Audio Engineering Society

Convention Paper

Presented at the 13th Regional Convention
2007 July 19–21 Tokyo, Japan

A linear prediction-based microphone array for speech dereverberation in a realistic sound field

Keisuke Kinoshita[1], Marc Delcroix[1], Tomohiro Nakatani[1], and Masato Miyoshi[1]

[1]*NTT Communication Science Labs. NTT Corporation*

ABSTRACT

A speech signal captured by a distant microphone is generally smeared by reverberation, which is known to severely degrade Automatic Speech Recognition (ASR) performance. One way to solve this problem is to dereverberate the observed signal prior to ASR. In this paper, we propose an efficient dereverberation method that employs multi-channel multi-step forward linear prediction. It could precisely estimate and suppress the late reflections that are known to be a major cause of ASR performance degradation. The algorithm performed good dereverberation with an amount of observed data corresponding to the duration of one speech utterance, in our case, less than 6 seconds. Experimental results showed substantial improvements in ASR performance under severe reverberant conditions. The algorithm could work below a real-time factor of 1.

1. INTRODUCTION

A speech signal captured by a distant microphone is generally smeared by reverberation. The reverberation is known to deteriorate the performance of Automatic Speech Recognition (ASR) severely. Thus, it is desirable to find a reliable way of mitigating the effect that reverberation has on ASR.

A major stream of research designed to find a way to cope with the reverberation problem involves estimating inverse filters for the room impulse responses that characterizes the acoustic paths between the speaker and the microphones. These inverse filters are designed to cancel out the distortion caused by the room impulse responses, and generally estimated using multiple microphones [1][2]. The dereverberation methods based on inverse filtering are developed with a rigorous mathematical formulation that potentially enables us to achieve almost perfect dereverberation. Therefore, they are viewed as very attractive solutions for solving the reverberation problem. However, these methods are known to pose a sensitivity problem in that background noise or a small change in the transfer function results in severe performance degradation [2][3].

In contrast to the inverse filtering methods, robust and practical approaches have been investigated to mitigate the effect of reverberation on ASR [4][5]. These methods primarily aim at suppressing late reflections, i.e. the latter parts of the reverberation, which are the most detrimental to ASR performance [6]. The critical assumption behind these studies is that the energy density of the room impulse response may decay exponentially. Late reflections are blindly estimated based on this simple model, and then removed by using Spectral Subtraction (SS) [7]. SS can be used because we can assume the target signal (direct signal and early reflections) and interference (late reflections) to be uncorrelated [8]. The remaining early reflections may not greatly affect the state-of-the-art ASR, because they can be

well handled with such techniques as Cepstral Mean Subtraction (CMS) [9]. Such dereverberation methods appear computationally simple and relatively robust to noise. However, since reverberation cannot be well-represented solely with such a simple model, i.e. an exponential decay model, the dereverberation performance is relatively poor, thus the ASR performance cannot be restored to the level of clean speech recognition.

In a previous study, we proposed a novel single channel dereverberation method that estimates the late reflections based on the concept of the inverse filtering, and performs SS to remove late reflection energy [10]. The estimation of late reflections was carried out with multi-step Linear Prediction (LP) [11]. By incorporating the mathematical formulations used in the inverse filtering into the estimation of the late reflection energy, the method is expected to suppress late reflection more effectively and keep the early reflections unchanged for subsequent CMS. Moreover, by excluding the phase information from the dereverberation operation as in [4],[5], the proposed method provides a robustness to certain errors that conventional sensitive inverse filtering methods cannot achieve. However, this 1-ch method could not appropriately handle a non-minimum phase impulse response [12], which is often the case in real reverberant environments. Consequently the performance is degraded, especially when the distance between the speaker and the microphone is relatively large.

In this paper, we extend the above-mentioned algorithm to a multi-channel scenario with a view to tackling more realistic situations, i.e. speech recorded in a reverberant room with distant microphones. By employing the multi-channel multi-step LP that can appropriately deal with non-minimum phase room impulse responses [12], we expect to estimate the late reflections more accurately, and further improve ASR performance.

2. MODEL OF OBSERVATION PROCESS

We consider the acoustic system shown in Fig 1. By denoting the observed signal as $\mathbf{x}(n)$, the Sylvester matrix of multi-channel room impulse responses as $\mathbf{H}$, and the source signal as $\mathbf{s}(n)$, the observation process can be formulated as:

$$\mathbf{x}(n) = \mathbf{H}\mathbf{s}(n), \tag{1}$$

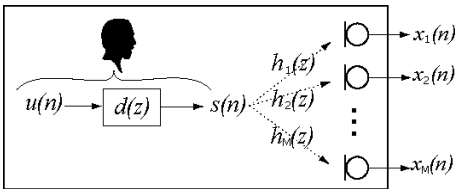

Fig. 1: Acoustic system: $s(n)$ is a source signal. $h_m(z)$ is the room transfer function between the speaker and the m-th microphone. $x_m(n)$ is a signal observed at the m-th microphone.

where

$$\begin{aligned}
\mathbf{s}(n) &= \left[s(n), s(n-1), \cdots, s(n-T-N+1))\right]^T, \\
\mathbf{x}_m(n) &= \left[x_m(n), x_m(n-1), \cdots, x_m(n-N+1)\right]^T, \\
\mathbf{x}(n) &= \left[\mathbf{x}_1^T(n), \mathbf{x}_2^T(n), \cdots, \mathbf{x}_M^T(n)\right]^T, \\
\mathbf{h}_m &= \left[h_m(0), h_m(1), \cdots, h_m(T-1)\right], \\
\mathbf{H}_m &= \begin{bmatrix} \mathbf{h}_m & 0 & \cdots & \cdots & 0 \\ 0 & \mathbf{h}_m & \ddots & & \vdots \\ \vdots & \ddots & \mathbf{h}_m & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \mathbf{h}_m \end{bmatrix}, \\
\mathbf{H} &= \left[\mathbf{H}_1^T, \mathbf{H}_2^T, \cdots, \mathbf{H}_M^T\right]^T.
\end{aligned}$$

$[\cdot]^T$ stands for the matrix transpose. The suffix m represents the signal at the m-th microphone. M is the number of microphones. The room impulse responses are assumed to be time-invariant.

3. PROPOSED DEREVERBERATION ALGORITHM

In this section, we explain the basic idea of the proposed dereverberation method. We first present our proposed algorithm for a case where $\mathbf{s}(n)$ is white noise. Then, we extend it to deal with a colored source such as speech.

3.1. White source case

3.1.1. Multi-channel multi-step linear prediction

Here, to estimate late reflections based on observed multi-channel signals, we introduce multi-channel multi-step LP [11]. Let L be the number of filter coefficients for each channel, D be the step size ($\simeq$ 30ms) [1], and M be the number of microphones, then the multi-channel multi-step LP is formulated

[1] A delay D is introduced to allow us to disregard the influence of the early reflections [10]

as follows:

$$x_i(n) = \sum_{m=1}^{M}\sum_{p=1}^{L} w_{m,i}(p)x_m(n-p-D) + e_i(n), \quad (i = 1, 2 \cdots M) \tag{2}$$

where $w_{m,i}(n)$ represents the prediction coefficients at the m-th microphone when we use the observed signal at the i-th microphone $x_i(n)$ as the prediction target, and $e_i(n)$ as the prediction error. The multi-channel multi-step LP calculates the late reflection component within $x_i(n)$. In this algorithm, we assume that the room transfer functions share no common zeros. The prediction coefficients $w_{m,i}(n)$ can be estimated by minimizing the mean square energy of the prediction error $e_i(n)$. Using a matrix notation, the minimization of the mean square energy of $e_i(n)$ leads to the following equation:

$$\underbrace{\mathrm{E}\{\mathbf{x}(n-D)\mathbf{x}^T(n-D)\}}_{\mathbf{R}}\mathbf{w}_i = \underbrace{\mathrm{E}\{\mathbf{x}(n-D)x_i(n)\}}_{\mathbf{r}_i}, \tag{3}$$

where

$$\mathbf{w}_i = [w_{1,i}(1), \cdots, w_{1,i}(L-1), \cdots w_{M,i}(1), \cdots, w_{M,i}(L-1)],$$

$\mathrm{E}\{\cdot\}$ represents the time average. $\mathbf{R}$ corresponds to the correlation matrix of the observed signal and $\mathbf{r}_i$ to the correlation vector. Thus, $\mathbf{w}_i$ can be obtained as:

$$\mathbf{w}_i = \mathbf{R}^+\mathbf{r}_i, \tag{4}$$

$(\cdot)^+$ indicates the Moore-Penrose pseudo-inverse. We use the algorithm proposed in [13] to solve eq. (4) efficiently.

3.1.2. Late reflection estimation and removal

The late reflection component at the i-th microphone, $\mathbf{l}_i(n)$, can be estimated by filtering the observed signals $\mathbf{x}(n)$ with the prediction filter $\mathbf{w}_i$ as:

$$\mathbf{l}_i(n) = \mathbf{x}^T(n)\mathbf{w}_i. \tag{5}$$

Then we subtract the estimated late reflection $\mathbf{l}_i(n)$ from the observed signal at the i-th microphone in the power spectral domain to obtain the estimate of clean speech, $\hat{\mathbf{s}}$ as:

$$|\mathcal{F}[\hat{\mathbf{s}}(n)]|^2 = |\mathcal{F}[\mathbf{x}_i(n)]|^2 - |\mathcal{F}[\mathbf{l}_i(n)]|^2, \tag{6}$$

where $\mathcal{F}[\cdot]$ corresponds to the short-time Fourier transform (STFT).

Now, let us examine the precision of the late reflection estimation. First, from eq. (4), we can obtain the following equation by making a similar calculation to that in [14].

$$\mathbf{w}_i = (\mathbf{H}^T)^+\mathbf{h}_{late,i}, \tag{7}$$

where

$$\mathbf{h}_{late,i} = \left[h_i(D), h_i(D+1), \cdots, h_i(T-1), 0, \cdots, 0\right]^T.$$

When $\mathbf{H}$ is full column rank, using eq. (7), we can develop eq. (5) as:

$$\begin{aligned}\mathbf{x}^T(n)\mathbf{w}_i &= \mathbf{s}^T(n)\mathbf{H}^T\mathbf{w}_i, \\ &= \mathbf{s}^T(n)\mathbf{H}^T(\mathbf{H}^T)^+\mathbf{h}_{late,i}, \quad (8)\\ &= \mathbf{s}^T(n)\mathbf{H}^T\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{h}_{late,i}, \quad (9)\\ &= \mathbf{s}^T(n)\mathbf{h}_{late,i}. \quad (10)\end{aligned}$$

Equation (10) simply indicates that the late reflections can be precisely estimated.

3.2. Colored source case

Here we consider the case of colored source such as a speech signal. Let us assume that a source signal $s(n)$ is modeled by an autoregressive (AR) process $\alpha(n)$ applied to white noise $u(n)$ as:

$$s(n) = \sum_{k=1}^{q}\alpha(k)s(n-k) + u(n), \tag{11}$$

Eq. (11) indicates that, when considering colored sources, $\mathbf{w}_i$ shown in eq. (7) would include not only the effect of the room impulse responses but also that of the source AR process. In this case, the late reflections may not be estimated accurately.

To cope with this problem, we suggest that pre-whitening should be performed before multi-step LP in order to remove the error introduced by $\alpha(n)$ in $\mathbf{w}_i$. In this paper, this pre-whitening was accomplished by using small order LP ($\simeq 20$ taps), because we assume that the order of $\alpha(n)$ is also generally about 20 taps.

In addition to the pre-whitening, the introduction of a delay D in multi-step LP also works to cope with colored source. In this paper, D is set at 360 taps, which we think is much larger than the order of $\alpha(n)$. If D is sufficiently larger than the order of $\alpha(n)$, this process hardly causes any distortion in the direct signal. This means that the original feature of the target signal would be well preserved even after the subtraction of late reflections. By incorporating pre-whitening and a delay in LP, we can also achieve good dereverberation with a colored source.

3.3. Schematic processing diagram

Figure 2 is a schematic diagram of the proposed method. Note that the duration of the input signal for this system has to be long enough for the

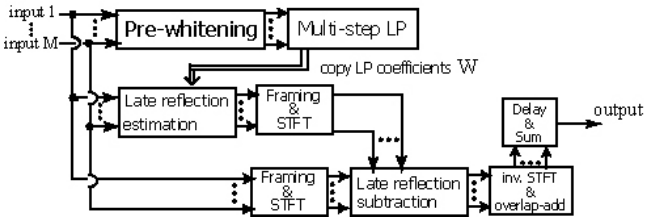

Fig. 2: Schematic diagram of multi-channel implementation

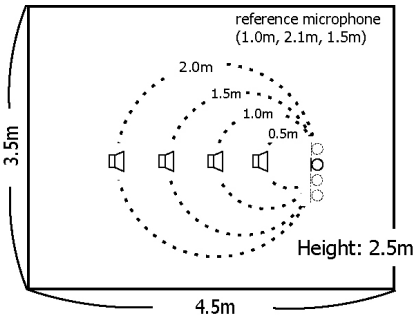

Fig. 3: Experimental setup

estimation of the prediction filter $\mathbf{w}_i$, i.e. one utterance long. First, pre-whitening is applied to the observed signals $\mathbf{x}(n)$. Using pre-whitened signals instead of observed signals in eq. (4), we calculate the prediction filter $\mathbf{w}_i$. Next, we estimate the late reflections $\mathbf{l}_i(n)$, using $\mathbf{w}_i$ and the observed signal as in eq. (5). Then we divide $\mathbf{l}_i(n)$ and $\mathbf{x}_i(n)$ into short-time frames with hamming windows, and calculate their power spectrum via STFT. Now, we subtract the power spectrum of the late reflections from that of the observed signal. The resultant spectrum is converted back to the time domain via inverse STFT (denoted as "inv. STFT" in the figure) and the overlap-add technique. To synthesize the time domain signal, we used the phase of the observed signals. This dereverberation procedure is repeated for all microphone signals ($i = 1, 2 \cdots M$). Finally, to further utilize the acoustic diversity provided by the multi-channel inputs, we incorporate the concept of the delay-sum beamformer [15]. We adjust the delays among the output signals and calculate their sum to obtain the resultant signal. The delays can be estimated, for example, based on the cross-correlation of the output signals. By employing the fast algorithm such as [16][13] for matrix inversion, whole process could work below a real-time factor of 1.

4. EVALUATION OF DEREVERBERATION PERFORMANCE IN TERMS OF ASR SCORE

In this section, we perform the dereverberation of speech recorded in a reverberant chamber, and compare the performance with the 1-ch algorithm [10].

4.1. Experimental conditions

4.1.1. Reverberation condition

Figure 3 summarizes the experimental setup. The conventional 1-ch method employed the microphone illustrated with the solid line, while the proposed method employed 3 extra microphones indicated with dotted lines. Each microphone was equally spaced at a distance of 0.2 m. We recorded the reverberant speech for 4 different speaker positions, with distances of 0.5, 1.0, 1.5 and 2.0 m between the reference microphone and the speaker. For each gender, one hundred Japanese sentences taken from JNAS database were played through a BOSE 101VM loudspeaker, and recorded using SONY ECM-77B omni-directional microphones with a sampling frequency of 12 kHz. The SNRs of the recordings were about 15 to 20 dB, and the RT_{60} reverberation time was about 0.5 sec. We applied high-pass filtering to the recordings before the dereverberation process to suppress the unwanted background noise, which was mainly concentrated below 200 Hz. After the high-pass filtering, the SNRs were about 30 dB. As a control, we also recorded the same utterances in a non-reverberant condition with a close microphone using the same experimental equipment.

4.1.2. ASR condition

The Japanese Newspaper Article Sentences (JNAS) corpus was used to investigate the effectiveness of the proposed method as a preprocessing algorithm for ASR. The ASR performance was evaluated in terms of word error rate (WER) averaged over genders and speakers. In the acoustic model, we used the following parameters : 12 order MFCCs + energy, their Δ and $\Delta\Delta$, 3 state HMMs, and 16 mixture Gaussian distributions. The model was trained on clean speech processed with CMS. The language model was a standard trigram trained on Japanese newspaper articles written over a ten-year period. The average duration of the test data was about 6 sec.

4.1.3. Parameters for dereverberation

The filter length L and the step-size D in eq. (2) for multi-step LP were 750 and 360, respectively. For the conventional 1-ch method, we used a filter length of 750 taps $\times$ 4 mics = 3000 taps. For multi-channel pre-whitening, we used a 20th order LP filter, which we calculated similarly to the approach described in [17]. No special parameters such as over-subtraction parameters or smoothing parameters were used for spectral subtraction. The length of the hamming window for DFT was 360 (=30 ms), and overlap-

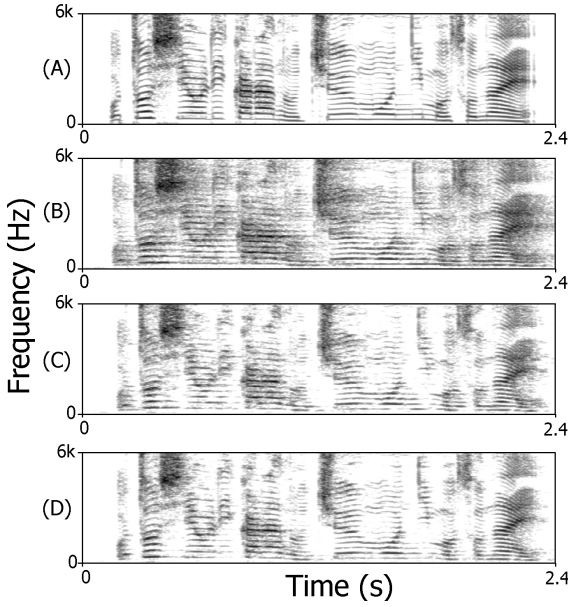

Fig. 4: Spectrograms in a reverberant environment when the distance between the microphones and speaker is set at 1.5 m: (A) clean speech, (B) recorded reverberant speech, (C) speech processed with the conventional algorithm [10], and (D) speech processed with the proposed 4-ch algorithm.

ping rate was 1/8. The dereverberation was performed utterance by utterance, which means that the length of the observed data used to estimate the LP coefficients $\mathbf{W}_i$ is equivalent to the duration of each input utterance.

4.2. Spectrogram improvement

Figure 4 shows a spectrogram of clean speech, reverberant speech recorded at a distance of 1.5 m, speech dereverberated by the conventional 1-ch algorithm, and speech dereverberated by the proposed 4-ch algorithm. All the speech is processed with CMS. We can clearly see the effect of the proposed method.

4.3. Dereverberation effect on ASR

Figure 5 shows the WER as a function of the distance between the microphone and the speaker. In the figure, "no proc." corresponds to the observed speech processed with CMS, "1 ch derev." to speech dereverberated with the conventional 1-ch algorithm, "4 ch derev." to speech dereverberated with the proposed 4-ch algorithm. In this experiment, the baseline performance was 4.9 %, which is the WER obtained with recordings made in a non-reverberant condition.

As expected, the 1-ch algorithm could not perform well when the distance between the speaker and the microphone was large, while the proposed 4-ch algorithm achieved a relatively stable performance for

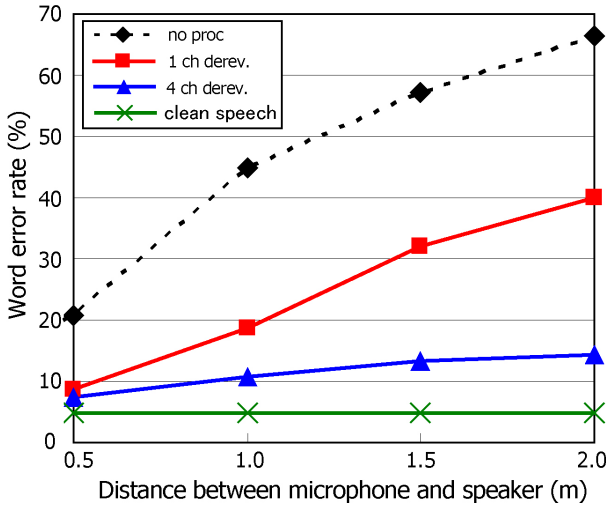

Fig. 5: Recognition experiment in reverberant environment: Recognition performance as a function of the distance between the microphone and the speaker

all reverberant conditions. The results indicate that the proposed method works well even in a severe reverberant environment.

5. SUMMARY

In this paper, we proposed a dereverberation method that primarily aims at suppressing late reflections that are found to be most detrimental to ASR performance. The proposed method first estimates late reflections using multi-channel multi-step linear prediction, and then suppresses them with subsequent spectral subtraction. Experimental results showed that the proposed method could achieve good dereverberation with an amount of observed data corresponding to the duration of one speech utterance, and could significantly improve the ASR performance even in a severe reverberant environment.

6. REFERENCES

[1] S. Gannot and M. Moonen, "Subspace methods for multi microphone speech dereverberation," *EURASIP Journal of Applied Signal Process.*, vol. 2003(11), pp. 1074–1090, 2003.

[2] G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal processing advances in wireless and mobile communications.* Upper Saddle River, NJ: Prentice Hall, 2001.

[3] B. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Trans. Speech Audio Process.*, vol. 8(3), pp. 311–319, 2000.

[4] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition," in *Proc. Hands-Free Communication and Microphone Arrays*, 2005.

[5] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 2003, pp. 844–847.

[6] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 2002, pp. 557–600.

[7] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. 27(2), pp. 113–120, 1979.

[8] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Efficient dereverberation framework for automatic speech recognition," in *Proc. Interspeech*, 2005, pp. 3145–3148.

[9] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55(6), pp. 1304–1312, 1974.

[10] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step linear prediction for single channel speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 2006, pp. 817–820.

[11] D. Gesbert and P. Duhamel, "Robust blind identification and equalization based on multistep predictors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 26(5), 1997, pp. 3621–3624.

[12] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Speech Audio Process.*, vol. 36(2), pp. 145–152, 1988.

[13] A. Varga and P. Benner, "Slicot - a subroutine library in systems and control theory," *Applied and Computational Control, Signal and Circuits*, vol. 1, pp. 499–539, 1999.

[14] M. Delcroix, T. Hikichi, and M. Miyoshi, "Blind dereverberation algorithm for speech signals based on multi-channel linear prediction," *Acoustical Science and Technology*, vol. 26(5), pp. 432–439, 2005.

[15] J. L. Flanagan, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78(11), pp. 1508–1518, 1985.

[16] D. Kressner and P. V. Dooren, "Factorizations and linear system solvers for matrices with toeplitz structure - slicot working note," TU Berlin, Tech. Rep., 2000.

[17] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. Workshop Acoust. Echo Noise Control*, vol. 1, 2003, pp. 99–102.