DISTANT-TALKING ROBUST SPEECH RECOGNITION USING LATE REFLECTION COMPONENTS OF ROOM IMPULSE RESPONSE

Randy Gomez, Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science
Nara Institute of Science and Technology , JAPAN
E-mail: {randy-g,even,sawatari,shikano}@is.naist.jp

ABSTRACT

We propose a robust and fast dereverberation technique for real-time speech recognition application. First, we effectively identify the late reflection components of the room impulse response. We use this information together with the concept of Spectral Subtraction (SS) to remove the late reflection components of the reverberant signal. In the absence of the clean speech in actual scenario, approximation is carried out in estimating the late reflection where the estimation error is corrected through multi-band SS. The multi-band coefficients are optimized during offline training and used in the actual online dereverberation. The proposed method performs better and faster than the relevant approach using Multi-LPC and reverberant matched model. Moreover the proposed method is robust to speaker and microphone locations.

Index Terms— Robustness, Speech Recognition, Dereverberation, Spectral Subtraction

1. INTRODUCTION

Reverberation degrades significantly the performance of distant-talking speech recognition applications. Thus, it is important to suppress the reverberation effects to minimize model mismatch prior to input to the recognizer. Technique such as inverse filtering is effective but take much computation time and precludes real-time application. In this research, we focus on a single channel real-time dereverberation framework which can be easily extended to multiple channels.

A novel approach based on this framework is proposed in [1]. This approach employs a numerical criterion based on minimum squared error through multi-step Linear Prediction Coefficients (LPC) to effectively estimate the late reflection and makes use of single-band SS to remove it from the observed signal. Although [1] works well in estimating the late reflection, this approach requires the complete reverberant utterance for processing since multi-step LPC's performance is directly proportional to the observed data. Thus, realtime speech recognition is difficult to realize.

In our proposed method, we extended and modified [1], resulting to a real-time dereverberation in realistic reverberant conditions. Instead of using multi-LPC, we devise an approach to effectively estimate the late reflection using the measured impulse response and suppress its effect through multi-band SS. Unlike in [1], the proposed method does not need to wait for the whole reverberant utterance to start processing thus, real-time implementation is possible.

2. SPECTRAL SUBTRACTION

A reverberant speech signal contains both the effects of the early and late reflections (when referring to early reflection we include by definition the direct signal). Although there exists a strong correlation due to articulatory constraints between the speech and the effects of the reverberant environment condition (i.e. early, and late reflections) this strong correlation is lost due to articulatory movements [2]. Thus, we can write

$$x(n) = x_E(n) + x_L(n), \tag{1}$$

where $x_E(n)$, $x_L(n)$ are the uncorrelated early and late reflection components of the reverberant signal $x(n)$. Denote $s(n)$ as clean speech, and suppose that given room impulse $h(n) = [h_E h_L]$ where its early coefficients h_E and late coefficients h_L are identified in advance, Equation 1 becomes

$$x(n) = s(n) * h_E + s * h_L. \tag{2}$$

Since $x_E(n)$,$x_L(n)$ are uncorrelated to some constraint [2], we can use SS [3] to remove $x_L(n)$. The target signal $x_E(n)$ becomes

$$x_E(n) = x(n) - x_L(n). \tag{3}$$

The reasons of removing $x_L(n)$ are the following:

(1) The late reflection has lower energy compared to the early reflections.

(2) The late reflection tends to be static over time and not so sensitive with microphone-to-speaker distance as opposed to the early reflection.

(3) The late reflection falls outside the framework in which the 3-state HMM is designed to handle.

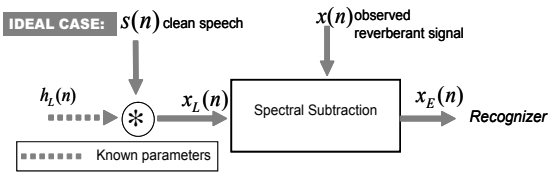

Fig. 1. Ideal dereverberation where clean speech signal is known

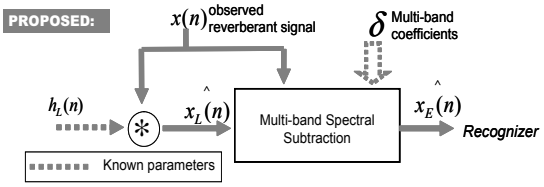

Fig. 2. Practical implementation of the ideal fast dereverberation.

Moreover, the early reverberation effects in target signal $x_E(n)$ can be handled by the 3-state HMM architecture through Cepstral Mean Normalization and adaptation techniques [4].

3. PROPOSED METHOD

Based on the SS concept given in Equation 3, a fast and simple dereverberation approach can be constructed as depicted in Figure 1. This figure shows that late reverberant components can be easily removed if late reflection impulse response h_L and clean speech $s(n)$ are given in order to estimate for $x_L(n)$. This approach is much faster and accurate than that of [1] since we use the exact impulse response boundary and not rely on multi-LPC which takes time to estimate for $x_L(n)$. Figure 1 is ideal in a sense that h_L and $s(n)$ are not available. In Figure 2, we show our proposed method which is an alternative implementation to that of the ideal case shown in Figure 1. It is possible to measure the room impulse response $h(n)$ and from that we can experimentally identify h_L which will be explained in Section 3.1. Likewise, we can assume that by using the actual reverberant signal $x(n)$ instead of $s(n)$ (note that $s(n)$ is not available) we can arrive to a crude estimate $\hat{x}_L(n)$ instead of the exact $x_L(n)$. Although this would result to significant estimation error, we can correct this through multi-band SS where multi-band coefficients $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_K\}$ are trained offline to minimize the error between $\hat{x}_L(n)$ and $x_L(n)$ as described in Section 3.3.

3.1. Identifying Impulse Response Boundary h_L

Suppose that we are able to measure the room impulse response $h(n)$ (see Section 4.1), we need to effectively find the boundary for h_L. In doing so, we varied the length of the impulse response in generating reverberant test data sets and perform recognition experiments using a clean model. The result of the experiment is shown in Figure 3, where the horizontal axis is the length of the impulse response and the vertical axis shows the recognition performance. It is obvious in this figure that the steep decrease in the performance starts at 70 ms which suggests the beginning of the effect of the late reflection $x_L(n)$. The steep decrease is attributed to the fact that the recognizer cannot deal with reverberation that falls outside the 3-state HMM framework. Thus this part is due to the late reflection h_L. Moreover, this figure shows that the recognizer is robust to the effects of the early part h_E which causes the early reflections $x_E(n)$.

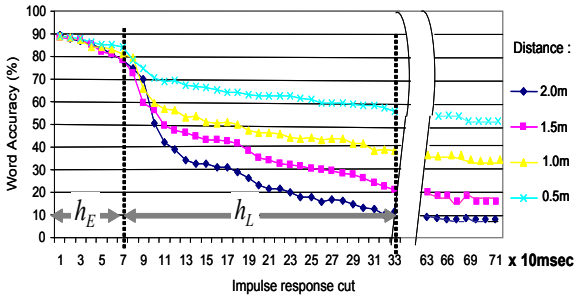

Fig. 3. Late Reflection Boundary Identification.

3.2. Estimating $\hat{x}_L(n)$ instead of $x_L(n)$

Since it is not feasible to estimate $x_L(n)$ because $s(n)$ is not available in the actual scenario, we made a crude assumption that we can instead estimate $\hat{x}_L(n) = x(n) * h_L$ using the observed reverberant signal $x(n)$ as shown in Figure 2. This assumption however, results to significant estimation error and would render the conventional single band SS to be inoperative since SS needs a good estimate of $x_L(n)$. To correct this error, we employ multi-band SS similar to that in [5]. We introduced an offline training scheme in computing the multi-band coefficients that minimize the error between $x_L(n)$ and the crude estimate $\hat{x}_L(n)$ which is discussed in Section 3.3.

3.3. Correcting Estimation Error Through Training of Multi-band Coefficients for SS

Although $s(n)$ is not available in the actual scenario, we can have access of this in the training database. Thus, we optimize the values of the multi-band coefficients offline in a form of training to minimize the error between $x_L(n)$ and $\hat{x}_L(n)$. Figure 4 shows this process. For each selected clean signal $s(n)$ in the database, the actual late reflection is $x_L(n) = h_L * s(n)$ and the crude estimate late reflection $\hat{x}_L(n) = h_L * h * s(n)$ are computed using the late part of the impulse response and the clean speech in the database. Next, the power spectral densities (PSD) $X_L(f)$ and $\hat{X}_L(f)$ of both signals are estimated using Welch's method. The window type, overlap and length of the frame are the same as the one used in the multi-band SS. Figure 5 shows an example of PSDs of both signals. For a given set of bands $\boldsymbol{B} = \{B_1, \ldots, B_K\}$, the coefficients $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_K\}$ are determined by minimizing the squared error in each band k

$$E_k = \sum_{f \in B_k} |X_L(f) - \delta_k \hat{X}_L(f)|^2. \qquad (4)$$

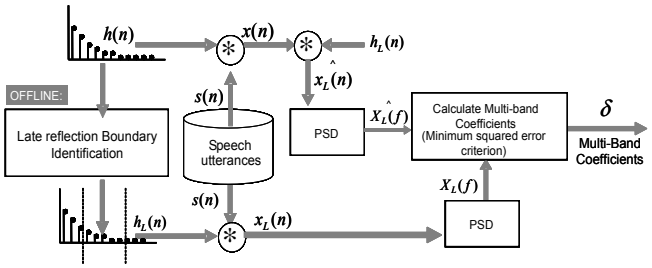

Fig. 4. Obtaining the values of the multi-band coefficients offline using the clean utterances.

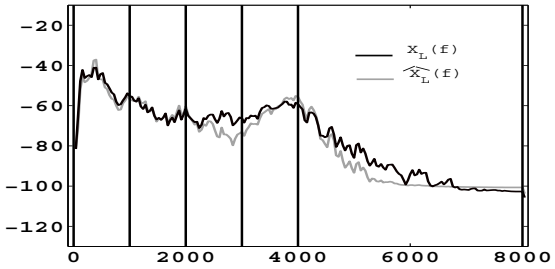

Fig. 5. Power spectral densities of the real late reverberant component $X_L(f)$ and estimated late reverberant component $\hat{X}_L(f)$.

Thus, in the actual multiband SS online using the optimized $\boldsymbol{\delta}$, the target signal $X_E(f)$ in frequency domain is given as,

$$|\hat{X}_E(f,\tau)| = \begin{cases} |X(f,\tau)|^\gamma - \delta_k|\hat{X}_L(f,\tau)|^\gamma \\ \text{if } |X(f,\tau)|^\gamma - \delta_k|\hat{X}_L(f,\tau)|^\gamma > 0 \\ \beta|\hat{X}_L(f,\tau)|^\gamma \\ \text{else} \end{cases} \quad (5)$$

for $f \in B_k$ with β the flooring coefficient and γ the power exponent as in conventional SS. We have tried different number of bands and finally choose the one used by the recognizer in obtaining the mel scale (see Section 4 Table 1). Moreover, the resulting $\boldsymbol{\delta}$ coefficients from training which is used in the actual multi-band SS are {3.430, 1.913, 1.647, 0.780, 0.664, 2.743, 2.655, 1.995, 1.699, 1.232, 1.794, 1.324} .

4. EXPERIMENTAL RESULTS

4.1. Experimental Conditions

We use the Time Stretched Pulse (TSP) method [6] to obtain the measurement of the actual room impulse response $h(n)$ and to simulate reverberant utterances for both the training and test data in the same manner as [1]. In this experiment we use a single channel directional microphone. The room set-up is shown in Figure 6 with source/speaker locations of 0.5m, 1.0m, 1.5m, and 2.0m respectively. Microphones are located with positions L2, L1, C, R2, and R1 respectively. Reverberation time of the measured impulse response is around 400ms. Reverberant signals are obtained using 6000-tap filter.

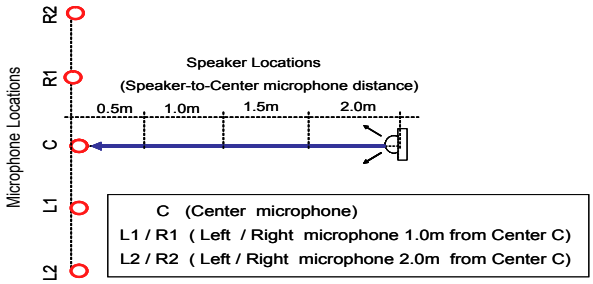

Fig. 6. Microphone-speaker set-up in acquiring room impulse response using TSP

Table 1. System specifications

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM , 2000 states
Training data	Adult and Senior by JNAS
Test data	Adult and Senior by JNAS

JULIUS [7] is used as a recognizer using Phonetically Tied Mixture (PTM) [8] model with 20K-word Japanese newspaper dictation task from JNAS [9] with a combined 561 speakers (male and female). The open test set constitutes 44 (male and female) speakers with a combined 200 utterances. Summary of the conditions used in recognition is given in Table 1.

4.2. Recognition Performance

In Figure 7 we show the basic recognition results at each speaker-to-center microphone distances 0.5m, 1.0m, 1.5m, and 2.0m. At each of this distances we also consider the 5 microphone positions R2, R1, C, L1, and L2 (refer to Figure 6 for room configuration). Figure 7 shows that the proposed method (A) outperforms the multi-LPC approach (B) in all cases. Moreover, the recognition performance improvement in using the proposed method is obvious as compared to the (C) and (D).

4.3. Robustness to Microphone Positions and Speaker Distances

A variation in speaker location would imply a variation of $\boldsymbol{\delta}$. The result shown in Figure 8 shows that the proposed method is independent of $\boldsymbol{\delta}$, thus robust to variation in location. When using only one set of $\boldsymbol{\delta}$ measured at the farthest microphone distance at 2.0m (we refer to this as robust $\boldsymbol{\delta}$), the recognition performance does not vary much as compared to using several

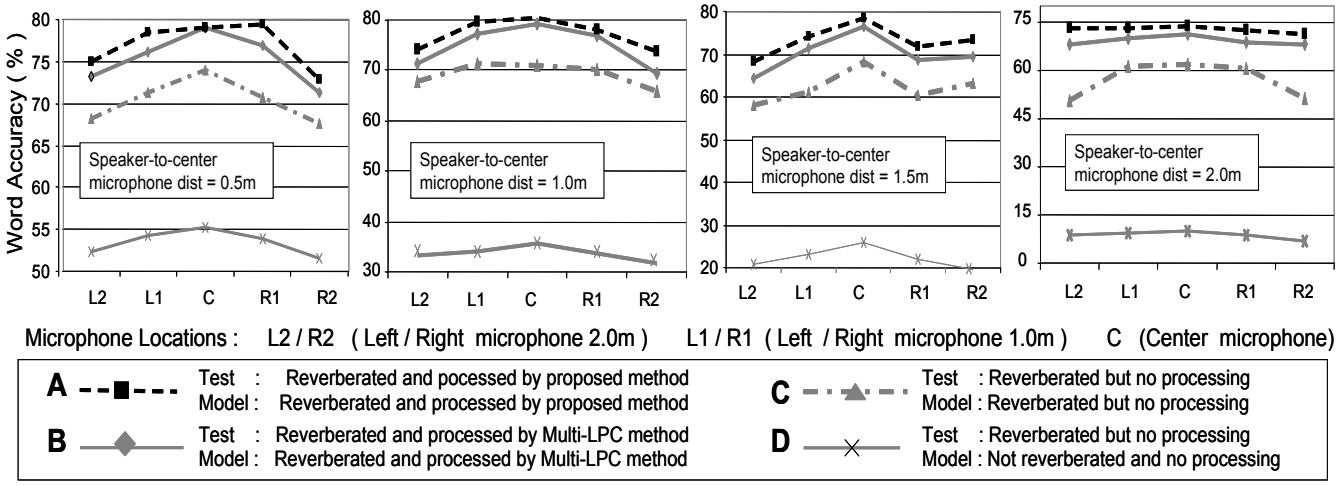

Fig. 7. Basic Recognition Performance.

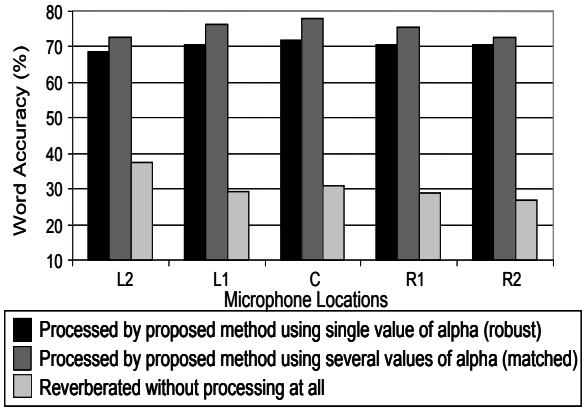

Fig. 8. Robustness of the Proposed Method.

matched $\boldsymbol{\delta}$. This points to the fact that $x_L(n)$ does not vary much as well.

5. CONCLUSION

Although the multi-LPC [1] is novel in a sense that it can adaptively estimate $x_L(n)$, real-time dereverberation for real-time speech recognition is not feasible. It is true that the proposed method requires a measurement of room impulse response in advance, but this trade-off is negligible since we are able to execute a fast and real-time dereverberation implementation which is not achieved in [1]. Moreover, since $x_L(n)$ does not vary so much with distance (as shown in Figure 8), we only need to measure a single impulse response and calculate a single set of $\boldsymbol{\delta}$. Currently we are expanding this research to using microphone arrays.

6. ACKNOWLEDGMENT

This work is supported by the Japanese MEXT e-Society project.

7. REFERENCES

[1] K. Kinoshita, T. Nakatani, and M. Miyoshi "Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation" *In Proceedings of ICASSP*, 2006

[2] K. Kinoshita, T. Nakatani, and M. Miyoshi "Efficient Dereverberation Framework For Automatic Speech Recognition" *In Proceedings of ICSLP* , Vol 1, pp 92-95, 2005

[3] S.F Boll "Suppression of Acoustic Noise in Speech using Spectral Subtraction" *IEEE Trans. on ASSP*, vol. 27(2), pp. 113–120, 1979

[4] C.J.Leggeter and Woodland "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models" *In Proceedings of Computer Speech and Language*, vol.9,pp.171-185, 1995

[5] S. Kamath, and P. Loizou "A Multi-Band Spectral Subtraction Method for enhancing Speech corrupted by colored Noise" *In Proceedings of ICASSP*, 2002

[6] Y. Suzuki, F. Asano, H.-Y. Kim, and Toshio Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses" *J. Acoust. Soc. Am. Vol.97(2), pp.-1119-1123, 1995*

[7] "Julius, an Open-Source Large Vocabulary CSR Engine - http://julius.sourceforge.jp"

[8] A. Lee, T. Kawahara, K. Takeda and K. Shikano, "A New Phonetic Tied-Mixture Model For Efficient Decoding" *In Proceedings of ICASSP* , pp. 1269-1272 2000.

[9] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research" *The Journal of Acoustical Society of Japan*, vol. 20, pp. 199-206, 1999