INCREMENTAL ESTIMATION OF REVERBERATION WITH UNCERTAINTY USING PRIOR KNOWLEDGE OF ROOM ACOUSTICS FOR SPEECH DEREVERBERATION

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 proposes a new dereverberation method that works with incremental processing. A major problem here is how to estimate parameters of the observation process reliably when only a very short observation is available, for example, at the beginning of each human utterance. For this purpose, the prior knowledge of the room acoustics is incorporated into the proposed method by emplying the prior probability density function representation. The proposed method integrates the prior knowledge and the information obtained from the observed signal based on the Bayes' rule, and achieves incremental dereverberation effectively.

Index Terms— Dereverberation, Room acoustics, Statistical signal processing, Incremental estimation, Speech enhancement

1. INTRODUCTION

Speech signals captured by distant microphones in an enclosed space will inevitably contain reverberant components because of reflections from the walls, the floor or the ceiling. These reverberant components have a detrimental effect on the quality of the signal and seriously degrade many applications including automatic speech recognition.

One way to overcome this problem is speech dereverberation, where the goal is to estimate parameters of unknown observation process, and to recover the original quality of the speech signals based on the estimated parameters [1, 2, 3]. It has been shown that the maximum likelihood (ML) estimation approach with a model of time-varying speech characteristics and that of the room acoustics represented by long-term autoregressive (AR) coefficients is promising for achieving effective dereverberation based only on a few seconds of observation [1]. In addition, a method has been proposed for implementing this speech dereverberation approach in the STFT domain in a computationally very efficient manner [4].

Although effective and efficient dereverberation methods have been developed as noted above, they only work with batch processing. They require the entire observed signal (or at least more than 1 sec observation) to be available in advance if we are to estimate the parameters of the observation process effectively. The dereverberation process may degrade the quality of a short observed signal. Therefore, we cannot use such methods for real time applications, for which the dereverberation needs to work incrementally from the beginning of each human utterance with very little algorithmic delay.

The goal of this paper is to develop a new dereverberation method that works with incremental processing. Here we assume that the beginning of each human utterance can be detected by a technique such as voice activity detection [5]. Then, the main problem of this task is how to estimate the parameters of the observation process reliably when only a very short observation is available. For this purpose, we propose a way of incorporating the prior probability density function (pdf) of the AR coefficients that represents the room acoustics into the proposed method. The prior pdf enables the proposed method to estimate the AR coefficients by means of the posterior pdf reliably even with little observation, and thus to achieve effective dereverberation with incremental processing. Note that such a prior pdf can be blindly obtained from only a few seconds of observation using existing dereverberation techniques, and thus we do not need to measure room impulse responses (RIR) in advance for the proposed method.

2. METHOD

Suppose a single speech signal is captured by a distant microphone, where the speech signal is known to be active from a certain starting time. Then, as discussed in [4], the observation process can be modeled in the short time Fourier transform (STFT) domain by separate long-term AR processes in individual frequency bins as[1]

$$x_{t,k} = \mathbf{c}_k^H \mathbf{x}_{t-1,k} + s_{t,k} \tag{1}$$

where t and k are frame and frequency indices of an STFT, H denotes the conjugate transposition of a matrix, a vector, or a scalar, and $x_{t,k}$ and $s_{t,k}$ are frequency bins of complex spectra corresponding to the STFTs of the observed and clean speech signals, respectively. $\mathbf{c}_k$ and $\mathbf{x}_{t-1,k}$ are vectors of length T that contain the long-term AR coefficients of the room acoustics and a past observed signal sequence preceding a frame t, respectively, defined as

$$\begin{aligned} \mathbf{c}_k &= [c_{1,k}^H, c_{2,k}^H, \ldots, c_{T,k}^H]^H, \\ \mathbf{x}_{t-1,k} &= [x_{t-1,k}^H, x_{t-2,k}^H, \ldots, x_{t-T,k}^H]^H. \end{aligned}$$

In (1), $r_{t,k} = \mathbf{c}_k^H \mathbf{x}_{t-1,k}$ represents an STFT of the reverberant component included in $x_{t,k}$.

2.1. Basic dereverberation scheme

In our approach, we assume $x_{t,k}$, $s_{t,k}$, $r_{t,k}$ and $c_{t,k}$ to be realizations of their respective random variables. Hereafter, we suppose that a time series of the observed signal, denoted by $\xi_{\tau,k} = \{x_{t,k}\}_{t\in\tau}$, at a frequency bin k within a certain time duration τ is available at an estimation step in the incremental processing. Then, as a subgoal of the dereverberation, our method first estimates the posterior

[1]Although the inversion of a single channel RIR is not precisely represented by a causal linear filter in general, an STFT representation is empirically confirmed to mitigate this modeling error. In addition, the discussion in this paper can be easily extended to multi-channel cases.

probability density function (pdf) of $\mathbf{c}_k$, denoted by $p_{c_k}(\mathbf{c}_k|\xi_{\tau,k})$, for all k given the observed signal. With this pdf, we can decompose the reverberation $r_{t,k}$ into two parts, namely its conditional mean given $\xi_{\tau,k}$ and the deviation from the mean, respectively denoted by $\bar{r}_{t,k}$ and $e_{t,k}$, as

$$r_{t,k} = \bar{r}_{t,k} + e_{t,k}, \quad (2)$$

$$\bar{r}_{t,k} = \bar{\mathbf{c}}_k^H \mathbf{x}_{t-1,k}, \quad (3)$$

where $\bar{\mathbf{c}}_k = E\{\mathbf{c}_k|\xi_{\tau,k}\}$ corresponds to the mean of $p_{c_k}(\mathbf{c}_k|\xi_{\tau,k})$. We can also derive the conditional variance of $e_{t,k}$ given $\xi_{\tau,k}$ as

$$E\{|e_{t,k}|^2|\xi_{\tau,k}\} = E\{|\bar{r}_{t,k} - r_{t,k}|^2|\xi_{\tau,k}\}, \quad (4)$$

$$= \mathbf{x}_{t-1,k}^H C_k \mathbf{x}_{t-1,k}, \quad (5)$$

where $C_k = E\{(\mathbf{c}_k - \bar{\mathbf{c}}_k)(\mathbf{c}_k - \bar{\mathbf{c}}_k)^H|\xi_{\tau,k}\}$ corresponds to the covariance matrix of $p_{c_k}(\mathbf{c}_k|\xi_{\tau,k})$.

The clean speech signal is then estimated based on the posterior pdf of $\mathbf{c}_k$ and (1) as follows. First, letting $\tilde{x}_{t,k} = x_{t,k} - \bar{r}_{t,k}$, the observation process in (1) can be rewritten as

$$\tilde{x}_{t,k} = s_{t,k} + e_{t,k}. \quad (6)$$

In (6), $e_{t,k}$ behaves like additive random noise with a mean of zero and a covariance matrix $E\{|e_{t,k}|^2|\xi_{\tau,k}\}$. Therefore, we can estimate the clean speech power spectrum based on (6) by subtracting the power spectrum of $e_{t,k}$ estimated as $E\{|e_{t,k}|^2|\xi_{\tau,k}\}$ from that of $\tilde{x}_{t,k}$ based on a technique such as spectral subtraction. For synthesizing the speech signal by means of the waveform, the use of overlap-add synthesis and substituting the phase of $\tilde{x}_{t,k}$ for that of the dereverberated signal was shown to be effective experimentally.

In the following, we discuss how to estimate the conditional expectation values, $\bar{\mathbf{c}}_k$ in (3) and C_k in (5), by introducing prior pdfs of the speech and room acoustics.

2.2. Posterior pdf estimation using prior information

We estimate the conditional expectation values, $\bar{\mathbf{c}}_k$ and C_k, as the mean and covariance of the posterior pdf of $\mathbf{c}_k$, $p_{c_k}(\mathbf{c}_k|\xi_{\tau,k})$. The pdf can be rewritten as

$$p_{c_k}(\mathbf{c}_k|\xi_{\tau,k}) = \frac{p_{c_k}(\mathbf{c}_k) p_{\xi_{\tau,k}}(\xi_{\tau,k}|\mathbf{c}_k)}{\int p_{c_k}(\mathbf{c}_k) p_{\xi_{\tau,k}}(\xi_{\tau,k}|\mathbf{c}_k) d\mathbf{c}_k}, \quad (7)$$

where $p_{c_k}(\mathbf{c}_k)$ is the prior pdf of $\mathbf{c}_k$, and according to (1), we can further rewrite $p_{\xi_{\tau,k}}(\xi_{\tau,k}|\mathbf{c}_k)$ in the above equation as

$$p_{\xi_{\tau,k}}(\xi_{\tau,k}|\mathbf{c}_k) = \prod_{t\in\tau} p_{x_{t,k}}(x_{t,k}|\mathbf{x}_{t-1,k}, \mathbf{c}_k),$$

$$= \prod_{t\in\tau} p_{s_{t,k}}(s_{t,k} = x_{t,k} - \mathbf{c}_k{}^H \mathbf{x}_{t-1,k}), \quad (8)$$

where $p_{s_{t,k}}(s_{t,k})$ is the prior pdf of the speech signal at a frame t. Based on (7) and (8), the posterior pdf $p_{c_k}(\mathbf{c}_k|\xi_{\tau,k})$ can be determined when $p_{c_k}(\mathbf{c}_k)$ and $p_{s_{t,k}}(s_{t,k})$ are given.

We introduce definitions of the two prior pdfs in the following.

2.2.1. Prior pdf of room acoustics

We first assume that an RIR from the speaker to the microphone can be viewed as a random variable that depends on, for example, the speaker location and room temperature. Then, we assume that the long-term AR coefficients can also be viewed as random variables that have uncertainty derived from that of the RIR and the system modeling errors in (1).

In this paper, the prior pdf of the long-term AR coefficients is defined as

$$p_{c_k}(\mathbf{c}_k) = \mathcal{N}(\mathbf{c}_k; 0, \Sigma_{c_k}), \quad (9)$$

where $\mathcal{N}(\mathbf{a}; \mu, \Sigma)$ denotes the pdf of a multivariate complex Gaussian process $\mathbf{a}$ with a mean μ and a covariance matrix Σ. Because the phase of an RIR varies greatly over different speaker and microphone locations, we assume the mean of each AR coefficient to be zero in the above pdf. On the other hand, we assume that Σ_{c_k} can be determined or estimated in advance as $\Sigma_{c_k} = E\{\mathbf{c}_k \mathbf{c}_k{}^H\}$. For the sake of simplicity, we further assume that Σ_{c_k} is diagonal, namely

$$\Sigma_{c_k} = \mathrm{diag}([\gamma_{1,k}^2, \ldots, \gamma_{T,k}^2]), \quad (10)$$

where $\gamma_{t,k}^2 = E\{|c_{t,k}|^2\}$, $\mathrm{diag}(\mathbf{a})$ 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 power-time envelope of the long-term AR coefficients. Because our preliminary experiments showed that the power-time envelope of the long-term AR coefficients in a room is relatively insensitive to differences in microphone and speaker locations, we assume that the above pdf can be used as a general pdf for the long-term AR coefficients in a room.

We can determine Σ_{c_k} in many ways. For example, we can measure several RIRs in a room with different measurement settings, derive the long-term AR coefficients corresponding to respective settings, and calculate Σ_{c_k} according to (10). Σ_{c_k} can also be determined simply by collecting certain sets of observed signals and by deriving long-term AR coefficients from each of them based on existing speech dereverberation algorithms such as [4]. As an alternative approach, Σ_{c_k} may be represented by a parametric model as used in [2], and the envelope can be determined simply by controlling the reverberation time.

2.2.2. Prior pdfs of speech signals

We adopt the time-varying multivariate Gaussian source model (TVGSM) as the prior pdfs of speech signals because it has been shown to be very effective for a conventional speech dereverberation method based on the long-term AR observation model [4]. With TVGSM, the speech signal is modeled as

$$p_{s_t}(\mathbf{s}_t) = \mathcal{N}(\mathbf{s}_t; 0, \Sigma_{s_t}), \quad (11)$$

where $\mathbf{s}_t = [(s_{t,1})^H, \ldots, (s_{t,K})^H]^H$ is a vector that contains frequency bins of the clean speech over all frequency bins at a frame t, and $\Sigma_{s_t} = E\{\mathbf{s}_t \mathbf{s}_t^H\}$. We further assume Σ_{s_t} is diagonal, namely,

$$\Sigma_{s_t} = \mathrm{diag}([\sigma_{t,1}^2, \ldots, \sigma_{t,K}^2]), \quad (12)$$

where $\sigma_{t,k}^2 = E\{|s_{t,k}|^2\}$. In other words, the above prior pdf is characterized solely by the power spectrum of the clean speech signal.

Because the power spectrum of the clean speech signal is not given in advance, we need to approximate it in some way with this approach. One way to do this is to adopt the power spectrum of the observed signal as the approximated value. This approach has also been shown to be very effective with the conventional method. With this approach, we may be able to refine the approximation further based on an iterative estimation scheme by determining the power spectrum based on the posterior pdf of the speech signal estimated in the preceding estimation steps based on the posterior pdf of the long-term AR coefficients.

2.2.3. Solution

Based on (8), (9), and (11), it is easily shown that (7) becomes a multivariate complex Gaussian pdf as

$$p_{c_k}(\mathbf{c}_k|\xi_{\tau,k}) = \mathcal{N}(\mathbf{c}_k;\bar{\mathbf{c}}_k, C_k), \quad (13)$$

$$\bar{\mathbf{c}}_k = C_k \sum_{t\in\tau} \frac{\mathbf{x}_{t-1,k} x_{t,k}^H}{\sigma_{t,k}^2}, \quad (14)$$

$$C_k = \left(\sum_{t\in\tau} \frac{\mathbf{x}_{t-1,k}\mathbf{x}_{t-1,k}^H}{\sigma_{t,k}^2} + \Sigma_{c_k}^{-1} \right)^{-1}. \quad (15)$$

With these estimates, the dereverberation can be accomplished by the method described in section 2.1.

2.3. Solution interpretation

According to the posterior pdf of $\mathbf{c}_k$ derived above, the contribution of the prior information on $\mathbf{c}_k$ to the dereverberation is represented by Σ_{c_k} in (15). Interestingly, this contribution becomes negligible as the number of observed frames in τ becomes large. Then, the conditional mean of the reverberation in (2), $\bar{r}_{t,k}$, given as (3), becomes identical to that of the conventional dereverberation method. In addition, the conditional covariance of the deviation of $r_{t,k}$ from the mean, $E\{|e_{t,k}|^2|\xi_{\tau,k}\}$ represented as (5), approaches zero. This means that the solution of the proposed method becomes identical to the STFT domain filtering employed by the conventional method when a sufficiently long observed signal is available.

In contrast, when we disregard the information obtained from the observation, that is, when we set the summations over the time frames in (14) and (15) at zero, the conditional mean and the covariance, respectively, become

$$\bar{r}_{t,k} = 0,$$

$$E\{|e_{t,k}|^2|\xi_{\tau,k}\} = \sum_{t'=1}^{T} \gamma_{t',k}^2 |x_{t-t',k}|^2,$$

where $E\{|e_{t,k}|^2|\xi_{\tau,k}\}$ is calculated as a convolution between the long-term AR coefficients and the past observed signal in the power spectral domain. According to these estimates, the dereverberation can be accomplished simply by reducing $E\{|e_{t,k}|^2|\xi_{\tau,k}\}$ from the observed signal in the power spectral domain. In this respect, the proposed method has certain correspondences with existing dereverberation methods that operate in the power spectral domain [2, 3, 6], and thus provides a new estimation scheme for this approach.

As a consequence, the proposed method achieves the dereverberation by a combination of filtering in the STFT domain and reverberation reduction in the power spectral domain.

2.4. Processing flow

We implemented the proposed method so that it dereverberates the observed signal incrementally from the beginning of the signal with certain algorithmic delay. The processing flow is summarized as

1. Set predetermined values at Σ_{c_k} for all k.
2. The observed signal is segmented into short time frames of a fixed length, and transformed into the STFT domain. The frames are then segmented into frame blocks of a fixed length.
3. Incrementally apply the following for each frame block.
   (a) Set the initial values of Σ_{s_t} as $\sigma_{t,k}^2 = |x_{t,k}|^2$.
   (b) Repeat the following for all k
      i. Update the summations in (14) and (15) by adding terms corresponding to the current block, and obtain $p_{c_k}(\mathbf{c}_k|\xi_{\tau,k})$ by (13).
      ii. Update Σ_{s_t} based on the posterior pdf of the speech signal, $p(s_{t,k}|\xi_{\tau,k})$.
   (c) Obtain $\tilde{x}_{t,k} = x_{t,k} - \bar{r}_{t,k}$, and estimate the clean speech power spectrum using spectral subtraction.
   (d) Synthesize the estimated speech signal in the current block by using the waveform.

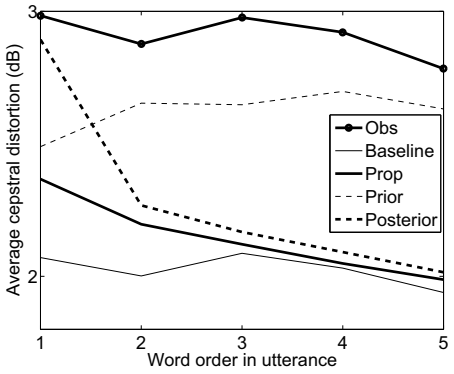

Fig. 1. Average cepstral distortions for the observed signals (Obs) and signals dereverberated by Baseline, Prop, Prior, and Posterior in relation to the word order in the utterances.

3. EXPERIMENTS

We evaluated the proposed method, which we refer to hereafter as "Prop", in comparison with the conventional method proposed in [4]. Because the conventional method performed with batch processing, it is referred to hereafter as "Baseline". With Baseline, the long-term AR coefficients were determined after a whole observed signal had been obtained, and the dereverberation of the whole signal was performed based on these AR coefficients. In contrast, with Prop, the posterior pdf of the long-term AR coefficients were estimated at each time block using the signals obtained by the time block, and the dereverberation of the block was performed based on this pdf. Therefore, the results obtained with Baseline are dealt with as the desired performance with incremental processing in this paper. Furthermore, we also tested the behavior of Prop in two different settings to examine the effect of the prior information. In one setting, Prop was performed without using any prior information on the room acoustics, that is, with $\Sigma_{c_k}^{-1}$ in (15) set at zero. This setting is referred to as "Posterior." In the other setting, Prop was performed only with prior information, that is, the summation terms in (14) and (15) were both set at zero. This setting is referred to as "Prior." Prop, Posterior, and Prior were all performed with incremental processing.

To test the effectiveness of each method, we prepared five utterances by two speakers (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. The observed signals were synthesized by convolving each utterance with a 1-ch RIR measured in a reverberant room with a reverberation time (RT60) of 0.5 sec. Dereverberation was performed for each utterance, and the performance was evaluated in terms of the cepstral

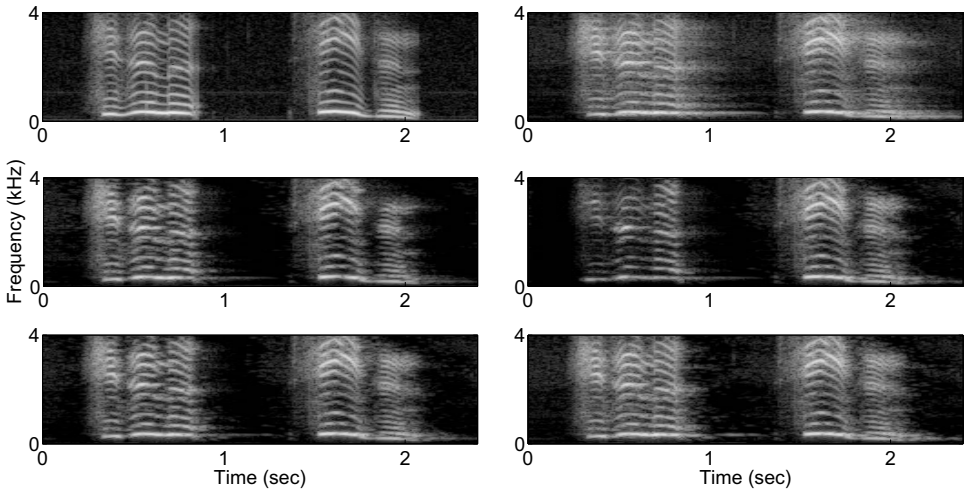

Fig. 2. Spectrograms of clean (top left), and reverberated (top right) signals, and signals dereverberated by Prior (middle left), Posterior (middle right), Prop (bottom left), and Baseline (bottom right). Only first two words in a female utterance are shown in the figure.

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 speech signal being evaluated and the original clean speech signal, and we adopted $D = 12$. To reduce the effect of the early reflections that remain in the dereverberated signals, we applied cepstral mean normalization to both 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. The analysis window size and the frame shift were set at 256 and 128, respectively. The block size of the incremental processing was set at 16 frames ($= 256$ ms). This corresponds to the period for updating the posterior pdf of $\bar{\mathbf{c}}_k$. To determine the prior pdf of the long-term AR coefficients for Prop, or Σ_{c_k}, we first applied Baseline to a female utterance convolved with an RIR that was measured at a different location in the same room, and set Σ_{c_k} based on (10) using the obtained AR coefficients. The order of the long-term AR process was set at 24 for each frequency bin. The iteration number for step 3-(b) was set at 3.

Figure 1 plots the average CDs of the observed signals, and those of the signals dereverberated by Baseline, Prop, Prior, and Posterior. The average CDs were calculated separately for the first to the fifth words over different utterances, and plotted according to the word order. The figure shows that Baseline was the best at reducing the average CDs. In contrast, Prop and Posterior were worse for the first words in the utterances than Baseline, but their performance quickly improved for the following words and approached that of Baseline. When comparing Prop and Posterior, Prop improved the quality of the first words much better than Posterior. Prior also reduced the average CDs stably from the beginning, but the improvement was relatively small. These results show that the prior information on the room acoustics enabled Prop to dereverberate utterances from their beginning with incremental processing and little algorithmic delay.

Figure 2 shows spectrograms of speech signals obtained before and after dereverberation. They clearly demonstrate that Prior reduced the reverberation energy effectively from the beginning of the utterance while Posterior quickly improved the dereverberation performance for the second word in the utterance. In contrast, Prop took effective advantage of these two methods, Prior and Posterior, for the incremental processing.

4. CONCLUSION

This paper described a way of incorporating the prior pdf of the long-term AR coefficients that characterizes a general acoustic property of a room into the proposed method to achieve speech dereverberation based on incremental processing with little algorithmic delay. The proposed method integrates the prior information and the information obtained from the observed signals based on the Bayes' rule. The experiments showed that the proposed method can effectively dereverberate the word sequence incrementally from its beginning without degrading the quality of the speech signal under a 0.5 sec reverberation time (RT60) condition.

5. REFERENCES

[1] T. Nakatani, B.H. Juang, T. Yoshioka, K. Kinoshita, and M. Miyoshi, "Importance of energy and spectral features in Gaussian source model for speech dereverberation," *Proc. WASPAA-2007*, pp. 299–302, 2007.

[2] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "A method based on the MTF concept for dereverberating the power envelope from the reverberant signal," *Proc. ICASSP-2003*, vol. 1, pp. 840–843, 2003.

[3] E.A.P. Habets, N. Gaubitch, and P.A. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," *Proc. ICASSP-2008*, pp. 4577–4580, 2008.

[4] 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.

[5] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," *Proc. ICASSP-2008*, pp. 4441–4444, 2008.

[6] R. Gomez, J. Even, H. Saruwatari, and K Shikano, "Distant-talking robust speech recognition using late reflection components of room impulse response," *Proc. ICASSP-2008*, pp. 4581–4584, 2008.