INTRODUCTION OF SPEECH LOG-SPECTRAL PRIORS INTO DEREVERBERATION BASED ON ITAKURA-SAITO DISTANCE MINIMIZATION

Yasuaki Iwata[⋆†] *Tomohiro Nakatani*[†]

[⋆]Graduate School of Information Science, Nagoya University, Aichi, Japan
[†]NTT Communication Science Labs., NTT Corporation, Kyoto, Japan

ABSTRACT

It has recently been shown that a multi-channel linear prediction can effectively achieve blind speech dereverberation based on maximum-likelihood (ML) estimation. This approach can estimate and cancel unknown reverberation processes from only a few seconds of observation. However, one problem with this approach is that speech distortion may increase if we iterate the dereverberation more than once based on Itakura-Saito (IS) distance minimization to further reduce the reverberation. To overcome this problem, we introduce speech log-spectral priors into this approach, and reformulate it based on maximum a posteriori (MAP) estimation. Two types of priors are introduced, a Gaussian mixture model (GMM) of speech log spectra, and a GMM of speech mel-frequency cepstral coefficients. In the formulation, we also propose a new versatile technique to integrate such log-spectral priors with the IS distance minimization in a computationally efficient manner. Preliminary experiments show the effectiveness of the proposed approach.

Index Terms— Dereverberation, probabilistic speech model, Itakura-Saito distance, Maximum a posteriori estimation, Gaussian mixture model

1. INTRODUCTION

Speech signals captured in an enclosed space such as a conference room will inevitably contain reverberant components because of reflections from the walls, the floor or the ceiling. As a result, the captured signals become less intelligible and often seriously degrade many speech applications, including automatic speech recognition.

To cope with this problem, dereverberation techniques have been studied that cancel out the reverberant components of the observed signals and recover the quality of the original speech signals [1, 2, 3]. Dereverberation based on multi-channel linear prediction (MCLP) in the frequency domain is one such technique [3]. This method performs dereverberation based on maximum likelihood (ML) estimation, where a time-varying Gaussian model (TVGM) estimated directly from the reverberant observation is used as an approximated speech probabilistic model. It has been shown that this method effectively achieves dereverberation based only on a relatively short observation with no prior knowledge of the reverberation process. An iterative estimation scheme was also proposed with this approach to further reduce the reverberation. In the scheme, not only the speech estimates but also the TVGM are updated alternately in each iteration, where the TVGM is updated so that the Itakura-Saito (IS) distance between the speech estimates and the TVGM is minimized.

However, there is a problem in the iterative estimation scheme. It may also increase distortion of speech particularly when the observation is short. As a result, we need to limit the number of iterations, which also limits the dereverberation performance. One cause of this problem is that the TVGM is updated based not on any prior knowledge of the source, and thus the power spectra represented by the TVGM may be updated to values that speech spectra never take. A method for modeling the speech power spectra using an autoregressive model has also been proposed [4], but the same problem occurs even with this method.

To overcome this problem, this paper proposes a way of introducing speech log-spectral priors trained in advance into MCLP based speech dereverberation. Based on the log-spectral priors, the TVGM can be updated to more appropriate one, therefore, the speech signals are less likely to be distorted by iterative estimation. In particular, we propose using a log-spectral priors represented by a GMM of speech log spectra (LS-based prior) and by a GMM of speech mel-frequency cepstral coefficients (MFCC-based prior). Here, one important issue is to find a way of efficiently solving the non-linear optimization problem that results from the introduction of the priors. This paper therefore proposes a method for combining the expectation-maximization (EM) algorithm and Newton's method to realize this optimization. Note that GMMs of speech log-spectra/MFCCs are widely used for automatic speech recognition as probabilistic models of speech features, however they have hardly been used for speech enhancement based on the IS distance minimization [5]. The proposed method, thus, can be viewed as a versatile technique that can be applied to the many other speech enhancement methods based on the IS distance minimization.

In the rest of this paper, Section 2 overviews the dereverberation method based on MCLP in the frequency domain. Section 3 describes the two proposed methods. Sections 4 and 5, respectively, provide experimental results and concluding remarks.

2. DEREVERBERATION BASED ON ITAKURA-SAITO DISTANCE MINIMIZATION

Suppose that a single speech source is captured by L_m microphones. Let $s_{t,f}$ and $x_{t,f}^{(m)}$ be the time-frequency (TF) domain representations of the source signal and the observed signal, respectively, where t and m are time frame and microphone indices, respectively. The observed signal can be represented as follows [3]:

$$x_{t,f}^{(m)} = \sum_{k=0}^{L_h-1} \left(h_{k,f}^{(m)}\right)^* s_{t-k,f} + e_{t,f}^{(m)} + b_{t,f}^{(m)}, \qquad (1)$$

where $h_{t,f}^{(m)}$, $e_{t,f}^{(m)}$, and $b_{t,f}^{(m)}$ are the room impulse response (RIR) of length L_h from the source to the mth microphone in the TF domain, the modeling errors of the RIR convolution in the TF domain, and the noise signal. " $(\cdot)^*$ " denotes a complex conjugate operation. Then, assuming $e_{t,f}^{(m)} = b_{t,f}^{(m)} = 0$ for simplicity's sake, the observed

signal at $m = 1$ can be represented in the MCLP form as

$$x_{t,f}^{(1)} = \bar{c}_f^{*T} \bar{x}_{t-D,f} + d_{t,f}^{(1)}, \tag{2}$$

$$d_{t,f}^{(1)} = \sum_{k=0}^{D-1} \left(h_{k,f}^{(1)}\right)^{*T} s_{t-k,f}, \tag{3}$$

where $\bar{c}_f$ is a regression coefficient vector of the order of L_c, D is a time duration corresponding to the early reflections of the reverberation, "$\bar{\cdot}$" denotes a vector symbol, "$(\cdot)^T$" denotes a vector/matrix non-conjugate transposition operation, and

$$\bar{x}_{t,f} = \left[\left(\bar{x}_{t,f}^{(1)}\right)^T, \left(\bar{x}_{t,f}^{(2)}\right)^T, \ldots, \left(\bar{x}_{t,f}^{(L_m)}\right)^T\right]^T,$$

$$\bar{x}_{t,f}^{(m)} = \left[x_{t,f}^{(m)}, x_{t-1,f}^{(m)}, \ldots, x_{t-L_c,f}^{(m)}\right]^T,$$

$$\bar{c}_f = \left[\left(\bar{c}_f^{(1)}\right)^T, \left(\bar{c}_f^{(2)}\right)^T, \ldots, \left(\bar{c}_f^{(L_m)}\right)^T\right]^T,$$

$$\bar{c}_f^{(m)} = \left[c_{1,f}^{(m)}, c_{2,f}^{(m)}, \ldots, c_{L_c,f}^{(m)}\right]^T.$$

The goal of the dereverberation method proposed in [3] is to estimate the regression coefficient vector $\bar{c}_f$ and then recover, based on (2), the desired signal $d_{t,f}^{(1)}$, which only contains the direct signal and the early reflections. To estimate $\bar{c}_f$, a likelihood function based on a probabilistic speech model is used as an optimization criterion. With the probabilistic model, the desired signal is assumed to follow a time-varying Gaussian distribution, and the probability density function (pdf) is defined as

$$p\left(d_{t,f}^{(1)}\right) = \mathcal{N}_c\left(d_{t,f}^{(1)}; 0, \sigma_{t,f}^2\right), \tag{4}$$

where $\mathcal{N}_c(\cdot)$ is a pdf of a complex Gaussian distribution, and $\sigma_{t,f}^2 = E\{d_{t,f}^{(1)} d_{t,f}^{(1)*}\}$ corresponds to the variance of the process, or the power spectrum of $d_{t,f}^{(1)}$. This model is capable of precisely representing the characteristics of any time-varying power spectra because $\sigma_{t,f}^2$ can take any value in each short-time frame. Because $\sigma_{t,f}^2$ is not given in advance, it is considered a parameter to be estimated.

Let $\sigma_f^2 = \{\sigma_{1,f}^2, \sigma_{2,f}^2, \ldots\}$ be a time series of $\sigma_{t,f}^2$ for all frames t at a frequency bin f, and $\theta_f = \{\bar{c}_f, \sigma_f^2\}$ be the parameter set to be estimated. Then, the log likelihood function can be derived as

$$\mathcal{L}(\theta) = \sum_t \log p\left(d_{t,f}^{(1)} = x_{t,f}^{(1)} - \bar{c}_f^{*T} \bar{x}_{t-D,f}; \theta_f\right), \tag{5}$$

$$= -\sum_t \left(\frac{\left|x_{t,f}^{(1)} - \bar{c}_f^{*T} \bar{x}_{t-D,f}\right|^2}{\sigma_{t,f}^2} + \log \sigma_{t,f}^2\right) + \text{const.} \tag{6}$$

Here, $(\cdot)$ in (6) is equivalent to the IS distance between $|d_{t,f}^{(1)}|^2$ and $\sigma_{t,f}^2$ for the estimation of $\sigma_{t,f}^2$ given $|d_{t,f}^{(1)}|^2$, and the minimization of this term corresponds to the maximization of the likelihood.

To estimate $\bar{c}_f$ and $\sigma_{t,f}^2$, the likelihood function can be maximized by iterating the following:

1. Initialize $\hat{\sigma}_{t,f}^2$ as $\hat{\sigma}_{t,f}^2 = |x_{t,f}|^2$.
2. Repeat the following until convergence.
   (a) Update $\hat{\bar{c}}_f$ as $\hat{\bar{c}}_f = \hat{\Phi}^+ \hat{\bar{\phi}}$.
   (b) Update $\hat{d}_{t,f}$ as $\hat{d}_{t,f} = x_{t,f}^{(1)} - \hat{\bar{c}}_f^{*T} \bar{x}_{t-D,f}$.
   (c) Update $\hat{\sigma}_{t,f}^2$ as $\hat{\sigma}_{t,f}^2 = \max\{|\hat{d}_{t,f}|^2, \epsilon_f\}$.

where "$\hat{\cdot}$" denotes an estimated value, ϵ_f is a small positive constant used to avoid zero division, "$(\cdot)^+$" is the Moore-Penrose pseudo-inverse, and

$$\hat{\Phi} = \sum_t \frac{\bar{x}_{t-D,f} \bar{x}_{t-D,f}^{*T}}{\hat{\sigma}_{t,f}^2}, \quad \hat{\bar{\phi}} = \sum_t \frac{\bar{x}_{t-D,f} x_{t,f}^{(1)*}}{\hat{\sigma}_{t,f}^2}. \tag{7}$$

It is important to note that the first iteration in the above procedure surely improves the quality of the speech, but the following iterations do not necessarily do so. Indeed, the quality often degrades, particularly when the observation is very short. This is because in the above 2(c), $\sigma_{t,f}^2$ is updated to a power spectrum of $|\hat{d}_{t,f}|^2$ based on the IS distance minimization with no spectral priors. Through this update, $\sigma_{t,f}^2$ may take a value that a speech power spectrum can never take. This is the problem with the conventional method.

3. DEREVERBERATION WITH LOG-SPECTRAL PRIORS

In this section, we describe our two proposed methods. To overcome the above problem, we introduce two speech log-spectral priors, the LS-based prior and the MFCC-based prior, into the proposed methods.

3.1. Dereverberation with LS-based prior

The first proposed method performs a dereverberation based on the maximum a posteriori (MAP) estimation with an LS-based prior. Let a time series of the speech log spectra be $\rho_f = \{\rho_{1,f}, \rho_{2,f}, \ldots\}$ where $\rho_{t,f} = \log \sigma_{t,f}$, and a conditional pdf of the parameters $\bar{c}_f$ and ρ_f given the time series of the observed signals $x_f = \{x_{1,f}, x_{2,f}, \ldots\}$ be represented as

$$p(\bar{c}_f, \rho_f | x_f) \propto p(x_f | \bar{c}_f, \rho_f) p(\rho_f), \tag{8}$$

where we disregarded the prior term $p(\bar{c}_f)$, assuming it to be uniform. We also disregarded the prior term $p(x_f)$ as a constant term. On the right side of (8), $p(x_f|\bar{c}_f, \rho_f)$ is equivalent to (6), and $p(\rho_f)$ is the prior of the speech log spectra.

Then, a speech log spectrum $\bar{\rho}_t = [\rho_{t,1}, \rho_{t,2}, \ldots]^T$ in each time frame is modeled by a multivariate GMM (log-spectral GMM) as

$$p(\bar{\rho}_t) = \sum_i^M p(i) \mathcal{N}(\bar{\rho}_t; \bar{\mu}_i, \Omega_i), \tag{9}$$

where M is the number of mixture components, and $p(i)$ is a mixture weight for $i = 1, \ldots, M$ that satisfies $\sum_{i=1}^M p(i) = 1$ and $p(i) \geq 0$. $\mathcal{N}(\cdot)$ is a multivariate Gaussian pdf, where $\bar{\mu}_i$ and Ω_i are the mean vector and the covariance matrix of $\bar{\rho}_t$ for each i, respectively. We assume that Ω_i is diagonal with its diagonal components, $\{\omega_{i,1}^2, \omega_{i,2}^2, \ldots\}$. This allows us to perform the MAP estimation separately in individual frequency bins by using (8).

Let $\theta_f = \{\bar{c}_f, \rho_f\}$ and $\theta = \{\theta_1, \theta_2, \ldots\}$ be the parameter set to be estimated. Then, we adopt the EM algorithm to estimate the parameter set θ_f that maximizes (8), letting i be a hidden variable. Accordingly, θ_f is estimated based on the following procedure:

1. Initialize θ.
2. Repeat the following until convergence.
   (a) E-step: Compute $Q(\theta_f|\hat{\theta})$.
   (b) M-step: Set $\hat{\theta}_f = \arg\max_{\theta_f} Q(\theta_f|\hat{\theta})$.

where $Q(\theta_f|\hat{\theta})$ is the auxiliary function defined as

$$Q(\theta_f|\hat{\theta}) = -\sum_t \left(\frac{\left| x_{t,f}^{(1)} - \bar{c}_f^{*T} \bar{x}_{t-D,f} \right|^2}{\exp(2\rho_{t,f})} + 2\rho_{t,f} + \sum_i^M z_{i,t} \frac{(\rho_{t,f} - \mu_{i,f})^2}{2\omega_{i,f}^2} \right), \quad (10)$$

$$z_{i,t} = \frac{p(i)\mathcal{N}(\hat{\bar{\rho}}_t; \bar{\mu}_i, \Omega_i)}{\sum_i^M p(i)\mathcal{N}(\hat{\bar{\rho}}_t; \bar{\mu}_i, \Omega_i)}. \quad (11)$$

Specifically, we perform the optimization as follows:

1. Initialize $\hat{\rho}_{t,f}$ as $\hat{\rho}_{t,f} = \log|x_{t,f}|$.
2. Repeat the following until convergence.
   (a) Update $\hat{\bar{c}}_f$ as $\hat{\bar{c}}_f = \hat{\Phi}^+ \hat{\bar{\phi}}$.
   (b) Update $\hat{d}_{t,f}$ as $\hat{d}_{t,f} = x_{t,f}^{(1)} - \hat{\bar{c}}_f^{*T} \bar{x}_{t-D,f}$.
   (c) Update $\hat{\rho}_{t,f}$ so that it satisfies $\partial Q / \partial \rho_{t,f} = 0$.

In the above procedure, $\hat{\rho}_{t,f}$ is updated using the prior of $\rho_{t,f}$ in 2(c), and only the way of updating this parameter is different from that of the conventional dereverberation method. For the update of $\hat{\rho}_{t,f}$, we adopt Newton's method, which can be accomplished in a computationally efficient way as explained in the following.

3.2. Update of $\rho_{t,f}$ with Newton's method

First, the equation, $\partial Q / \partial \rho_{t,f} = 0$, can be rewritten as

$$\exp(x) + x + a = 0, \quad (12)$$

by setting

$$x = -2\rho_{t,f} + \log \frac{4|d_{t,f}^{(1)}|^2}{\sum_i^M \frac{z_{i,t}}{\omega_{i,f}^2}},$$

$$a = \frac{2}{\sum_i^M \frac{z_{i,t}}{\omega_{i,f}^2}} \left(\sum_i^M \frac{z_{i,t}\mu_{i,f}^2}{\omega_{i,f}^2} - 2 \right) - \log \frac{4|d_{t,f}^{(1)}|^2}{\sum_i^M \frac{z_{i,t}}{\omega_{i,f}^2}}.$$

Because this is a one-dimensional nonlinear optimization problem, it can be solved by Newton's method in a computationally efficient way. It is guaranteed that Newton's method converges to a unique solution because the left side of (12) is concave and monotonically increasing. Furthermore, we set the initial value x_0 of x at

$$x_0 = \begin{cases} \log(-a) & \text{for } a \le -1/2 \\ -a & \text{for } a > -1/2 \end{cases} \quad (13)$$

because (12) around the solution can be roughly approximated as $\exp(x) + a = 0$ when a is small, and as $x + a = 0$ when a is large. Then, our preliminary experiments based on various settings showed that Newton's method always converged after only two iterations to solutions that were sufficiently close to the true ones.

3.3. Dereverberation with MFCC-based prior

The second proposed method performs the dereverberation based on MAP estimation with an MFCC-based prior. Let $\bar{m}_t = [m_{t,1}, m_{t,2}, \dots]^T$ be an MFCC in each time frame, then we model the relationship between $\bar{\rho}_t$ and $\bar{m}_t$ by a linear regression model (LRM) as

$$\bar{\rho}_t = \mathbf{A}\bar{m}_t + \bar{b} + \bar{e} \quad , \quad p(\bar{e}) = \mathcal{N}(\bar{e}; \bar{0}, \Gamma), \quad (14)$$

where $\mathbf{A}$ is an (order of $\bar{\rho}_t$)×(order of $\bar{m}_t$) matrix and $\bar{b}$ is an (order of $\bar{m}_t$)-dimensional column vector. $\bar{e} = [e_1, e_2, \dots]^T$ represents the modeling error, and we assume it follows a Gaussian pdf with a mean vector $\bar{0}$ and a diagonal covariance matrix Γ with its diagonal components, $\{\gamma_1^2, \gamma_2^2, \dots\}$. We further assume that the values of $\mathbf{A}$, $\bar{b}$, and γ_f^2 can be fixed in advance using a certain speech database.

Let the time series of $\bar{x}_t$, $\bar{\rho}_t$, and $\bar{m}_t$ be $x = \{\bar{x}_1, \bar{x}_2, \dots\}$, $\rho = \{\bar{\rho}_1, \bar{\rho}_2, \dots\}$, and $m = \{\bar{m}_1, \bar{m}_2, \dots\}$, respectively, then, the conditional pdf of the parameters $c = \{\bar{c}_1, \bar{c}_2, \dots\}$, ρ, and m given x is represented as follows:

$$p(c, \rho, m|x) \propto p(x|c, \rho)p(\rho|m)p(m), \quad (15)$$

where we again disregarded the priors $p(c)$ and $p(x)$. On the right side of (15), $p(x|c, \rho)$ is equivalent to the sum of (6) for all f, $p(\rho|m)$ is the posterior of the speech log spectra given the MFCCs, which can be defined by (14), and $p(m)$ is the prior of the MFCCs.

With this method, we model an MFCC $\bar{m}_t$ in each time frame by using a multivariate GMM (MFCC-GMM) as follows:

$$p(\bar{m}_t) = \sum_i^M p(i)\mathcal{N}(\bar{m}_t; \bar{\mu}_i, \Omega_i), \quad (16)$$

where $\bar{\mu}_i$ and Ω_i are the mean vector and the covariance matrix of $\bar{m}_t$ for each i, respectively. We assume that Ω_i is diagonal with its diagonal components, $\{\omega_{i,1}^2, \omega_{i,2}^2, \dots\}$.

We again adopt the EM algorithm to maximize (15) and to estimate the parameter set composed of c, ρ, and m. The auxiliary function is defined as

$$Q(\theta|\hat{\theta}) = -\sum_t \left\{ \sum_f \left(\frac{\left| x_{t,f}^{(1)} - \bar{c}_f^{*T} \bar{x}_{t-D,f} \right|^2}{\exp(2\rho_{t,f})} + 2\rho_{t,f} + \frac{\{\rho_{t,f} - (\bar{a}_f \bar{m}_t + b_f)\}^2}{2\gamma_f^2} \right) + \sum_i z_{i,t} \sum_l \frac{(m_{t,l} - \mu_{i,l})^2}{2\omega_{i,l}^2} \right\}, \quad (17)$$

$$z_{i,t} = \frac{p(i)\mathcal{N}(\hat{\bar{m}}_t; \bar{\mu}_i, \Omega_i)}{\sum_i p(i)\mathcal{N}(\hat{\bar{m}}_t; \bar{\mu}_i, \Omega_i)}, \quad (18)$$

where $\bar{a}_f$ and b_f are a vector and a scalar in the fth row of $\mathbf{A}$ and $\bar{b}$, respectively. With the auxiliary function (17), we perform iterative optimization as follows:

1. Initialize $\hat{\rho}_{t,f}$ as $\hat{\rho}_{t,f} = \log|x_{t,f}|$.
2. Initialize $\hat{\bar{m}}_t$ as $\hat{\bar{m}}_t = \mathbf{A}^+(\hat{\bar{\rho}}_t - \bar{b})$.
3. Repeat the following until convergence.
   (a) Update $\hat{\bar{c}}_f$ as $\hat{\bar{c}}_f = \hat{\Phi}^+ \hat{\bar{\phi}}$.
   (b) Update $\hat{d}_{t,f}$ as $\hat{d}_{t,f} = x_{t,f}^{(1)} - \hat{\bar{c}}_f^{*T} \bar{x}_{t-D,f}$.
   (c) Update $\hat{\bar{m}}_t$ as $\hat{\bar{m}}_t = \hat{\Psi}^{-1} \hat{\bar{\psi}}$.
   (d) Update $\hat{\rho}_{t,f}$ so that it satisfies $\partial Q / \partial \rho_{t,f} = 0$.

where

$$\hat{\Psi} = \mathbf{A}^T \Gamma^{-1} \mathbf{A} + \sum_i z_{i,t} \Omega_i^{-1}, \quad (19)$$

$$\hat{\bar{\psi}} = \mathbf{A}^T \Gamma^{-1} (\bar{\rho}_t - \bar{b}) + \sum_i z_{i,t} \Omega_i^{-1} \bar{\mu}_i. \quad (20)$$

As in 3(c) of the above procedure, $\hat{\bar{m}}_t$ can be updated with a closed-form equation because the auxiliary function is in a quadratic form with respect to $\hat{\bar{m}}_t$. For the update of $\hat{\rho}_{t,f}$ in 3(d), we can again use Newton's method in the same way as described in section 3.2. As a whole, the estimation can again be accomplished in a computationally efficient manner.

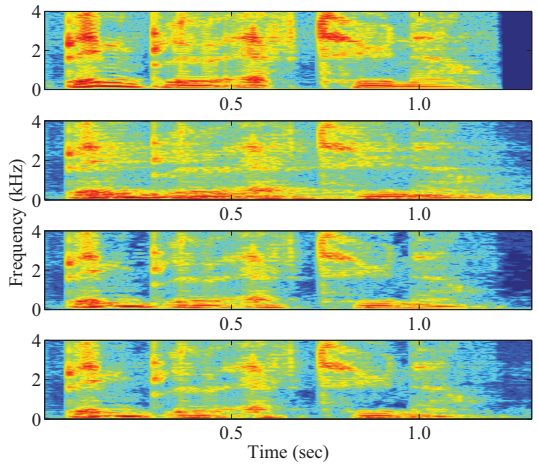

Fig. 1. Example spectrograms of clean (top) and reverberated (2nd) signals, and signals dereverberated with an LS-based prior (3rd) and an MFCC-based prior (bottom) when the number of iterations was five.

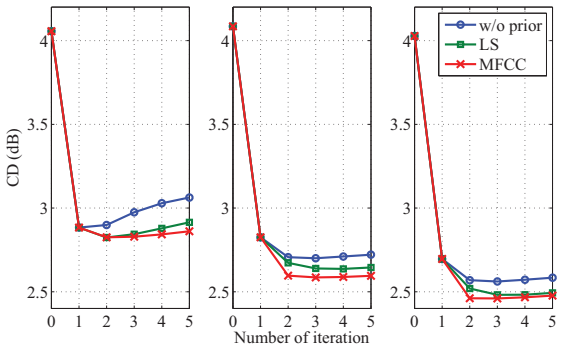

Fig. 2. Average CDs obtained by the conventional method with no spectral priors and by two proposed methods (LS and MFCC), depending on the number of EM iterations when the length of the observation was 1.15 sec (left), 2.3 sec (center), and 4.6 sec (right).

4. PRELIMINARY EXPERIMENT

To test the effectiveness of the proposed methods, we used 30 utterances of a single male speaker for evaluation. The length of an utterance was 2.3 sec on average (min: 2.09 sec and max: 2.8 sec). The observed signals were synthesized by convolving each utterance with 2-ch room impulse responses (RIR) measured in a reverberant room with a reverberation time (RT60) of 0.5 sec. To evaluate the dependence of dereverberation performance on the length of the observation, we prepared two more utterance sets. In one set, each utterance in the above original set was separated into two utterances at its center time, resulting in 60 utterances (average length: 1.15 sec). In the other set, we concatenated 30 pairs of same utterances in the original data set, resulting in 30 utterances (average length: 4.6 sec). In both sets, the RIRs were convolved separately with each utterance. Dereverberation was performed for each utterance, and the performance was evaluated in terms of the cepstral distortion (CD). See [3] for the definition of a CD.

The log-spectral GMM in the first proposed method and the LRM and the MFCC-GMM in the second proposed method were trained on 500 utterances from the same speaker[1]. We set the orders of the log spectrum $\bar{\rho}_t$ and the MFCC $\bar{m}_t$ at 257 and 13, respectively, and set the number of mixture components M in the GMMs of both proposed methods at 256.

Example spectrograms of a speech signal before and after dereverberation are shown in Fig.1, and indicate that the time and frequency structure of the signal was clearly recovered by both proposed methods. Figure 2 compares the conventional dereverberation method with no spectral priors and the two proposed methods in terms of CDs depending on the number of EM iterations. The iteration number 0 means the observed signal. Next, the CDs obtained with all 3 methods were the same after the first iteration because the speech log-spectral priors were not used at the first update of the regression coefficients $\hat{\bar{c}}_f$. After the second iteration, the method without priors gradually increased the CDs with increased iteration number particularly in the shortest observation case. In contrast, the two proposed methods improved the CDs much more than the conventional method with no priors when the iteration number exceeded one. In particular, they barely increased the CDs even after several iterations using the shortest observation. These results clearly demonstrate the effectiveness of the introduction of the speech log-spectral priors.

[1]Here, to exclude the effect of early reflections in the preliminary experiments, each utterance was convolved with the early-reflection components of the above RIRs before the training.

5. CONCLUSION

We proposed dereverberation methods based on the IS distance minimization with two different speech log-spectral priors, namely an LS-based prior and an MFCC-based prior. Preliminary experiments revealed that, in terms of cepstral distortion, the two proposed methods can improve the quality of the dereverberated signals much more than the conventional dereverberation method with no spectral priors after more than one EM iteration. Future work should include a comprehensive evaluation of the proposed methods under different observation conditions, including noisy observation cases. In addition, it would be very interesting to use the proposed approach to combine the IS distance measure with GMM-based log-spectral priors and the other speech enhancement approaches based on IS distance minimization.

6. REFERENCES

[1] B.W. Gillespie, H.S. Malvar, and D.A.F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. ICASSP-2011*, vol. 6, pp. 3701–3704, 2001.

[2] P.A. Naylor and N.D. Gaubitch (Eds.). *Speech Dereverberation*, Springer 2010.

[3] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[4] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, 2009.

[5] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Computation*, 21(3), 793-830, Mar., 2009.