BLIND SPEECH DEREVERBERATION WITH MULTI-CHANNEL LINEAR PREDICTION BASED ON SHORT TIME FOURIER TRANSFORM REPRESENTATION

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

It has recently been shown that the use of the time-varying nature of speech signals allows us to achieve high quality speech dereverberation based on multi-channel linear prediction (MCLP). However, this approach requires a huge computing cost for calculating large covariance matrices in the time domain. In addition, we face the important problem of how to combine the speech dereverberation efficiently with many other useful speech enhancement techniques in the short time Fourier transform (STFT) domain. As the first step to overcoming these problems, this paper presents methods for implementing MCLP based speech dereverberation that allow it to work in the STFT domain with much less computing cost. The effectiveness of the present methods is confirmed by experiments in terms of the recovered signal quality and the computing time.

Index Terms— Dereverberation, Likelihood maximization, Short time Fourier transform, probabilistic speech model, Inverse filtering

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. These reverberant components have a detrimental effect on the quality of the signal and often seriously degrade many applications including automatic speech recognition.

Inverse filtering is a technique that has been studied to 減輕 mitigate the reverberation problem. It cancels out the reverberation by inverting the room impulse response (RIR), which can be considered an aggregate 聚集 of all the reflections with corresponding delays [1, 2]. In relation to this approach, the use of the time-varying characteristics of short time speech segments has recently been shown to be very important for blindly estimating a dereverberation filter that suppresses reverberation [3, 4]. This has led to a probabilistic model based formulation of multi-channel linear prediction (MCLP), where the objective is to design a filter (as part of an overall probabilistic model) that would turn reverberant speech into something that is probabilistically more like clean speech [4]. Experiments showed that this new approach allows us to achieve high quality speech dereverberation based only on a few seconds of observation.

Although important mechanisms for effective speech dereverberation have been presented as described above, fundamental problems remain that 阻止 preclude their use for real applications. One problem is that it is not easy to combine these mechanisms in an efficient manner with many useful speech enhancement techniques, such as Wiener filtering and frequency domain blind source separation, in the short time Fourier transform (STFT) domain. This is because the MCLP based dereverberation methods that have been proposed can work only in the time domain. On the other hand, we need to calculate very large covariance matrices on the observed signals in order to exploit the time-varying nature of speech signals in an optimal way, but this will inevitably and 過分的 prohibitively increase the computing cost of the dereverberation process.

As the first step to overcoming the above problems, we present new methods for implementing MCLP based dereverberation so that it can function in the STFT domain with much less computing cost. One important issue for this implementation involves finding a way to calculate the convolution with a long prediction filter precisely using the STFTs. We describe two different approaches to this problem and present two corresponding dereverberation methods, which we refer to as the method with window effect reduction (MWER) and the method with window effect compensation (MWEC). First, we show that the convolution in the time domain can be calculated approximately as convolutions in individual frequency bins in the STFT domain. With this approximation, MWER efficiently estimates the dereverberation filter based on the covariance matrices calculated separately in individual frequency bins. By contrast, with MWEC, we present a method to compensate precisely for the approximation errors found with MWER, and introduce an efficient conjugate gradient optimization 方案 scheme into the dereverberation filter estimation so that it does not require the calculation of the covariance matrix.

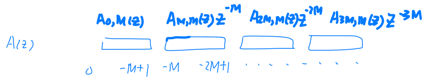
It may be important to note that STFT can also be viewed as a set of subband filters [5], and convolution is implemented in each subband separately with this 解釋 interpretation. Therefore, the dereverberation methods described in this paper can also be implemented using a subband processing technique without mojor modifications. An advantage of our approach in this paper is that the resultant terms are certain to be STFTs of the corresponding time-domain signal, and thus it is easy to combine our approach directly to the other signal processing techniques that work in the STFT domain.

In the 剩下的 remainder of this paper, section 2 overviews MCLP based dereverberation in the time domain. Section 3 describes the two methods for dereverberation in the STFT domain. Sections 4 and 5, respectively, provide experimental results and concluding remarks.

2. DEREVERBERATION IN TIME DOMAIN

Suppose that a single speech source is captured by two microphones. Let s_t and $x_t^{(l)}$ be the digitized sequences of the source signal and the observed signal, respectively, where t and l are the time and microphone indices, respectively. With MCLP, the room transfer functions in different channels are assumed to have no common zeros. Furthermore, we assume[1] that the microphone closest to the source is given in advance as $l = 1$ with the first tap of the RIR for $l = 1$

[1] These assumptions can be easily mitigated in practical applications, and the discussion here can be easily extended to more than two microphones.



being 1. Then, the relationship between the source and the observed signals can be written in MCLP as [6]

$$x_t^{(1)} = \sum_{l=1}^{2}\sum_{\tau=1}^{K} c_\tau^{(l)} x_{t-\tau}^{(l)} + s_t, \tag{1}$$

where $c_t^{(l)}$ is an MCLP prediction coefficient. In (1), the current observed signal, $x_t^{(1)}$, is predicted by sequences of past observed signals, and the source s_t is taken as the residual signal. Note that $[\underset{x_t^{(1)}}{1}\ -\underset{x_{t-1}^{(1)}\sim x_{t-K}^{(1)}}{\bar{c}^{(1)}}\ \underset{x_t^{(2)}}{0}\ -\underset{x_{t-1}^{(2)}\sim x_{t-K}^{(2)}}{\bar{c}^{(2)}}]$ where $\bar{c}^{(l)} = [c_1^{(l)}\ c_2^{(l)}\ \ldots\ c_K^{(l)}]$ can be viewed as an inverse filter that satisfies $s_t = x_t^{(1)} - \sum_l \sum_\tau c_\tau^{(l)} x_{t-\tau}^{(l)}$.

We need to introduce a certain optimization criterion to determine the prediction coefficients $\bar{c}^{(l)}$ for the dereverberation. With the probabilistic model formulation [4], a likelihood function based on a probabilistic speech model is introduced as this criterion. A time-varying Gaussian source model has been shown to be effective as this model. Its probability density function (pdf) is defined as

$$p_s(\bar{s}_t) = \mathcal{N}(\bar{s}_t; 0, \mathbf{r}_t), \tag{2}$$

where $\bar{s}_t = [s_t\ s_{t+1}\ \ldots\ s_{t+N-1}]^T$ is a vector representing a short time frame s_t of length N, and is assumed to be a stationary Gaussian process with a zero mean and the autocorrelation function $\mathbf{r}_t = E(\bar{s}_t\bar{s}_t^T)$ within the short time frame, and $\mathbf{r}_t$ is assumed to vary over the short time frames. Then, the dereverberation is defined as the problem of finding a set of parameters, $\theta = \{\bar{c}, \mathbf{r}\}$ where $\mathbf{r} = \{\mathbf{r}_t\}$ for all t, that maximizes the following likelihood function.

$$\mathcal{L}(\theta) = \sum_t \log p_x(\bar{x}_t^{(1)} | \{\bar{x}_{t-\tau}^{(l)}\}_{\tau>1,l=1,2}; \theta) \tag{3}$$

$$= \sum_t \log \mathcal{N}\left(\bar{x}_t^{(1)}; \sum_{l=1}^{2}\sum_{\tau=1}^{K} c_\tau^{(l)} \bar{x}_{t-\tau}^{(l)}, \mathbf{r}_t\right), \tag{4}$$

where $\bar{x}_t^{(l)}$ is a short time segment of $x_t^{(l)}$, $p_x(\cdot)$ is the posteriori pdf of $\bar{x}_t^{(l)}$ given the past observed signals. (3) is easily rewritten as (4) according to (1) and (2).

Effective dereverberation algorithms have been derived based on (4) with further 參數化 parameterization of $\mathbf{r}_t$ [4]. However, all such algorithms incur huge computing cost to calculate the covariance matrix $\mathbf{d}(\mathbf{r})$ below and its inverse $\mathbf{d}^{-1}(\mathbf{r})$,

$$\mathbf{d}(\mathbf{r}) = \sum_t \mathbf{x}_{t-1}^T \mathbf{r}_t^{-1} \mathbf{x}_{t-1}, \tag{5}$$

where $\mathbf{x}_{t-1} = [\bar{x}_{t-1}^{(1)}\ \bar{x}_{t-2}^{(1)}\ \ldots\ \bar{x}_{t-K}^{(1)}\ \bar{x}_{t-1}^{(2)}\ \bar{x}_{t-2}^{(2)}\ \ldots\ \bar{x}_{t-K}^{(2)}]$. This is because the prediction filter should be long and, in theory, at least as long as the room impulse response.

3. DEREVERBERATION IN STFT DOMAIN

We describe two dereverberation methods in the STFT domain, which we refer to as MWER and MWEC, in the following subsections.

3.1. Method with window effect reduction (MWER)

Suppose that $y_t^{(l)}$ is a signal obtained by 捲積 convolving an observed signal $x_t^{(l)}$ with a prediction filter $c_t^{(l)}$. Then, $y_t^{(l)}$ within a short time analysis window can be represented in the Z-domain as

$$W_N(Y^{(l)}(Z)Z^t) = W_N(C^{(l)}(Z)X^{(l)}(Z)Z^t). \tag{6}$$

where $Y^{(l)}(Z) = C^{(l)}(Z)X^{(l)}(Z)$ and $W_N(\cdot)$ is a window function of length N. $W_N(A(Z))$ extracts terms from Z^0 to Z^{-N+1} in $A(Z)$, modifies their coefficients in proportion to the window shape, and discards all other terms outside the window. Z^t is a time shift operator that shifts the short time frame starting at time t into the window function. Now, let us represent $A_{t,M}(Z) = W_M^R(A(Z)Z^t)$ where $W_M^R(\cdot)$ is a rectangular window of length M. Obviously, $A(Z) = \sum_\tau A_{\tau M,M}(Z)Z^{-\tau M}$. Then, (6) can be rewritten as

$$\underbrace{W_N(\overbrace{Y_{t,N}^{(l)}(Z)}^{W_N(Y^{(l)}(Z)Z^t)})}_{Y_{t,N}^{(l)}(Z)\text{ 多項一次沒變}} = W_N(\sum_{\tau=0}^{K_R} \overbrace{C_{\tau M,M}^{(l)}(Z)Z^{-\tau M}}^{C^{(l)}(Z)} X^{(l)}(Z)Z^t),$$

$$= \sum_{\tau=0}^{K_R} W_N(C_{\tau M,M}^{(l)}(Z)X^{(l)}(Z)Z^{t-\tau M}),$$

$$= \sum_{\tau=0}^{K_R} W_N(C_{\tau M,M}^{(l)}(Z)X_{t-M+1-\tau M,M+N-1}^{(l)}(Z)Z^{M-1}), \tag{7}$$

where $K_R \approx K/M$. The argument of the window function in (7) is a convolution of short time segments of $X^{(l)}(Z)$ and $C^{(l)}(Z)$. Now, we introduce an approximation according to a common practice in short time speech analysis, namely, that the convolution of a short time segment with a filter can be approximated by the product of the STFTs of the signal and the filter in the STFT domain when the filter is much shorter than the analysis window. We can use this approximation in the above equation when M is much smaller than N, and (7) can be rewritten on the unit circle as

$$W_N(Y_{t,N}^{(l)}(Z)) \approx \sum_{\tau=0}^{K_R} W_N^R(C_{\tau M,M}^{(l)}(Z))W_N(X_{t-\tau M,N}^{(l)}(Z)). \tag{8}$$

With discrete STFT representation, it becomes

for filter M : NWIN · hopsize N: NFFT

for signal M: hopsize N: NWIN NFFT

$$Y_n^{(l)} \approx \sum_{\tau=0}^{K_R} \text{diag}(X_{n-\tau}^{(l)})C_\tau^{(l)}, \tag{9}$$

where n and τ are frame indices, $Y_n^{(l)}$, $C_n^{(l)}$, and $X_n^{(l)}$, respectively, are vectors that contain frequency bins of the STFTs corresponding to windowed $Y^{(l)}(Z)$, $C^{(l)}(Z)$, and $X^{(l)}(Z)$, and diag(X) is a diagonal matrix that contains the elements of X as its diagonal components. As a consequence, the convolution in the time domain is represented as those of STFTs in individual frequency bins. Because M is equal to the frame shift in (8), the frame shift needs to be sufficiently small compared with the window length N with this approximation.

By applying STFT to both sides of (1) using (9) and introducing a certain delay d to the past observed signals on the right hand side, we obtain

$$X_n^{(1)} = \overbrace{\sum_{l=1}^{2}\sum_{\tau=d}^{K_R} \text{diag}(X_{n-\tau}^{(l)})C_\tau^{(l)}}^{\text{MCLP model}} + \tilde{S}_n. \tag{10}$$

Here, the delay d was introduced to prevent the STFTs of the current and past observed signals from sharing the same signal in the time domain. Because of this delay, MCLP cannot predict early reflections of reverberation included within a short time frame [7], and thus they remain in the residue. To clarify this, we denote the residue of MCLP as $\tilde{S}_n$ in (10). Note that early reflections remaining after dereverberation can be appropriately handled in many signal processing techniques based on the STFT representation [7].

The likelihood function can be defined in a similar way to the time domain algorithm. We again adopt the time-varying Gaussian process as the speech model, and define it as

$$p(\tilde{S}_n) = \mathcal{N}(\tilde{S}_n; 0, \Psi_n), \tag{11}$$

where $\Psi_n = E(\tilde{S}_n\tilde{S}_n^{*T})$ is the covariance matrix of $\tilde{S}_n$ with '$*$' representing a complex conjugate, and we assume Ψ_n is diagonal for the sake of simplicity. Then, we can rewrite the pdf of the speech separately in each frequency bin as $p(\tilde{S}_{m,n}) = \mathcal{N}(\tilde{S}_{m,n}; 0, \psi^2_{m,n})$, where $\tilde{S}_{m,n}$ is the m-th bin of $\tilde{S}_n$ and $\psi^2_{m,n} = E(\tilde{S}_{m,n}\tilde{S}^*_{m,n})$ is the variance of the frequency bin m at frame n. Let $\theta = \{\bar{C}_m, \bar{\psi}^2_m\}$ be an estimation parameter set composed of a vector of the prediction coefficients, $\bar{C}_m = [C^{(1)}_{m,d}\ C^{(1)}_{m,d+1} \cdots C^{(1)}_{m,K_R}\ C^{(2)}_{m,d}\ C^{(2)}_{m,d+1} \cdots C^{(2)}_{m,K_R}]^T$, and a time series of the source variances, $\bar{\psi}^2_m = \{\psi^2_{m,n}\}$, for all n at frequency bin m. Then, the likelihood function in each bin can be defined as

$$\begin{aligned} \mathcal{L}_m(\theta) &= \sum_n \log p(X^{(l)}_{m,n} | \mathbf{X}_{m,n-d}; \theta), \qquad (12) \\ &= \sum_n \log \mathcal{N}(X^{(l)}_{m,n}; \mathbf{X}_{m,n-d}\bar{C}_m, \psi^2_{m,n}), \end{aligned}$$

where $\mathbf{X}_{m,n-d} = [\mathbf{X}^{(1)}_{m,n-d}\ \mathbf{X}^{(2)}_{m,n-d}]$ and $\mathbf{X}^{(l)}_{m,n-d} = [X^{(l)}_{m,n-d}\ X^{(l)}_{m,n-d-1} \ldots X^{(l)}_{m,n-K_R}]$. The algorithm for maximizing (12) can be derived in a similar way to that of the time domain MCLP with a time-varying white Gaussian source model [4]. The iterative maximization algorithm is summarized as follows.

1. Set initial values as $C^{(l)}_{m,n} = 0$ and $\tilde{S}_n = X^{(1)}_n$.
2. Repeat the following until convergence
   $\hat{\psi}^2_{m,n} = \tilde{S}_{m,n}\tilde{S}^*_{m,n} \ \rightarrow\ \psi^2_{m,n}$,
   $\hat{\bar{C}}_m = (\sum_n \frac{\mathbf{X}^{*T}_{m,n-d}\mathbf{X}_{m,n-d}}{\psi^2_{m,n}})^{-1} \sum_n \frac{\mathbf{X}^{*T}_{m,n-d}X^{(1)}_{m,n}}{\psi^2_{m,n}} \ \rightarrow\ \bar{C}_m$,
   $\hat{\tilde{S}}_{m,n} = X^{(1)}_{m,n} - \mathbf{X}_{m,n-d}\bar{C}_m \ \rightarrow\ \tilde{S}_{m,n}$.

We still need to calculate the covariance matrix similar to (5) above, however, we can greatly reduce the size of the matrix and thus the computing cost compared with those of the time domain algorithm.

It is important to note that, although MWER was designed so that the effect of the circular convolution can be reduced, such effect is not explicitly avoided in the optimization process, and thus may remain in the resultant prediction filter. This is also the case with MWEC described in the next subsection. This means that the STFT representation contains a certain redundancy for the prediction filter parameter space compared with the time-domain representation. Because it is not easy to analyze theoretically how advantageously or disadvantageously such a redundancy may function with actual dereverberation, we investigate this experimentally in section 4.

3.2. Method with window effect compensation (MWEC)

By taking the window effect carefully into account, we can more precisely calculate (7) based on the STFT representation, without introducing the approximation (8). First we set $N = M$, which allows us to rewrite (7) in a rather simple form as

$$W_N(Y^{(l)}_{t,N}(Z)) = W_N(\sum_{\tau=0}^{K_C} F_\tau(Z)), \ \text{ where}$$

$$F_\tau(Z) = C^{(l)}_{\tau N,N}(Z)X^{(l)}_{t-\tau N,N}(Z) + C^{(l)}_{\tau N,N}(Z)X^{(l)}_{t-(\tau+1)N,N}(Z)Z^N,$$

where $K_C \approx K/N$. $F_\tau(Z)$ is a sum of convolutions between short time segments of the same length N, and thus can be calculated as the product of their STFTs obtained by discrete Fourier transformation (DFT) with $2N$ points. Z^N functions as a time shift operator with the order N. Because the segment length is N, we can calculate the time shift appropriately by using short time analysis with $2N$ DFT points. On the other hand, it is known that the window function can be represented as a circular convolution using the STFT representation. Consequently, the above equation can be rewritten using the STFT representation as

$$\begin{aligned} \mathbf{W}_N Y^{(l)}_n &= \mathbf{W}_N \sum_{\tau=0}^{K_C} \text{diag}(\tilde{X}^{(l)}_{m-\tau})C^{(l)}_\tau, \\ \tilde{X}^{(l)}_n &= X^{(l)}_n + \text{diag}(G)X^{(l)}_{n-1}. \end{aligned}$$

where $Y^{(l)}_n$, $C^{(l)}_n$, and $X^{(l)}_n$ are STFTs corresponding to $Y^{(l)}_{nN,N}(Z)$, $C^{(l)}_{nN,N}(Z)$, and $X^{(l)}_{nN,N}(Z)$, respectively, calculated with a rectangular window of length N and DFT points of N_p $(\geq 2N)$, G is the circular time shift operator defined as $[1\ e^{j2\pi N/N_p}\ e^{j4\pi N/N_p}\ \ldots\ e^{j2\pi N(N_p-1)/N_p}]^T$, and $\mathbf{W}_N$ is an Hermitian Toeplitz matrix that has an STFT of the window function in its first column.

Similar to (10), we can define MCLP with this approach as

$$\mathbf{W}_N X^{(1)}_n = \mathbf{W}_N \sum_{l=1}^{2} \sum_{\tau=1}^{K_C} \text{diag}(\tilde{X}^{(l)}_{n-\tau})C^{(l)}_\tau + \tilde{S}_n,$$

and we adopt the same source model as (11). The likelihood function, however, cannot be evaluated separately in each frequency bin because the window effect over frequency bins is taken into account with this approach. Let $\theta = \{\bar{C}, \psi^2_n\}$ be the estimation parameter set, where $\bar{C} = [(\bar{C}^{(1)})^T\ (\bar{C}^{(2)})^T]^T$ with $\bar{C}^{(l)} = [(\bar{C}^{(l)}_1)^T\ (\bar{C}^{(l)}_2)^T\ \ldots\ (\bar{C}^{(l)}_{K_C})^T]^T$, and let $\mathbb{X}_n = [\mathbb{X}^{(1)}_n\ \mathbb{X}^{(2)}_n]$ where $\mathbb{X}^{(l)}_n = [\text{diag}(\tilde{X}^{(l)}_n)\ \text{diag}(\tilde{X}^{(l)}_{n-1})\ \ldots\ \text{diag}(\tilde{X}^{(l)}_{n-K_C+1})]$, then the likelihood function is defined as

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_n \log p(\mathbf{W}_N X^{(1)}_n | \mathbb{X}_{n-1}; \theta), \\ &= \sum_n \log \mathcal{N}(\mathbf{W}_N X^{(1)}_n; \mathbf{W}_N \mathbb{X}_{n-1}\bar{C}, \Psi_n). \qquad (13) \end{aligned}$$

Although a repetitive maximization algorithm similar to MWER can also be derived for MWEC, it still requires a huge computing cost to calculate the covariance matrix without disregarding the window effect. Instead, we can maximize the above function efficiently based on the conjugate gradient method because it does not require us to calculate the covariance matrix. The resultant algorithm is summarized as follows.

1. Set initial values as $C^{(l)}_{m,n} = 0$ and $\tilde{S}_n = \mathbf{W_N}X^{(1)}_n$.
2. Repeat the following until convergence
   (a) $\hat{\psi}^2_{m,n} = \tilde{S}_{m,n}\tilde{S}^*_{m,n} \ \rightarrow\ \psi^2_{m,n}$
   (b) $r = \sum_n \mathbb{X}^{*T}_{n-1}\mathbf{W}^{*T}_N \Psi^{-1}_n \tilde{S}_n$
   (c) $p = r$
   (d) Repeat the following until convergence
      i. $q_n = \mathbf{W}_N \mathbb{X}_{n-1} p$,
      ii. $\alpha = r^{*T}r / \sum_n q^{*T}_n \Psi^{-1} q_n$,
      iii. $\hat{\bar{C}} = \bar{C} + \alpha p \ \rightarrow\ \bar{C}$,
      iv. $r' = r - \alpha \sum_n \mathbb{X}^{*T}_{n-1}\mathbf{W}^{*T}_N \Psi^{-1}_n q_n$,
      v. $\beta = (r')^{*T}r'/(r^{*T}r)$,
      vi. $p = r' + \beta p,\ r = r'$
   (e) $\hat{\tilde{S}}_n = \mathbf{W_N}(X^{(1)}_n - \mathbb{X}_{n-1}\bar{C}) \ \rightarrow\ \tilde{S}_n$

With the above procedure, we can calculate q_n efficiently at step 2(d)i because $\bar{C}$ in each frequency bin is very short, and we can also calculate the remaining procedure efficiently because q_n is as small as $\tilde{S}_n$, namely much smaller than the covariance matrix.

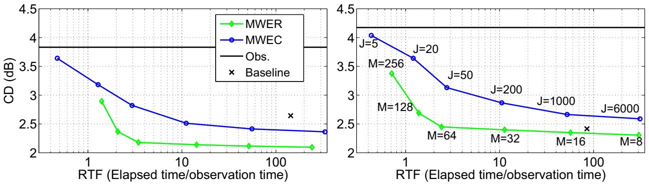


Fig. 1. Average cepstral distortions (CD) of the signals dereverberated by MWER and MWEC, average CDs of the observed signals (Obs.) and average CDs of the signals dereverberated by the time-domain algorithm (Baseline) when using u1 (left panel) and u5 (right panel) and controlling the real time factors (RTF).

4. PRELIMINARY EXPERIMENT

To test the effectiveness of the present methods, we prepared two utterance sets, u1 and u5, in which each utterance is composed of one word and five word sequences, respectively. Each set contains two utterances extracted from the ATR word utterance database, and spoken by a male and a female (MAU and FKM), respectively. The observed signals were synthesized by convolving each utterance with 2-ch RIRs 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 distortion (CD) of the recovered signals and the real time factor (RTF) of the dereverberation processing. A CD in dB is defined as

$$\mathrm{CD} = (10/\ln 10)\sqrt{2\sum_{k=0}^{D}(\hat{c}_k - c_k)^2},$$

where $\hat{c}_k$ and c_k are, respectively, cepstral coefficients of the speech signal being evaluated and the original clean speech signal, and we adopted $D = 12$. Distortions in the energy time pattern and spectral envelope were evaluated with this measure. The RTF is defined as the ratio of the computing time required for the dereverberation processing to the time duration of the observed signal. The present methods were both implemented with MATLAB, and the computing time was measured by a MATLAB interpreter on a linux computer. The computing time was controlled by changing the frame shift as $M = 256$, 128, 64, 32, 16, and 8 taps for MWER, and by increasing the iteration number of step 2(d) in the optimization algorithm as $J = 5$, 20, 50, 200, 1000, and 6000 for MWEC. The sampling rate and the STFT frame size were set at 8 kHz and $N = 256$, respectively, and the prediction filter length was set at $K_R \approx 3000/M$ for MWER and $K_C = 12$ for MWEC. The number of DFT points was set at 256 for MWER and 512 for MWEC. No a priori training was employed for the speech models in this experiment. The variances of the source models, $\psi^2_{m,n}$, were estimated from the initial source estimates, and not subsequently updated, that is, the iteration number of step 2 in MWER and MWEC was set at 1.

Figure 1 plots the CDs of the signals dereverberated by MWER and MWEC, the CDs of the observed signals, and the CDs of the signals dereverberated by the time-domain MCLP using a likelihood function corresponding to that of MWEC (Baseline). The figure shows that both present methods are able to reduce the CDs more effectively as the RTF becomes larger, and they converge to almost the same CDs. Interestingly, they are even smaller than Baseline, especially when the observed signal is short. We guess that the redundancy of the prediction filter parameter space with MWER and MWEC might function advantageously to make the resultant signal more like clean speech according to the probabilistic speech model.

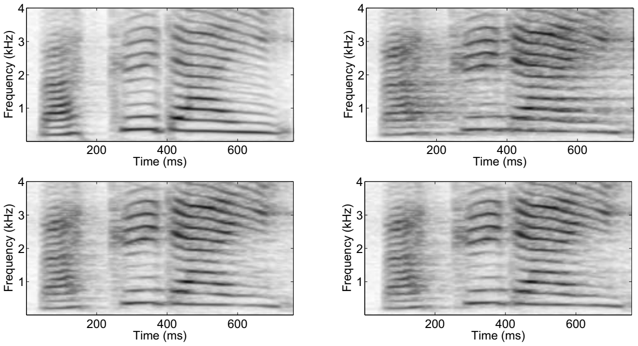


Fig. 2. Spectrograms of clean (top left), reverberated (top right) signals, and signals dereverberated by MWER with $M = 64$ (bottom left), and by MWEC with $J = 1000$ (bottom right) using u5.

Spectrograms of speech signals obtained before and after dereverberation shown in Fig. 2 indicate that the time and frequency structure of the signal was clearly recovered by MWER and MWEC. When we compared MWER and MWEC, the former was clearly superior in terms of RTFs. In particular, MWER was able to attain almost the best CDs with a frame shift of $M = 64$ that corresponds to an RTF of about 3. By contrast, MWEC required a much higher computing cost to realize its best performance although it demonstrated a certain dereverberation effect even with a low computing cost. We need to develop a more efficient optimization method for MWEC than the conjugate gradient method.

5. CONCLUSION

We presented two methods for implementing MCLP based speech dereverberation in the STFT domain, which we refer to as the method with window effect reduction (MWER) and the method with window effect compensation (MWEC). Preliminary experiments revealed that, in terms of the cepstral distortion of the dereverberated signals, both methods were comparable to even better than MCLP based dereverberation, which operates in the time domain. In addition, MWER was much better in terms of computing cost. MWER was able to achieve high quality speech dereverberation with a real time factor of about 3. Future work will include a comprehensive evaluation of the present methods in comparison with the time-domain algorithm.

6. REFERENCES

[1] B. Gillespie and L. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," *ICASSP-2003*, vol. 1, pp. 676–679, 2003.

[2] P.A. Naylor and N.D. Gaubitch, "Speech dereverberation," *IWAENC-05*, 2005.

[3] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Second-order statistics based dereverberation by using nonstationarity of speech," *IWAENC-2006*, 2006.

[4] 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, *WASPAA-2007*, 2007.

[5] M.R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," IEEE Trans. ASSP, vol. 24, No. 3, pp. 243–248, 1976.

[6] M. Miyoshi, "Estimating AR parameter-sets for linear-recurrent signals in convolutive mixtures," *ICA-2003*, pp. 585–589, 2003.

[7] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," *ICASSP-2006*, vol. 1, pp. 817–820, 2006.