Blind Dereverberation Based on Spectral Subtraction by Multi-channel LMS Algorithm for Distant-talking Speech Recognition

L. Wang[1], *S. Nakagawa*[1], *N. Kitaoka*[2]

[1]Department of Information and Computer Sciences, Toyohashi University of Technology, Japan
[2]Department of Media Science, Nagoya University, Japan
{wang,nakagawa}@slp.ics.tut.ac.jp, kitaoka@nagoya-u.jp

Abstract

In this paper, we propose a blind dereverberation method based on spectral subtraction by Multi-Channel Least Mean Square (MCLMS) algorithm for distant-talking speech recognition. In a distant-talking environment, the length of channel impulse response is longer than the short-term spectral analysis window. Therefore, the channel distortion is no more of multiplicative nature in a linear spectral domain, rather it is convolutional, and conventional Cepstral Mean Normalization (CMN) is not effective to compensate for the late reverberation under these conditions. By treating the late reverberation as additive noise, a noise reduction technique based on spectral subtraction is proposed to estimate power spectrum of the clean speech using power spectra of the distorted speech and the unknown impulse responses. To estimate the power spectra of the impulse responses, a Variable Step-Size Unconstrained MCLMS (VSS-UMCLMS) algorithm for identifying the impulse responses in a time domain is extended to the spectral domain. We conducted the experiments on distorted speech signal simulated by convolving multi-channel impulse responses with clean speech. An average relative recognition error reduction of 17.8% over conventional CMN under various severe reverberant conditions was achieved using only 0.6 second speech data to estimate the spectrum of the impulse response.

Index Terms: distant-talking speech recognition, blind dereverberation, Multi-channel LMS, spectral subtraction.

1. Introduction

Hands-free speech recognition has been more and more popular in some special environments such as an office or a cabin of a car. Unfortunately, in a distant-talking environment, channel distortion may drastically degrade speech recognition performance.

Compensating an input feature is the main way to reduce a mismatch between the pratical environment and the training environment. Cepstral Mean Normalization (CMN) has been used to reduce channel distortion as a simple and effective way of normalizing the feature space [1]. In order to be effective for CMN, the length of the channel impulse response needs to be shorter than the short-term spectral analysis window. However, the duration of the impulse response of reverberation usually has a much longer tail in a distant-talking environment. Therefore, the conventional CMN is not effective under these conditions. Several studies have focused on decreasing the above problem. Raut et al. [2] used preceding states as units of preceding speech segments, and by estimating their contributions to the current state using a maximum likelihood function, they adapted the models accordingly. However, model adaptation from *a priori* training data make it less practice to use. A reverberation compensation method for speaker recognition using spectral subtraction in which the late reverberation was treated as additive noise was proposed in [3]. However, the drawback of this approach is that the optimum parameters for spectrum subtraction are empirically estimated on a development dataset and the late reverberation cannot be subtracted well since the late reverberation is not modelled precisely. In [4, 5], a novel dereverberation method utilizing multi-step forward linear prediction were proposed. They estimated the linear prediction coefficients in a time domain and suppress amplitude of late reflections using spectral subtraction in a spectral domain.

In this paper, we propose a blind reverberation reduction method based on spectral subtraction by adaptive Multi-Channel Least Mean Square (MCLMS) algorithm for distant-talking speech recognition. Speech captured by distant-talking microphones is distorted by the reverberation. With long impulse response, the spectrum of the distorted speech is approximated by convolving the spectrum of clean speech with the spectrum of impulse response. We treat the late reverberation as additive noise, and a noise reduction technique based on spectral subtraction can be easily applied to compensate for the late reverberation. By excluding the phase information from the dereverberation operation as in [6, 5], the dereveration reduction on a power spectrum domain provided a robustness to certain errors that conventional sensitive inverse filtering method could not achieve. The compensation parameter (that is, the spectrum of the impulse response) for spectral subtraction is required. In [7, 8], an adaptive MCLMS algorithm was proposed to blindly identify the channel impulse response in a time domain. In this paper, we extend this method to blindly estimate the spectrum of impulse response for the spectral subtraction in a frequency domain.

2. Dereverberation Based on Spectral Subtraction

When speech $s[t]$ is corrupted by convolutional noise $h[t]$ and additive noise $n[t]$, the observed speech $x[t]$ becomes

$$x[t] = h[t] \otimes s[t] + n[t]. \quad (1)$$

In this paper, additive noise is ignored for simplification, so Eq. (1) becomes $x[t] = h[t] \otimes s[t]$.

To analyze the effect of impulse response, the impulse response $h[t]$ can be separated into two parts $h_{early}[t]$ and $h_{late}[t]$ as [3]

$$h_{early}[t] = \begin{cases} h[t] & \text{t < T} \\ 0 & \text{otherwise} \end{cases}, h_{late}[t] = \begin{cases} h[t+T] & \text{t} \geq 0 \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where T is the length of the spectral analysis window, and $h[t] = h_{early}[t] + \delta(t-T) \otimes h_{late}[t]$. $\delta()$ is a dirac delta function (that is, a unit impulse function). The formula (1) can be rewritten as

$$x[t] = s[t] \otimes h_{early}[t] + s[t-T] \otimes h_{late}[t], \quad (3)$$

where the early effect is within a frame (analysis window), and the late effect is over multiple frames.

When the length of impulse response is much shorter than analysis window size T used for short-time Fourier transform (STFT), STFT of distorted speech equals STFT of clean speech multiply by STFT of impulse response $h[t]$ (in this case, $h[t] = h_{early}[t]$). However, when the length of impulse response is much longer than an analysis window size, STFT of distorted speech is usually approximated by

$$\begin{aligned} X(t,\omega) &\approx S(t,\omega) \otimes H(w) \\ &= S(t,\omega)H(0,\omega) + \sum_{d=1}^{D-1} S(t-d,\omega)H(d,\omega). \quad (4) \end{aligned}$$

where $H(d,\omega)$ denotes the part of $H(\omega)$ corresponding to frame delay d. That is to say, with long impulse response, the channel distortion is no more of multiplicative nature in a linear spectral domain, rather it is convolutional [2].

In [3], the early term of Eq. (3) was compensated by the conventional CMN, whereas the late term of Eq. (3) was treated as additive noise, and a noise reduction technique based on spectrum subtraction was applied as

$$\hat{S}(t,\omega) = max(X(t,w) - \alpha \cdot g(\omega)X(t-T,\omega), \beta \cdot X(t,\omega)), \quad (5)$$

where α is the noise overestimation factor, and β is the spectral floor parameter to avoid negative or underflow values. However, the drawback of this approach is that the optimum parameters α, β, and $g(\omega)$ for the spectrum subtraction is empirically estimate on a development dataset and the STFT of late effect of impulse response as the second term of the right-hand side of Eq. (4) is not straightforward subtracted since the late reverberation is not modelled precisely.

In this paper, we propose a dereverberation method based on spectral subtraction to estimate the STFT of the clean speech $\hat{S}(t,\omega)$ based on Eq. (4), and the spectrum of the impulse response for the spectral subtraction is blindly estimated using the method described in Section 3. Assuming that phases of different frames is noncorrelated for simplification, the power spectrum of Eq. (4) can be approximated as

$$|X(t,\omega)|^2 \approx |S(t,\omega)|^2|H(0,\omega)|^2 + \sum_{d=1}^{D-1} |S(t-d,\omega)|^2|H(d,\omega)|^2. \quad (6)$$

The power spectrum of clean speech $|\hat{S}(t,\omega)|^2$ can be estimated as

$$|\hat{S}(t,\omega)|^2 = \frac{max(|S(t,\omega)|^2 - \alpha \cdot \sum_{d=1}^{D-1} |\hat{S}(t-d,\omega)|^2|H(d,\omega)|^2, \beta \cdot |X(t,\omega)|^2)}{|H(0,\omega)|^2}, \quad (7)$$

where $H(d,\omega), d = 0, 1...D-1$ is the STFT of impulse response which can be calculated from known impulse response or can be blindly estimated.

3. Compensation Parameter Estimation for Spectral Subtraction by Multi-channel LMS Algorithm

3.1. Adaptive Multi-channel LMS Algorithm for Blind Channel Identification in Time Domain

In [7, 8], an adaptive multi-channel LMS algorithm for blind Single-Input Multiple-Output (SIMO) system identification was proposed.

Before introducing the MCLMS algorithm for the blind channel identification, we express what SIMO systems are *blind identifiable*. According to [9], the following two assumptions are made to guarantee an identifiable system:

1. The polynomials formed from $h_n, n = 1, 2, ..., N$ where h_n is *n-th* impulse response and N is the channel number, are co-prime, i.e., the channel transfer functions $H_n(z)$ do not share any common zeros;
2. The autocorrelation matrix $\mathbf{R}_{ss} = E\{s(k)s^T(k)\}$ of input signal is of full rank (such that the single-input multiple-output (SIMO) system can be fully excited).

In the absence of additive noise, we can take advantage of the fact that

$$x_i * h_j = s * h_i * h_j = x_j * h_i, i, j = 1, 2, ..., N, i \neq j, \quad (8)$$

and have the following relation at time k:

$$\mathbf{x}_i^T(t)\mathbf{h}_j(t) = \mathbf{x}_j^T(t)\mathbf{h}_i(t), i, j = 1, 2, ..., N, i \neq j, \quad (9)$$

where $h_i(t)$ is *i-th* impulse response at time t and

$$\mathbf{x}_n(t) = [x_n(t)\ x_n(t-1)\ ...\ x_n(t-L+1)]^T, n = 1, 2, .., N, \quad (10)$$

where $x_n(t)$ is speech signal received from *n-th* channel at time t and L is the number of taps of the impulse response.

When the estimation of channel impulse responses deviates from the true value, an error vector is produced:

$$e_{ij}(t+1) = \mathbf{x}_i^T(t+1)\mathbf{h}_j(t) - \mathbf{x}_j^T(t)\mathbf{h}_i(t), i, j = 1, 2, ..., N, i \neq j. \quad (11)$$

This error can be used to define a cost function as

$$\mathbf{J}(t+1) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \epsilon_{ij}^2(t+1) \quad (12)$$

$$\epsilon_{ij}(t+1) = \frac{e_{ij}(t+1)}{||\mathbf{h}(t)||} \quad (13)$$

$$\mathbf{h}(t) = [\mathbf{h}_1^T(t)\ \ \mathbf{h}_2^T(t)\ \ ...\ \ \mathbf{h}_N^T(t)]^T \quad (14)$$

$$\mathbf{h}_n(t) = [h_n(t,0)\ \ h_n(t,1)\ \ ...\ \ h_n(t,L-1)]^T \quad (15)$$

where $h_n(t,l)$ is *l-th* tap of *n-th* impulse response at time t.

By minimizing the cost function $\mathbf{J}(t+1)$ of Eq. (12), impulse response is blindly derived. There are various methods to minimize the cost function $\mathbf{J}(t+1)$, for example, constrained MCLMS algorithm, constrained Multi-Channel Newton (MCN) algorithm and Variable Step-Size Unconstrained (VSS-UMCLMS) algorithm and so forth [7, 8]. Among these methods, the VSS-UMCLMS achieves a nice balance between complexity and convergence speed [8]. Moreover, the VSS-UMCLMS is more practical and much easier to use since the step size does not have to be specified in advance. Therefore, in this paper, we apply VSS-UMCLMS algorithm to identify the multi-channel impulse responses. The details of the VSS-UMCLMS were described in [8].

Table 1: Detail record conditions for impulse responses measurement. "angle": recorded direction between microphone and loudspeaker. "RT60 (second)": reverberation time in room. "S": small, "L": large.

array no	array type	room	angle	RT60
1	linear	tatami-floored room (S)	120^o	0.47
2	circle	tatami-floored room (S)	120^o	0.47
3	circle	tatami-floored room (L)	130^o	0.60
4	circle	tatami-floored room (L)	90^o	0.60
5	linear	Conference room	50^o	0.78
6	linear	echo room (panel)	70^o	1.30

3.2. Extending MCLMS Algorithm to Compensation Parameter Estimation for Spectral Subtraction

To blindly estimate the compensation parameter (that is, the spectrum of impulse response), we extend the MCLMS algorithm mentioned in Section 3.1 in a time domain to a frequency domain in this section.

The spectrum of distorted signal is a convolution operation of the spectrum of clean speech and that of impulse response as shown in Eq. (4). The spectrum of the impulse response is dependent on frequency ω, and the varibale ω is omitted for simplification. Thus, in the absence of additive noise, the spectra of distorted signals have the following relation at frame t on the frequency domain:

$$\mathbf{X}_i^T(t)\mathbf{H}_j = \mathbf{X}_j^T(t)\mathbf{H}_i, \quad i,j = 1,2,...,N, \quad i \neq j, \qquad (16)$$

Where $\mathbf{X}_n(t) = [X_n(t) \quad X_n(t-1) \quad ... \quad X_n(t-D+1)]^T$ is a D-dimention vector of spectra of the distorted speech recieved from *n-th* channel at frame t, $X_n(t)$ is the spectrum of the distorted speech recieved from *n-th* channel at frame t for frequency ω, $\mathbf{H}_n = [H_n(0) \quad H_n(1) \quad ... \quad H_n(d) \quad ... \quad H_n(D-1)]^T, d = 0,1,...,D-1$ is a D-dimension vector of spectra of the impulse response, and $H_n(d)$ is the spectrum of the impulse response for frequency ω.

Using Eq. (16) in place of Eq. (9), the spectra of the impulse responses can be blindly estimated by the VSS-UMCLMS mentioned in Section 3.1.

4. Experiments

4.1. Experimental setup

Multi-channel distorted speech signals simulated by convolving multi-channel impulse responses with clean speech were used to evaluate our proposed algorithm. Six kinds of multi-channel impulse responses measured in various acoustical reverberant environments were selected from the RWCP sound scene database [10]. Four-channel circle type or linear type microphone array was taken from a circle + linear type microphone array (30 channels). A four-channel circle type microphone array has a diameter of 30 cm, and 4 microphones are located at equal 90^o intervals. Four microphones of a linear microphone array are located at 11.32 cm intervals. Impulse responses were measured at several positions which were 2 m distance from the microphone array. The sampling frequency was 48 kHz. Table 1 shows the detail record conditions for six kinds of 4 channels microphone array.

For clean speech, twenty male speakers each with a close-microphone uttered 100 isolated words. The 100 isolated words are phonetic balance common isolated words selected from Tohoku University and Panasonic isolated spoken word database [11]. The average time of all utterances was about 0.6 second. The sampling frequency was 12 kHz. The impulse responses sampled at 48 kHz were downsampling to 12 kHz to convolve with clean speech. The frame length was 21.3 ms, and the frame shift was 8 ms with a 256 point Hamming window. Then, 116 Japanese speaker-independent syllable-based HMMs (strictly speaking, mora-unit HMMs [12]) were trained using 27992 utterances read by 175 male speakers (JNAS corpus). Each continuous-density HMM had 5 states, 4 with pdfs of output probability. Each pdf consisted of 4 Gaussians with full-covariance matrices. The feature space comprised 10 MFCCs. First- and second-order derivatives of the cepstra plus first and second derivatives of the power component were also included.

4.2. Experimental results and discussion

In this paper, only the speech signal from the first channel of each microphone array was performed for speech recognition. For our proposed method, at first speech signals from multiple microphones (2 microphones or 4 microphones) were used to blindly identify the compensation parameters for the spectral subtraction (that is, the spectra of the channel impulse responses), and then the spectrum of the first channel impulse response was used to compensate for the reverberation of the speech signal from the first channel.

The number of reverberant window D in Eq. (4) was set to 8. The length of the hamming window for DFT was 256 (=21.3 ms), and the overlaping rate was 1/2. No special parameters such as over-subtraction parameters were used for spectral subtraction ($\alpha = 1$), except that the subtracted value was controlled so that it did not become negative ($\beta = 0.15$). The speech recognition performance for clean isolated words was 96.0%.

Table 2 shows the experimental results for speech recognition. CMN performed on distorted speech was used as baseline. For given impulse response, LSE (Least Square Error) based inverse filtering [13] using single channel impulse response was used to recover the reverberant speech, which is an ideal condition. In practice, for LSE base inverse filtering, it could not appropriately deal with a non-minimum phase impulse response [13], whose case is often in real reverberant environments. Therefore, the speech recognition performance was not very accurate even using the known impulse response. There are many other more precise inverse filtering techniques such as [13, 14] and so forth. We will use the more precise inverse filtering techniques as ideal condition in the future. For our proposed method, the dereverberant speech of the first channel was obtained by using the proposed reverberation compensation technique based on the spectral subtraction, and then CMN was also performed on the dereverberant speech. The proposed method remarkably improved the speech recognition performance. By using 4 microphones to estimate the spectrum of the impulse response, the improvement was less than that of 2 microphones. The more parameters needed to estimate may result in degrade performance and we are still investigating the other reasons. By using 2 microphones to estimate the spectrum of the impulse response, different results were obtained by individual 2-channel microphone array, and the results with bold font were the best results. The different results obtained from different arrays may be complementary, thus we combined these results by a so-called *maximum-summation-likelihood method (MSLM)* proposed in [15]. The *MSLM* is to use the maximum summation likelihood of recognition results from different 2-channel

Table 2: Speech recogniton performacne for reverberant speech. Only the speech data of first channel was evaluated. For proposed method, the first channel speech was compensated by the impulse response of the first channel.

distorted speech #	No processing	CMN	proposed method (2 or 4 microphones were used to estimate the spectrum of impulse response)					inverse filtering
			4 microphones	(mic1, mic2)	(mic1, mic3)	(mic1, mic4)	*MSLM*	
1	46.0	64.2	69.9	67.5	**67.5**	66.5	69.5	77.4
2	48.4	64.2	63.4	65.9	66.1	**69.0**	68.4	71.8
3	53.3	62.8	66.8	69.2	**70.2**	64.1	70.1	77.3
4	48.7	65.1	69.5	**70.5**	70.1	67.9	70.4	76.2
5	43.5	56.2	63.4	**65.2**	62.1	62.5	66.3	70.9
6	40.3	54.7	62.5	62.5	**62.7**	61.0	63.9	72.4
Ave.	46.7	61.2	65.9	**66.8**	66.5	65.2	68.1	74.3

arrays to obtain the final result. It significantly improved the speech recognition performance for all severe reverberant conditions. An average relative recogniton error reduction of 17.8% over the conventional CMN was achieved.

5. Conclusions and Future Work

In this paper, we proposed a blind reverberation reduction method based on spectral subtraction by MCLMS algorithm for distant-talking speech recognition. In a distant-talking environment, the length of channel impulse response is longer than the short-term spectral analysis window. Therefore, the channel distortion is no more of multiplicative nature in a linear spectral domain, rather it is convolutional. We treated the late reverberation as additive noise, and a noise reduction technique based on spectrum subtraction was proposed to estimate the clean power spectrum. Power spectrum of impulse respone was necessary to estimate the clean power spectrum. To estimate the power spectra of the impulse responses, a VSS-UMCLMS algorithm for identifying the impulse responses in a time domain was extended to the spectral domain. Our proposed algorithm was evaluated by distorted speech signals simulated by convolving multi-channel impulse responses with clean speech taken from Tohoku University and Panasonic isolated spoken word database. The experimental results showed that an average relative recognition error reduction of 17.8% over the conventional CMN under various severe reverberant conditions was achieved using only a isolated word (about 0.6 second) to estimate the spectrum of the impulse response.

The proposed method relies on the assumption that there are no zeros common to all channels. However, it is known that room impulse responses have a large number of zeros close to the unit circle on the z-plane. If the channels present numerically overlapping zeros, the dereverberation performance would perform poorly. [5] indicated that spatial information can be used to deal with the problem of overlapping zeros. We try to use the spatial information to deal with the same problem of overlapping zeros of our method in the future.

6. Acknowledgements

This work was partially supported by The Global COE Program "Frontiers of Intelligent Human Sensing", from the ministry of Education, Culture, Sports, Science and Technology.

7. References

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acous. Speech Signal Processing, Vol. 29, No. 2, pp. 254–272, 1981.

[2] C. Raut, T. Nishimoto, S. Sagayama, "Adaptation for long convolutional distortion by maximum likelihood based state filtering approach," Proc. of ICASSP-2006, Vol. 1, pp. 1133–1136, 2006.

[3] Q. Jin, T. Schultz and A. Waibel, "Far-field speaker recognition," IEEE Trans. ASLP, Vol. 15, No. 7, pp. 2023–2032, 2007.

[4] M. Delcroix, T. Hikichi, M. Miyoshi, "On a blind speech dereverberation using multi-channel linear prediction," IEICE Trans. Fundamentls, E89-A(10), pp. 2837–2846, October 2006.

[5] M. Delcroix, T. Hikichi, M. Miyoshi, "Precise dereverberation using multi-channel linear prediction," IEEE Trans. ASLP, 15(2), pp. 430–440, February 2007.

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

[7] Y. Huang and J. Benesty, "Adaptive multichannel least mean square and Newton algorithms for blind channel identification," Signal Processing, Vol. 82, pp. 1127-1138, Aug. 2002.

[8] Y. Huang, J. Benesty and J. Chen, "Acoustic MIMO Signal Processing", Springer, 2006.

[9] M. Xu, L. Tong and T. Kailath, "A least-squares approach to blind channel identification", IEEE Trans. Signal Processing, Vol. 43, pp. 2982–2993, Dec. 1995.

[10] http://www.slt.atr.co.jp/ tnishi/DB/micarray/indexe.htm.

[11] S. Makino, K. Niyada, Y. Mafune and K. Kido, "Tohoku University and Panasonic isolated spoken word database," Journal of the Acoustical Society of Japan, Vol. 48, No. 12, pp. 899–905, Dec. 1992. (in Japanese)

[12] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition," Proc. International Workshop on Automatic Speech Recognition and Understanding, pp. 393–396, 1999.

[13] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-36, No.2, pp. 145–152, 1988.

[14] T. Hikichi, M. Delcroix, M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," EURASIP J. APS, vol.2007, Article-ID 34013, April 2007.

[15] L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN", EURASIP J. Appl. Signal Process., Vol. 2006, Article ID 95491, pp. 1–11., 2006.