SPECTRAL SUBTRACTION STEERED BY MULTI-STEP FORWARD LINEAR PREDICTION FOR SINGLE CHANNEL SPEECH DEREVERBERATION

Keisuke Kinoshita *Tomohiro Nakatani* *Masato Miyoshi*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{kinoshita,nak,miyo}@cslab.kecl.ntt.co.jp

ABSTRACT

A speech signal captured by a distant microphone is generally smeared by reverberation, which severely degrades Automatic Speech Recognition (ASR) performance. In this paper, we propose a novel dereverberation method utilizing multi-step forward linear prediction. It precisely estimates and suppresses the late reflections, which constitute a major cause of ASR performance degradation. Our experimental results showed that the proposed method can improve ASR performance significantly even without using special adaptation methods such as multi-condition acoustic model training.

1. INTRODUCTION

A speech signal captured by a distant microphone is generally smeared by reverberation, and can be modeled as:

$$x(n) = \sum_{i=0}^{\infty} h(i)s(n-i), \tag{1}$$

where $s(n)$ refers to clean speech (source signal), and $h(n)$ to a room impulse response. Reverberation is known to degrade both Automatic Speech Recognition (ASR) performance and speech intelligibility severely. In particular, in a reverberant environment with a reverberation time (RT) of more than 0.5 seconds, the ASR performance cannot be improved even with an acoustic model trained with a matched reverberation condition [1]. Therefore, before ASR, the speech should be preprocessed with dereverberation.

Considerable research has been undertaken with a view to improving the ASR performance. Some researchers have proposed methods that attempt to estimate and equalize acoustic poles in a room sound field [2][3]. Others have proposed a method that estimates an inverse filter based on the harmonic structure of speech [4][5]. To extend the capability of the method proposed in [4][5] and make it more suitable for ASR, [6] introduced a dereverberation framework that makes extensive use of a speech property, namely sparseness. It is designed to suppress the late reflection component of reverberant speech, because it is mostly late reflections that degrade the recognition performance [7]. The experimental results showed a substantial improvement in speech recognition performance when the method was combined with the multi-condition acoustic model, if sufficiently long observed signals are provided (i.e. several tens of seconds).

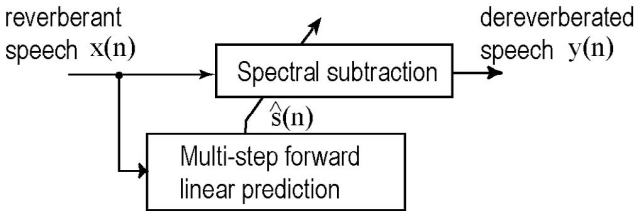

Fig. 1. Schematic diagram of proposed method

Although [6] improves the speech recognition performance, the heuristic way of estimating late reflection was not controlled by any concrete criteria to determine how much late reflection it would suppress. Certainly, it is preferable if the dereverberation method has cost functions or numerical principles to control its own performance. Moreover, since this approach relies excessively on the sparseness of the speech, it is not good at handling portions of speech where energy exists all the time.

In this paper, we propose a novel dereverberation method based on multi-step forward linear prediction that effectively suppresses the late reflection components. Unlike [6], it provides us with a numerical criterion, least mean square principle, for controling the dereverberation performance, and so may allow us to achieve more accurate dereverberation to make ASR possible even with a clean acoustic model.

2. SPECTRAL SUBTRACTION OPERATED BY MULTI-STEP FORWARD LINEAR PREDICTION

In this section, we introduce our overall dereverberation framework, and multi-step forward linear prediction, which is its essential component.

2.1. Schematic processing diagram

In Fig. 1, first the late reflection energy is estimated from an observed signal by multi-step forward linear prediction. Next, this estimated energy is used as a reference interference amplitude in the context of spectral subtraction [8], and subtracted from an observed signal to obtain the dereverberated speech. It is valid to use spectral subtraction here, if we could assume that the direct signal (target signal [1]) and late

[1]Precisely speaking, the target signal includes the direct signal as well as early reflections.

reflection (interference) are statistically independent [2]. Although early reflections remain in the dereverberated speech, we can expect that they do not affect the ASR performance, because they can be well handled with such techniques as Cepstral Mean Normalization (CMN) [9] or Maximum Likelihood Linear Regression (MLLR) [10].

In the following sections, we first introduce speech modeling and pre-whitening which allows us to predict late reflection more precisely. Next, we introduce multi-step forward linear prediction, and describe how it estimates the late reflection energy.

2.2. Speech modeling and pre-whitening

First, let us assume that source signal (speech signal) $s(n)$ is produced through a FIR filter $d(z)$ from white noise $u(n)$ as in eq. (2).

$$s(n) = \sum_{k=0}^{P} d(k)u(n-k), \tag{2}$$

where d(n) is the time-domain representation of d(z). Then, according to eq. (1), observed reverberant speech $x(n)$ can be expressed as:

$$\begin{aligned} x(n) &= \sum_{j=0}^{\infty}\sum_{k=0}^{P} h(j)d(k)u(n-j-k), \\ &= \sum_{l=0}^{\infty} g(n)u(n-l), \end{aligned}$$

where $h(n)$ corresponds to the room impulse response.

By applying an effective pre-whitening to the observed signal, we can reasonably assume $g(n) \simeq h(n)$. In this paper, hereafter, we assume that the observed signal is preprocessed with pre-whitening, thus d(n) is sufficiently equalized before the process of multi-step forward linear prediction.

2.3. Multi-step forward linear prediction

Let M be the number of filter coefficients, and D be the step-size (i.e. delay), then multi-step forward linear prediction can be formulated as follows [3].

$$x(n) = \sum_{i=1}^{M} \alpha(i)x(n-i-D) + e(n), \tag{3}$$

where $\alpha(i)$ are linear prediction (LP) coefficients, $x(n)$ is the pre-whitened observed signal, and $e(n)$ is prediction error. $\alpha(i)$ is to estimated by minimizing the mean square energy of prediction error $e(n)$. If the room impulse response is *minimum-phase*, it precisely predicts the late reflection component that arrives at a microphone D-tap later than the direct signal.

[2]The validity of using spectral subtraction to suppress late reflections is discussed in [6].

[3]When D is zero, the equation is the same as for ordinary linear prediction.

Now we describe how we estimate the late reflection energy using eq. (3) focusing particularly on the case of a *non-minimum phase* room impulse response. Hereafter, we use a Z-domain representation for simplicity. If we define $g_d(z)$ to be the direct signal and early reflection part, and $g_r(z)$ to be the late reflection part of $g(z)$ as:

$$g(z) \triangleq g_d(z) + z^{-D} g_r(z), \tag{4}$$

we can formulate the closed-form solution of $\alpha(z)$ obtained with eq. (3) as [11][12]:

$$\begin{aligned} \alpha(z) &= \frac{z^{-D} g_r(z)}{\hat{g}(z)}, \\ g(z) &\triangleq g_{min}(z) \cdot g_{max}(z), \\ \hat{g}(z) &\triangleq g_{min}(z) \cdot min[g_{max}(z)], \end{aligned}$$

where $g_{min}(z)$ and $g_{max}(z)$, respectively, stand for the minimum and maximum-phase components of $g(z)$. $min[g_{max}(z)]$ is the minimum-phase representation of $g_{max}(z)$, which is obtained by reflecting all the zeros of $g_{max}(z)$ to the inside of a unit circle on the Z-plane.

In this framework, $\alpha(z)$ is not an inverse filter for the room impulse response. However, using $\alpha(z)$ we can *predict* the late reflection as follows. Let us apply $\alpha(z)$ to the observed signal to obtain a predicted late reflection.

$$\begin{aligned} u(z) \cdot [g(z) \cdot \alpha(z)] &= u(z) \cdot \left[\frac{g(z) \cdot z^{-D} g_r(z)}{\hat{g}(z)} \right], \\ &= u(z) \cdot \left[\frac{g_{max}(z)}{min[g_{max}(z)]} \cdot z^{-D} g_r(z) \right], \qquad (5) \\ &= \hat{u}(z) \cdot z^{-D} g_r(z), \qquad (6) \end{aligned}$$

If we focus on $g_{max}(z)/min[g_{max}(z)]$ in eq. (5), the term can be clearly characterized as an all-pass filter, which is known to have a unit spectral magnitude and only introduce phase distortion into the input signal. Therefore, we see that $\hat{u}(z)$ in eq. (6), which is the product of white noise $u(z)$ and all pass filter $g_{max}(z)/min[g_{max}(z)]$, represents the white noise characteristics. Consequently, eq. (6) indicates that the magnitude of late reflection is precisely predicted, while its phase information is totally contaminated.

It is interesting that conversely we can also say that, even if the room impulse response is non-minimum phase, eq. (6) indicates that it is possible to recover the late reflection *amplitude* [4]. That is, the result of eq. (6) can be directly used as a reference signal for the type of algorithm that only requires the *amplitude* information such as spectral subtraction. In fact, as in [6], since it is theoretically valid to use spectral subtraction to suppress the late reflections, we propose using the result of eq. (6) as a reference *amplitude* of late reflections for spectral subtraction, as in eq. (7).

$$|u(z)g_d(z)| \simeq |u(z)g(z)| - |\hat{u}(z) \cdot z^{-D} g_r(z)| \tag{7}$$

[4]In a real situation, some degree of error will be introduced by the correlation of speech contained in $g_r(z)$.

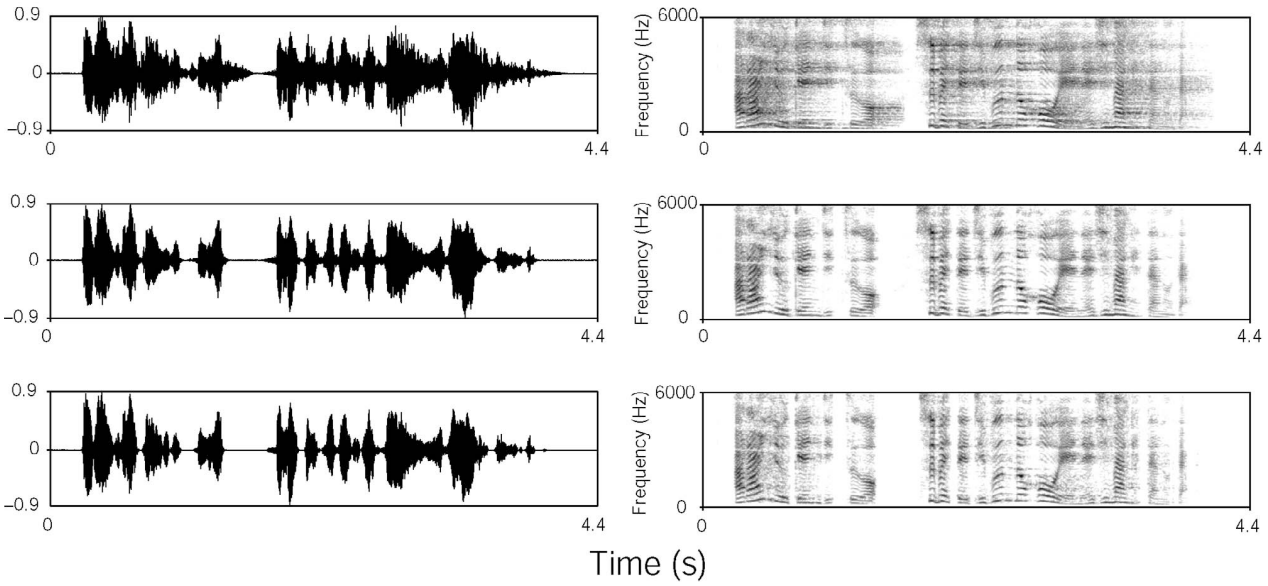

Fig. 2. Waveform and spectrogram of reverberant speech (top), artificially synthesized dereverberated speech as a ideal case (middle), dereverberated speech obtained with the proposed method (bottom)

2.4. Interpretation of overall dereverberation framework

- The dereverberation framework is essentially based on the concept of inverse filtering, except that it is designed to ignore the phase information. In other words, this framework can be seen as a precise inverse filtering method for amplitude information that is the essential speech feature for ASR. By sacrificing the phase information, the dereverberation might find a degree of robustness that conventional inverse filtering methods could not achieve.

- In contrast with [6], the proposed method explicitly provides us with a concrete numerical principle, namely the least mean square principle embedded in the linear prediction, and it always guarantees the dereverberation performance.

3. EXPERIMENTS

In this section, we evaluate the effectiveness of the proposed method by means of waveforms, spectrograms and ASR scores.

3.1. Waveform and spectrogram improvement

3.1.1. Experimental conditions

One spoken Japanese sentence was obtained from ATR data set B as the training data for the proposed method. The signals were sampled at 12 kHz and quantized with 16-bit resolution. To simulate a reverberant environment, the sentence was convolved with a 3000-tap artificial impulse response generated with a random sequence. The total duration of the reverberant sentence was about 4.5 sec.

The experimentally determined filter length M and the step-size D in eq. (3) for multi-step forward linear prediction were 5000 and 300, respectively. We employed CMN as pre-whitening before applying multi-step forward linear prediction. The window length of CMN was 300. No special parameters were used for spectral subtraction, except that the subtracted value was controlled so that it did not become negative.

3.1.2. Results

Figure 2 shows the waveform and the spectrogram of each speech. For comparison, in the middle of Fig. 2, we have included an artificially synthesized dereverberated speech, which was generated with a correct late reflection component. We can clearly see the effect of the proposed method in both the waveform and the spectrogram.

3.2. Dereverberation effect on ASR

3.2.1. Experimental conditions

We investigated the effectiveness of the proposed method as a preprocessing algorithm for ASR, using the Japanese Newspaper Article Sentences (JNAS) corpus. The ASR performance was evaluated in terms of word accuracy. In the acoustic model, we used the following parameters : 12 order MFCCs + energy, their delta and delta-delta, 3 state HMMs, and 6 mixture Gaussian distributions. We prepared two kinds of acoustic models. The model trained on clean speech processed with CMN is referred to as "CMN model", while the one trained on clean speech processed with CMN and proposed method is referred to as "CMN+derev model." CMN model was used to recognize speech processed only with CMN,

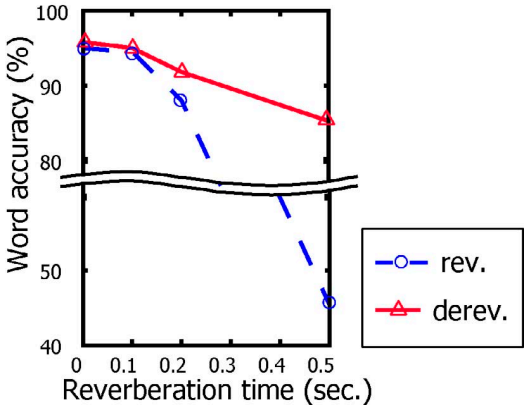

Fig. 3. Recognition performance as a function of the reverberation time. Word accuracy is improved by [0.7 0.9 3.3 39.8] at reverberation time of 0.0, 0.1, 0.2 and 0.5 sec., respectively.

while CMN+derev model was for the dereverberated speech. The language model was standard trigram trained on Japanese newspaper articles written over a ten-year period. The training and test set for the recognition task is summarized in table 1.

Table 1. Training and test set for speaker-independent acoustic model

Training	20103 utterances, 33 hours (131 speakers)
Test	100 utterances, 1578 words (22 speakers)

Reverberant speech was simulated by convolving clean speech with each of four impulse responses (Reverberation time: 0.0, 0.1, 0.2, 0.5) that were measured in a reverberant room in advance. The parameters for the dereverberation procedure were the same as for the previous experiment, The average duration of the training data for dereverberation was about 6 sec.

3.2.2. Results

Figure 3 shows the average word accuracy obtained with each recognition target. The recognition of reverberant and dereverberated speech is indicated as "rev." and "derev." respectively. The results revealed a substantial improvement of ASR performance even in severely reverberant environment. Note that the proposed method does not degrade ASR performance even when it is applied to the clean speech.

4. CONCLUSION

A speech signal captured by a distant microphone is generally smeared by reverberation, which severely degrades the Automatic Speech Recognition (ASR) performance. In this paper, we propose a novel dereverberation method that utilizes multi-step forward linear prediction. It precisely estimates the amplitude of late reflections, and suppresses them with a subsequent spectral subtraction. Our experimental results showed that the proposed method can achieve excellent dereverberation that can significantly improve the ASR performance.

5. REFERENCES

[1] B. Kingsbury, N. Morgan, "Recognizing reverberant speech with Rasta-Plp," *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1259-1262, 1997

[2] T. Hikichi and M. Miyoshi, "Blind algorithm for calculating common poles based on linear prediction," *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 89-92, 2004.

[3] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 476-488, 2003.

[4] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 92–95, 2003

[5] K. Kinoshita, T. Nakatani and M. Miyoshi, "Fast estimation of a precise dereverberation filter based on speech harmonicity," *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 1073-1076, 2005

[6] K. Kinoshita, T. Nakatani and M. Miyoshi, "Efficient dereverberation framework for automatic speech recognition," *Proc. of Interspeech2005*, vol. 1, pp. 92–95, 2005

[7] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol.1 pp. 557-600 2002

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27(2), pp. 113-120, 1979

[9] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of Acoustical Society of America*, vol. 55(6), pp. 1304-1312, 1974.

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[11] T. Kailath, A. H. Sayed and B. Hassibi, "Linear estimation," *Prentice Hall*, 2000.

[12] T. Hikichi, M. Delcroix and M. Miyoshi, "Blind dereverberation based on estimates of signal transmission channels without precise information of channel order," *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 1069-1072, 2005.