = \sum_l |x_{t,k}^{(l)}|^2$ is the power sum of the reverberant signals, $x_{t,k}^{(l)}$, for all l, and this can be estimated by subtracting the psd of the stationary background noise, N_k, from that of the observed signal, $|y_{t,k}|^2$. Note that the first term on the right hand side of (12) is the estimated psd of the sum of the late reflections from all the speech signals.

Once the psd of $\tilde{n}_{t,k}$ was obtained, the denoised and dereverberated signal, $\hat{s}_{t,k}$, could be obtained by using a simple spectral subtraction technique, which is defined, for example, as

$$\hat{s}_{t,k} = G_{t,k} y_{t,k}, \qquad (13)$$

$$G_{t,k} = \max\left\{\alpha, \frac{|y_{t,k}| - \beta\sqrt{E\{|\tilde{n}_{t,k}|^2|\psi_{t,k}\}}}{|y_{t,k}|}\right\}, \qquad (14)$$

where α and β are positive constants that control the flooring effect of the spectral subtraction. We adopted $\alpha = 0.05$ and $\beta = 0.95$ for all the experiments described in this paper.

2.4. Processing flow of proposed method

An example processing flow of the proposed method can be summarized as follows.

1. Execute the following in advance.
   (a) The stationary noise psd, N_k, is estimated during non-speech periods.
   (b) The prior pdf, $p(\mathbf{c}_k)$, is estimated using an existing dereverberation method [6] from more than a few seconds of a speech signal uttered at an arbitrary talker location.
2. Execute the following for a captured signal $y_{t,k}$ at each frame.
   (a) N_k is subtracted from $|y_{t,k}|^2$ to obtain $X_{t,k}$.
   (b) $E\{|\tilde{n}_{t,k}|^2|\psi_{t,k}\}$ is estimated as (12).
   (c) $\hat{s}_{t,k}$ is estimated by (13) and (14).

Clearly, the above step 2 can be implemented as real-time processing with a rather small processing delay once step 1 has been completed in advance.

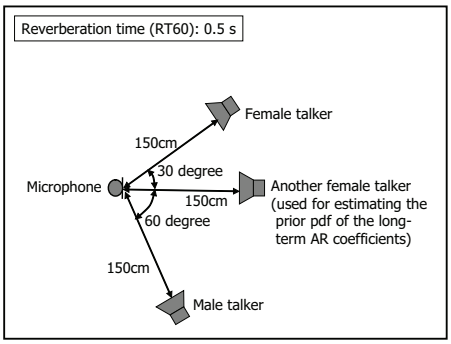

Fig. 1. Assumed recording conditions

3. EXPERIMENTS

We evaluated the proposed method, hereafter referred to as NR+RR (NR and RR stand for Noise Reduction and Reverberation Reduction, respectively), in comparison with three speech enhancement methods: 1) the noise reduction pre-processor used in the proposed method, referred to as NR, 2) the dereverberation method, which is proposed in [6] and performed with batch processing, referred to as NR+Batch, and 3) the dereverberation method, which is proposed in [7] and performed with incremental processing using prior information on the room acoustics, referred to as NR+aRR (aRR stands for adaptive Reverberation Reduction). In the experiments, the noise reduction pre-processing was also performed for NR+Batch and NR+aRR. With NR+Batch, the long-term AR coefficients were estimated after an entire observed signal was obtained and the entire signal was dereverberated using these coefficients. In contrast, with NR+aRR, the long-term AR coefficients were updated adaptively to the observed signal at each time frame as described in [7] and dereverberation was performed incrementally using the updated AR coefficients[2]. The long-term AR coefficients obtained by NR+Batch and NR+aRR were used to estimate the late reflections in the observed signal based on linear filtering, and spectral subtraction was performed to reduce the energy of the late reflections from the observed signal. Spectral subtraction based dereverberation was adopted here because it performed better under multi-talker conditions with a single microphone than the linear filtering based dereverberation that was originally used for NR+Batch and NR+aRR.

To test the effectiveness of each method, we first prepared five utterances by two talkers (a male and a female, a total of ten utterances). Each utterance was composed of a five-word sequence, where each word was extracted from the ATR word utterance database. Then, the male and female utterances were, respectively, convolved with different 1-ch RIRs measured in a reverberant room with a reverberation time (RT60) of 0.5 sec as shown in Fig. 1, and mixed according to two different settings, referred to as AT and ST. (AT and ST stand for Alternate Talkers and Simultaneous Talkers, respectively.) The male and female utterances appeared alternately in the mixture with AT and simultaneously with ST. Stationary white Gaussian noise was added to each mixture with different reverberant signal to noise power ratios (RSNR), namely ∞, 30, 20, and 10 dB. As the results, 40 observed signals (5 utterance pairs $\times$ 2 mixture settings $\times$ 4 RSNR conditions) were prepared.

[2]Note that NR+aRR requires a much greater computing cost for its adaptive processing than NR+RR.

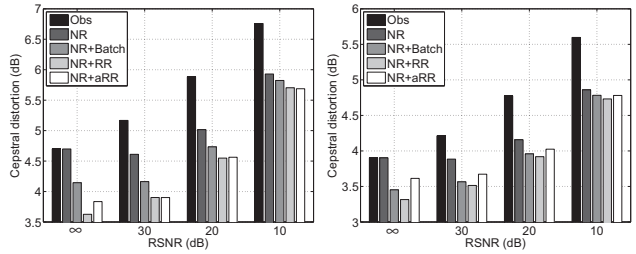

Fig. 2. Average cepstral distortions of the observed signals (Obs) and signals obtained using NR, NR+Batch, NR+RR (=proposed method), and NR+aRR with AT (left panel) and ST (right panel) with different reverberant signal to noise power ratios (RSNR).

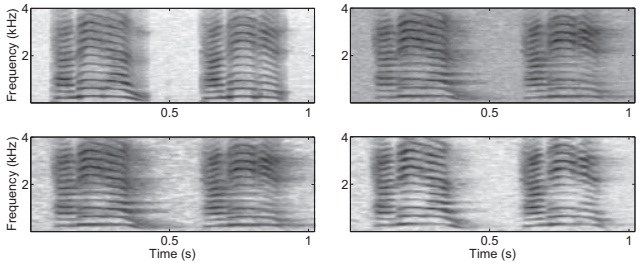

Fig. 3. Example spectrograms of clean (top left), and observed (top right) signals with AT, and signals processed by NR (bottom left), and NR+RR (bottom right) when the RSNR was set at 30 dB. Only two words from a female utterance are shown in the figure.

Dereverberation was performed for each observed signal, and the performance was evaluated in terms of the cepstral distortion (CD) of the recovered signals. CD in dB is defined as

$$\mathrm{CD} = (10/\ln 10)\sqrt{(\hat{\beta}_0 - \beta_0)^2 + 2\sum_{k=1}^{D}(\hat{\beta}_k - \beta_k)^2},$$

where $\hat{\beta}_k$ and β_k are, respectively, the cepstral coefficients of the signal being evaluated and those of the desired signal, and we adopted $D = 12$. The desired signals were set as mixtures of the original clean speech signals because the speech enhancement in this paper only reduces the background noise and the late reflections. In addition, to reduce the effect of the early reflections that remain in the dereverberated signals, we applied cepstral mean normalization to all signals before calculating the CDs. Distortions in the energy time pattern and spectral envelope were evaluated with this measure. The sampling rate was set at 8 kHz. For the time-frequency analysis, we adopted a complex subband filter representation [8], and set the number of subbands and the oversampling rate at 128 and 2, respectively. The order of the long-term AR process was set at 24 for each subband. We set $d = 1$ in (2). To determine the prior pdf of the long-term AR coefficients, or Σ_{c_k}, we first applied NR+Batch to a different female utterance convolved with an RIR that was measured for a different talker location in the same room (see Fig. 1).

Figure 2 shows the average CDs of the observed signals (Obs), and those of the signals obtained using NR, NR+Batch, the proposed method (NR+RR), and NR+aRR under different RSNR conditions. The figure shows that the proposed method (NR+RR) was the best at reducing the average CDs under almost all the RSNR conditions. Although NR+Batch and NR+aRR were consistently better than NR, they never outperformed the proposed method (NR+RR). We presume that this is because these two methods model the reverberation process by a single channel (quasi-)stationary long-term AR process that does not fit the multi-talker conditions well. In contrast, it is interesting to see that NR+aRR achieved almost the same performance as the proposed method (NR+RR) with AT when RSNR was 30 dB or lower. This suggests that the adaptive estimation of the long-term AR coefficients by NR+aRR tracked speaker alternation fairly well.

Figure 3 shows example spectrograms of speech obtained before and after speech enhancement with AT. They clearly demonstrate that the proposed method (NR+RR) reduced the energy of both background noise and reverberation effectively. To demonstrate the audible quality of the processed signals, we also prepared a web site containing sound examples obtained in the experiments [9].

4. CONCLUSION

This paper proposed a new method for enhancing speech signals captured in noisy reverberant multi-talker environments by a distant microphone. Specifically, the proposed method provides us with a methodology for reducing the late reflection energy of reverberation included in the observed signal by exploiting a prior knowledge of room acoustics, represented by employing the prior pdf of the long-term AR coefficients, no matter how many talkers are in the room. Once the psd of the stationary background noise and the prior pdf of the long-term AR coefficients are obtained based on a few seconds of observation, the proposed method can perform speech enhancement in real time based on a conventional noise suppression scheme. Our experiments showed the effectiveness of the proposed method under conditions in which two talkers speak alternately and simultaneously.

5. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, vol. 36, no. 2, pp. 145–152, 1988.

[2] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[3] E.A.P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," *Proc. ICASSP-2005*, vol. IV, pp. 173–4580, 2005.

[4] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," *Proc. ICASSP-2006*, vol. I, pp. 817–820, 2006.

[5] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Dereverberation by using time-variant nature of speech production system," *EURASIP Journal on Advances in Signal Process.*, vol. 2007, Article ID 65698, 15 pages doi:10.1155/2007/65698, 2007.

[6] 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," *Proc. ICASSP-2008*, pp. 85–88, 2008.

[7] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.H. Juang, "Incremental estimation of reverberation with uncertainty using prior knowledge of room acoustics for speech dereverberation," *Proc. IWAENC-2008*, pp. 85–88, 2008.

[8] S. Weiss and R.W. Stewart, "Fast implementation of oversampled modulated filter banks," *IEE Electronics Letters*, vol. 36, no. 17, pp. 1502–1503, 2000.

[9] "http://www.kecl.ntt.co.jp/icl/signal/nakatani/sound-demos/dm3/derev-demos3.html," .