Robust Speech Recognition in Reverberant Environment by Optimizing Multi-band Spectral Subtraction

Randy Gomez and Tatsuya Kawahara
Kyoto University, Academic Center for Computing and Media Studies (ACCMS),
Sakyo-ku, Kyoto 606-8501, JAPAN

Abstract

Reverberant environment poses a problem in speech recognition application where performance degrades drastically depending on the extent of reverberation. Thus, it is important to employ front-end speech processing, such as dereverberation to minimize its effect. Most dereverberation techniques used to address this problem enhance the reverberant waveform prior to speech recognition. Although the speech quality is improved, this approach treats the front-end speech enhancement and the recognizer independently. In this paper, we present an approach that treats both dereverberation and speech recognition interdependently. In our proposed approach, the dereverberation parameters are optimized to improve the likelihood of the acoustic model. The system is capable of adaptively fine-tuning these parameters jointly with acoustic model training. Additional optimization is also implemented during decoding of the test utterances. Experimental results show that the proposed method significantly improves the recognition performance over the conventional approach with a relative improvement of 5%.

1 Introduction

In hands-free speech recognition applications, the observed speech signal at the microphone is smeared by a phenomenon known as reverberation. This is due to the reflection of the speech signal inside a closed space (i.e. room). The smearing varies significantly with the property and dimension of the room. The recognition performance of a reverberant test utterance using a reverberant model is significantly degraded compared to the performance of non-reverberant test utterance with a non-reverberant model. Thus, it is imperative to counter the negative effect of reverberation both the test data and the acoustic model.

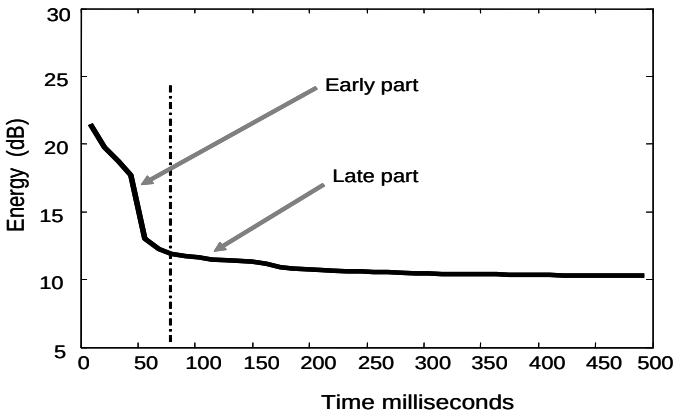

Figure 1: Measured impulse response energy.

We have proposed a single channel framework dereverberation technique based on multi-band Spectral Subtraction (SS) [1][2]. Similar approach based on single-band SS has been proposed in the work of [3]. In the multi-band SS dereverberation technique, the late reflection of the observed reverberant signal is suppressed through multi-band SS, whereas the early reverberant part (early reflection), more likely to vary with microphone-speaker distance, is handled through Cepstrum Mean Normalization (CMN) [4] [5]. The extent of suppressing the effects of the late reverberant signal is a function of the multi-band coefficients which are optimized using Minimum Mean Square Error (MMSE) criterion. Although this scheme works well, this criterion is inclined in optimizing the effect of dereverberation in the waveform level. Typically, this is a speech enhancement approach which improves the quality of the signal prior to acoustic modeling and recognition. This set-up treats the speech enhancement and recognition independently.

In this paper, we propose to treat these two interdependently by optimizing the dereverberation parameters based on the speech recognizer. Instead of just using the MMSE, we modified the criterion to directly optimize the likelihood of the recognizer. In this paper,

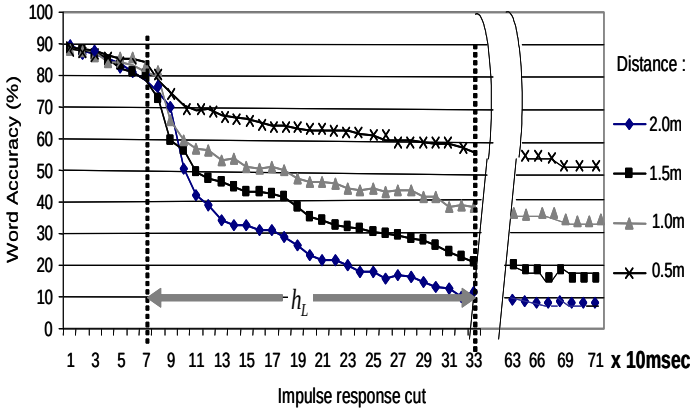

Figure 2: Late reflection boundary identification using recognition experiments and the measured room impulse response.

the optimization process of the dereverberation parameters is embedded in the acoustic model training. As a result, the dereverberation parameters are updated together with the acoustic model. This kind of approach, where front-end speech processing is optimized for recognition is shown to be effective with promising results in microphone array applications [6][7] and in Vocal Tract Length Normalization (VTLN) [8][9][10].

The organization of the paper is as follows; in section 2, we discuss the background of reverberation including its mathematical model as a function of early and late reflection. We also discuss the concept of multi-band SS based on the MMSE criterion as a dereverberation scheme. In section 3, we present the optimization in the acoustic model training phase. This involves optimization of the multi-band SS parameters based on the likelihood. In section 4, the optimization during decoding is presented. Experimental results are given in section 5, and we will conclude this paper in section 6.

2 Dereverberation Scheme

In this section, we discuss the significance of the room impulse response and its effect in the context of early and late reflection. In addition, we explain its characteristics relative to the Hidden Markov Model (HMM) structure. Consequently, we present the mathematical concept of multi-band Spectral Subtraction as a derverberation technique used in suppressing the effects of the late reflection.

2.1 Reverberation and Impulse Response

A reverberant speech signal contains the effects due to the early and late reflection. Room impulse response gives a good insight of reverberation and is often used to experimentally create a reverberant speech. When referring to the early reflection, we include by definition the direct speech signal and the overlapping of speech at earlier time. The late reflection however, is the collective overlapping of reflected speech at much later time. The following are the characteristics of the early and late reflection based on the energy plot of the measured impulse response $h(n)$shown in Figure 1:

(1) Early reflection has higher energy compared to the late reflection. Thus the speech signal in this region is dominant.

(2) Early reflection has a more dynamic value as compared to the late reflection which tend to be static over time. This characteristic implies that the effect of the late reflection can be approximately treated as constant. Since late reflection is a result of the overlapping of the speech signal in a much later time, a static energy means that as the distance between the speaker and the microphone increases, the characteristic of the late reflection remains relatively the same. Hence, a single impulse response measurement is enough to represent the different microphone-speaker locations. This treatment cannot be applied to the early reflection as its dynamic nature suggests that is sensitive to microphone-to-speaker locations.

(3) When considering a 3-state HMM architecture which has a 25 msec window and 10 msec window period, the early reflection occurs within the HMM architecture is designed to handle. Whereas, late reflection falls outside of the analysis framework.

Based on the arguments above, it is reasonable to argue that it would be beneficial to remove only the effect of late reflection through signal processing (i.e. using Spectral Subtraction) and retain the effect of the early reflection. The latter is more dependent with speaker-microphone distance, thus removing it together with the late reflection would require different impulse response measurement depending on the different microphone-speaker locations. In addition, the early reflection can be handled by the model-based system (HMM) through Cepstral Mean Normalization [4] [5].

2.2 Spectral Subtraction-based Dereverberation

In this section we outline the conventional dereverberation technique based on multi-band SS [1][2]. The speech signal has a strong correlation within each local time frame due to articulatory constraints. However, this correlation is lost according to articulatory movements [3]. As a result, it is established that early and late reflection are uncorrelated. Thus the reverberant speech signal $x(n)$ can be modeled as

$$x(n) = x_E(n) + x_L(n), \quad (1)$$

where $x_E(n)$, $x_L(n)$ are the uncorrelated early and late reflection components of the reverberant signal $x(n)$. If we denote $s(n)$ as clean speech, and the measured room impulse as $h(n) = [h_E(n), h_L(n)]$ where early components $h_E(n)$ and late components $h_L(n)$ of the whole sample $h(n)$ are identified in advance, Eq (1) can be written as,

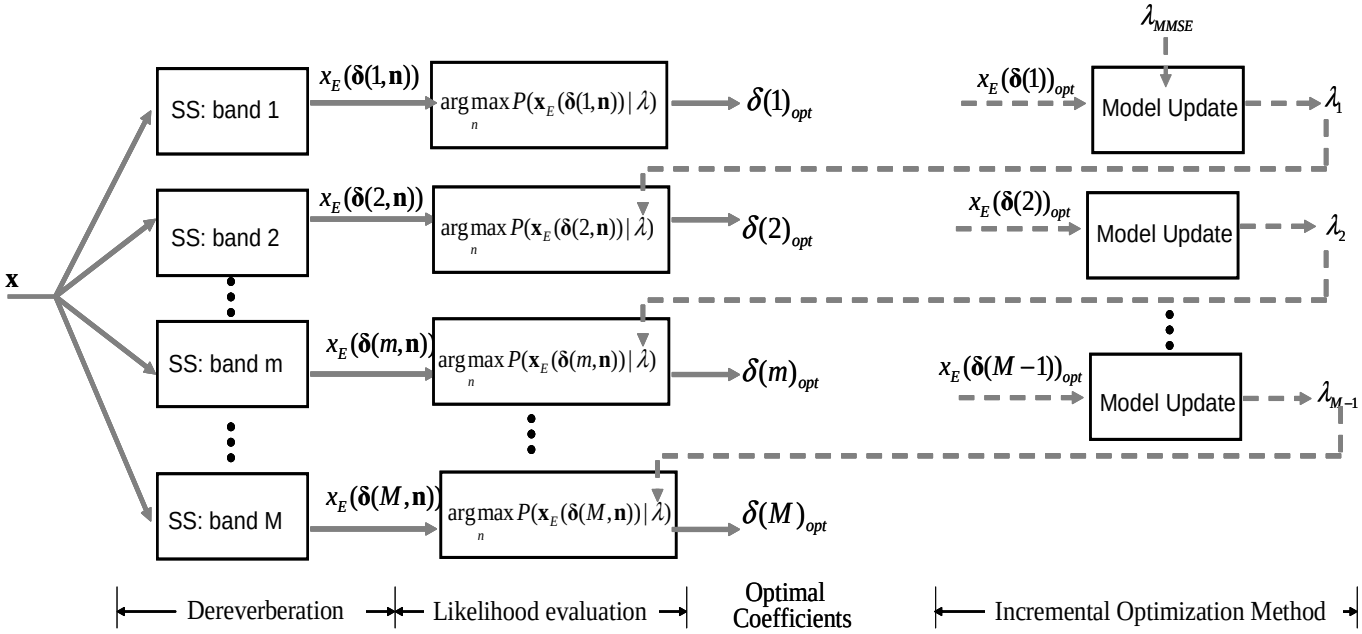

Figure 3: Block diagram of the proposed optimization technique in the acoustic training phase.

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

The boundary of the early and late reflection is very important in our model. Figure 2 is used in identifying the said boundary, where the horizontal axis represents 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. The steep decrease is attributed to the fact that the recognizer cannot deal with reverberation that fall outside of the 3-state HMM structure (i.e. caused by $x_L(n)$). Moreover, the insignificant decrease in the recognition performance within 70msec suggest that the recognizer can handle the effect due to $x_E(n)$.

In the SS-based dereverberation, we are only interested in recovering $x_E(n)$ from $x(n)$. Thus, we use spectral subtraction to remove the effect of $x_L(n)$. Theoretically, it is possible to remove entirely the effect of the whole impulse response $h(n)$, but robustness to the microphone-speaker location cannot be achieved since the early components $h_E(n)$ have high energy and is dependent on the distance between the microphone and speaker as explained in [1] [2]. In the multi-band SS approach, the effect of $x_E(n)$ is addressed through Cepstral Mean Normalization (CMN), which can be handled by the recognizer as it falls within the frame. Thus, only $x_L(n)$ is removed through the multi-band SS as its effect falls outside the frame in which the recognizer operates. The power spectra of $x_E(n)$ can be obtained through the multi-band SS,

$$|X_E(f,\tau)| = \begin{cases} |X(f,\tau)|^2 - \delta_k |X_L(f,\tau)|^2 \\ \quad \text{if } |X(f,\tau)|^2 - \delta_k |X_L(f,\tau)|^2 > 0 \\ \\ \beta |X_L(f,\tau)|^2 \quad otherwise \end{cases} \tag{3}$$

for $f \in B_k$ where B_k is the corresponding band, with β the flooring coefficient. $|X(f,\tau)|^2$ and $|X_L(f,\tau)|^2$ are the power spectra of the reverberant signal and its late reflection, respectively. The values of $\boldsymbol{\delta}$ coefficients are derived through an offline training which minimizes the error of the estimate $|X_L(f,\tau)|$ under the MMSE criterion. Details in the choice of the number of bands, the values of $\boldsymbol{\delta}$ coefficients (through offline training), and the effective identification of the late components of the impulse response $h_L(n)$ are discussed in [1] [2].

3 Optimization of Dereverberation Parameters for Acoustic Modeling

The conventional approach adopts MMSE in deriving the coefficients used in dereverberation. The derived coefficients are used to process the reverberant signal, and then the acoustic model is trained using the enhanced data. We present two methods that optimize the dereverberation parameters jointly with acoustic modeling. This principle is also applied during actual recognition which will be discussed in Section 4. The two methods are explained as follows:

3.1 Batch Optimization Method

The proposed optimization of the multi-band SS is shown in Fig. 3. We opt to optimize each band sequentially starting from the first band $m = 1$ to $m = M$. The band coefficient to be optimized is allowed to change

Table 1: System specifications

Sampling frequency	16 kHz
Window Frame length	25 ms
Window Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCCs, 12-order Δ MFCCs 1-order Δ E
HMM	8000 Gaussian pdfs
Training database	Male and Female Adult by JNAS
Test data	Male and Female Adult by JNAS

Table 2: Basic Recognition Results

Methods	200 msec	600 msec
(A) No processing	68.6 %	44.0%
(B) Conventional: MMSE	**80.1** %	**62.3**%
(C) Batch (training only)	81.3 %	64.3%
(D) Incremental (training only)	82.4 %	65.4%
(E) Batch (training/decoding)	83.1 %	66.1%
(F) Incremental (training/decoding)	**84.5** %	**67.5**%

within a close neighborhood $n\triangle$ where $n = 1...N$ and $\triangle = 0.02$. The reverberant observation data $\boldsymbol{x}$ is dereverberated using the multi-band SS. The rest of the bands are fixed to the MMSE-based estimates except for the band to be optimized. Thus, if the band to be optimized is band $m = 1$, we generate a set of coefficients $\boldsymbol{\delta}(1, n) = [\, \delta(1)_{MMSE} + n\,\triangle,\, \delta(2)_{MMSE},\, \delta(m)_{MMSE}, ...,\, \delta(M)_{MMSE}]$, and execute SS using the generated coefficients. The resulting data $x_E(\delta(1, n))$ are evaluated using the HMM-based acoustic model which is trained with data processed with MMSE-based SS parameters, denoted as $\lambda = \lambda_{MMSE}$. A Likelihood score is computed for each of the data processed with different SS conditions. Based on this result, $\delta(m)_{opt}$ that has the corresponding highest likelihood score is selected. The whole process from SS to likelihood evaluation is applied to all M bands independently. After all of the bands are optimized, the set of optimal SS coefficients $[\delta(1)_{opt}, ..., \delta(M)_{opt}]$ is used to process the reverberant data and proceed to acoustic model training. The resulting acoustic model will be used in the actual recognition.

3.2 Incremental Optimization Method

We extend the above *batch optimization method*. The additional process introduced is shown in dashed lines in Fig 3. Right after the optimal coefficient of band 1 is found, the acoustic model is re-estimated using the updated SS parameters. The newly re-estimated model λ_1 is then used in the likelihood evaluation block for band 2, and this process is iterated until $\delta(M)_{opt}$ is found for the Mth band. This approach, referred to as *incremental optimization method*, has the same principle with the *batch method*, except for the incremental updates of the HMM parameter λ in every band. In the *batch method*, we fixed $\lambda = \lambda_{MMSE}$ all throughout the bands. The incremental re-estimation allows us to treat each band interdependently in a sequential manner as opposed to the *batch optimization method* where each band is treated independently.

4 Optimal Parameter Selection During Decoding

Further optimization is implemented during actual recognition. Using the acoustic model processed with the optimal multi-band SS parameters in section 3, we evaluate a likelihood given a dereverberated test utterance. The reverberant test data are processed in the same manner as the optimization of the bands in the acoustic training phase, producing a set of processed utterances. These utterances are then evaluated with the acoustic model. The corresponding multi-band coefficient that gives the highest likelihood is selected for each band which is similar to that shown in Fig 3, and used for the final recognition. Since the dereverberation based on the multi-band SS depends on the room impulse response measurement, it is possible that the initial condition of the room impulse response used in training the model is not maintained in the actual recognition. Thus, the additional optimization during decoding is beneficial to the system in minimizing the mismatch between the actual test data and the acoustic model.

5 Experimental Evaluation

For evaluation of the proposed method, we used the training database from Japanese Newspaper Article Sentence (JNAS) corpus. The test set is composed of 200 utterances taken outside of the training database. Recognition experiments are carried out on the Japanese dictation task with 20K-word vocabulary. System specifi-

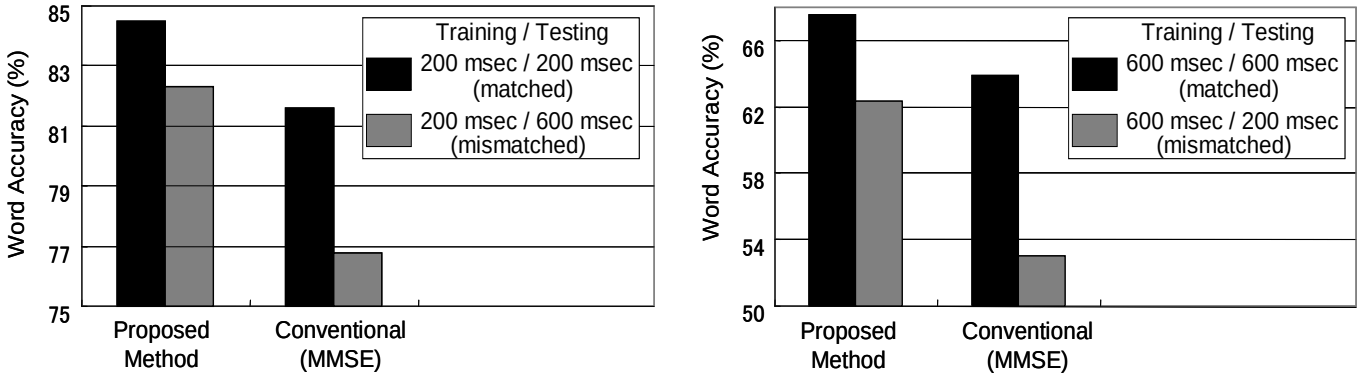

Figure 4: Test for robustness

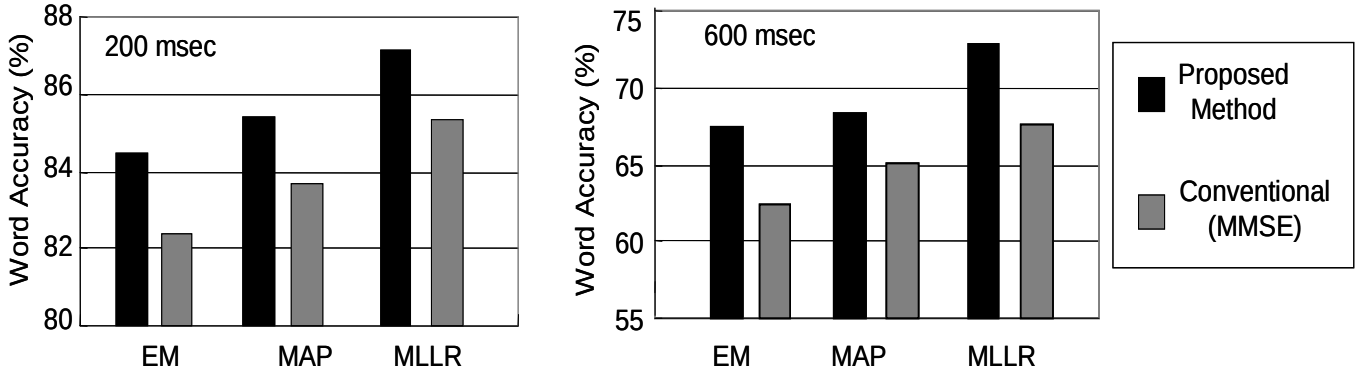

Figure 5: Performance when used in adaptation

cation is summarized in Table 1. The language model is a standard word trigram model. We experimented using two reverberant conditions: 200 msec and 600 msec. Reverberant data were made by convolving the clean database with the measured room impulse response [11]. The measured room impulse response contained flutter echo which is inherent of the actual room acoustics. In this experiment we use total number of bands $M = 5$ which is consistent to that of the former work [1][2].

5.1 Recognition Performance

Table 2 shows the basic recognition performance (word accuracy) of the proposed method in 200 msec and 600 msec reverberant conditions. (A) is the performance for reverberant test data (without dereverberation) using a clean acoustic model. (B) is for the conventional MMSE-based approach when both the test and training data are dereverberated with the conventional MMSE-based SS. (C) and (D) are the results of the proposed optimization for the batch and incremental methods, respectively. It is confirmed that the proposed front-end dereverberation optimization considering acoustic likelihood is more effective than the conventional MMSE-based method. And the incremental model update performs better than the batch training. In (E) and (F), we show that the performance of the system is further improved when optimization is also applied in the decoding process. Thus, optimizing dereverberation in both the acoustic modeling phase and decoding phase result in a synergetic effect in improving recognition accuracy. As a whole, we have achieved a relative 5% improvement over the baseline MMSE-based method.

5.2 Robustness of the Proposed Method

We also performed experiments regarding the robustness of the proposed approach. In real environment condition, it is possible that room impulse response may have considerably changed due to the additional presence/absence of physical fixtures inside the room which were absent during the measurement causing a mismatch between the acoustic model and the test data. By using different impulse responses in creating the reverberant test data and the training data, we simulate a mismatch of the reverberant condition and investigate the robustness of the proposed method as shown in Fig. 4. It is apparent that the change in the recognition performance from (matched) to (mismatched) is much smaller under the proposed method than in the conventional approach using MMSE criterion. We note that unlike the conventional method, the proposed approach is capable of optimizing the dereverberation parameters during the actual recognition which can further minimize mismatch.

5.3 Evaluation with MAP and MLLR

Then, we extend the proposed optimization technique to the adaptation scheme like MAP and MLLR. In this

case, we execute an iterative MAP and MLLR, and in each iteration we optimize the dereverberation parameters together with the 50 adaptation utterances. Recognition results shown in Figure 5 demonstrates that the proposed approach is effective in conjunction with adaptation, especially with MLLR, and the advantage over the conventional method is maintained after the adaptation.

5.4 Faster Implementation of the Proposed Optimization Technique

The proposed optimization process outlined in Fig 3 that uses HMM in evaluating the likelihood is confirmed to be effective in optimizing the dereverberation parameters. However, this process takes a lot of time and it is desirable to replicate the same performance in a shorter period of time. We try to use Gaussian Mixture Model (GMM) with 64 mixture components instead of HMM in finding the optimal parameters. A separate HMM is trained/updated only after the optimal parameters are found through GMM. This means that GMM is used for the optimization process and HMM is used for the actual speech recognition. This approach has been shown to be effective in VTLN [10].

In Fig. 6, we show the result for using both GMM and HMM in finding the optimal multi-band SS parameters. We can observe a negligible difference in word accuracy between GMM and HMM. With the GMM implementation, we reduced optimization time up to 10%. This implementation makes decoding in section 4 practical.

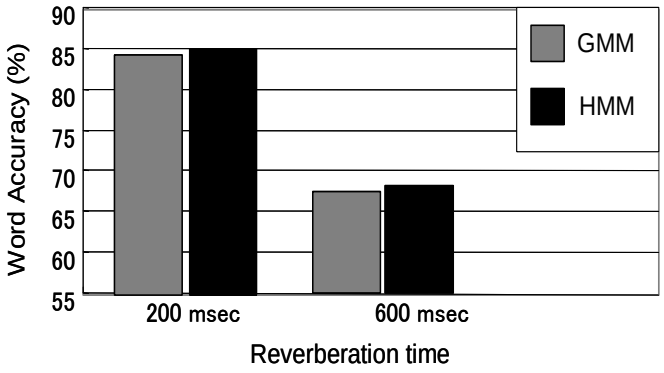

Figure 6: Performance comparison between GMM and HMM in optimizing the multi-band coefficients

6 Conclusion

We have presented the front-end dereverberation technique which is optimized based on the likelihood of the speech recognizer. The proposed is applied to the acoustic model training phase and the actual decoding phase. Both effects are confirmed, realizing significantly better performance than the conventional MMSE-based method which optimizes the parameters independent of speech recognition. We have also presented a method of speeding up the optimization process through the use of GMM which renders the decoding to be fast. In our future works, we will expand the current approach to an unknown room impulse response, where we can replace the room acoustics dependency with recognizer-based optimization in enhancing the reverberant speech signal for robust speech recognition.

References

[1] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, 2008

[2] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, 2008

[3] K. Kinoshita, T. Nakatani and M. Miyoshi, "Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, 2006

[4] A. Acero and R.M. Stern, "Environmental Robustness in Automatic Speech Recognition" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, pp 849-852 1990

[5] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition" *Kluwer Academic Publishers, Boston, MA*, 1993

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

[6] M. Seltzer, "Speech-Recognizer-Based Optimization for Microphone Array Processing" *IEEE Signal Processing Letters*, Vol. 10, No. 3, 2003

[7] M. Seltzer and R. Stern, "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments" *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 6, 2006

[8] L. Lee and R. Rose, "Speaker Normalization using Efficient Frequency Warping Procedures" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, pp 353-356, 1996

[9] D.Pye and P.C.Woodland, "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition" *In Proceeding of International Conference on Acoustics Speech and Signal Processing*, pp 1047-1050, 1997

[10] L. Welling, H. Ney, and S. Kanthak, "Speaker Adaptive Modeling by Vocal Tract Normalization" *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 10, No. 6, 2002

[11] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses" *Journal of Acoustical Society of America. Vol.97(2), pp.-1119-1123, 1995*