音声認識器の尤度を用いた残響抑圧パラメータの教師なし最適化

ゴメス・ランディ　　河原　達也

京都大学　学術情報メディアセンター
〒 606-8501 京都市左京区吉田本町

あらまし　残響下での音声認識は非常に困難なタスクである。従来の残響抑圧手法の大半が、音声認識器とは独立に音声波形を修復するものであった。これに対して本研究では、音声認識に用いる音響モデルの尤度が大きくなるように、残響抑圧のパラメータを最適化するアプローチを提案し、これをスペクトルサブトラクションに基づく方法に適用する。本手法により、残響抑圧と音響モデルの学習を統合して行うことができる。さらに音声認識（デコーディング）時にも、テストデータに対して残響抑圧パラメータの最適化を行う。大語彙連続音声認識による評価実験の結果、提案手法が従来手法に比べて、認識性能を有意に改善することを確認した。

キーワード： 音声認識、頑健性、残響抑圧

Unsupervised Optimization of Dereverberation Parameters using Likelihood of Speech Recognizer

Randy Gomez　**Tatsuya Kawahara**
School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

Abstract　Speech recognition under reverberant condition is a difficult task. Most dereverberation techniques used to address this problem enhance the reverberant waveform independent to that of the speech recognizer. In this paper, we expanded and improved the conventional Spectral Subtraction-based (SS) dereverberation technique. In our proposed approach, the multi-band SS parameters are optimized to improve the recognition performance. Moreover, the system is capable of adaptively fine-tuning these parameters in the acoustic modeling phase. Experimental results show that the proposed method significantly improves the recognition performance over the conventional approach.

Keywords: Automatic Speech Recognition, Robustness, Dereverberation

1 Introduction

Reverberation is a phenomenon caused by overlapping of signals due to reflection attributed by room environment. This degrades the performance of distant-talking speech recognition applications. Thus, it is imperative to minimize its effect. We have proposed a dereverberation approach based on multi-band Spectral Subtraction (SS) [1][2]. This method employs SS similar to that of [3] by removing only the late components of the reverberant speech signal. The multi-band coefficients 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. The criterion is modified to directly optimize the likelihood of the recognizer. In addition, we embed the optimization process 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 [4][5] and in Vocal Tract Length Normalization (VTLN) [6][7][8].

The organization of the paper is as follows; in section 2, we show the overview of the multi-band SS 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 Spectral Subtraction-based Dereverberation

In this section we outline the conventional dereverberation technique based on multi-band SS [1][2]. The reverberant speech signal is modeled as

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

where $x_E(n)$, $x_L(n)$ are the uncorrelated early and late reflection components of the reverberant signal $x(n)$. 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,

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

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,

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

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 conventiohal 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

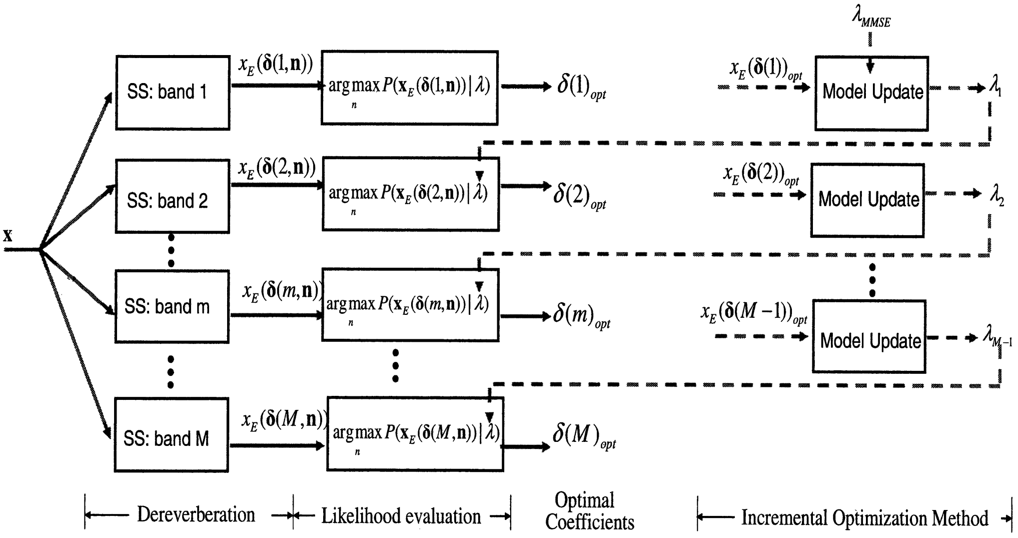
λ_{MMSE}
x
SS: band 1
SS: band 2
SS: band m
SS: band M
$x_E(\boldsymbol{\delta}(1,\mathbf{n}))$
$x_E(\boldsymbol{\delta}(2,\mathbf{n}))$
$x_E(\boldsymbol{\delta}(m,\mathbf{n}))$
$x_E(\boldsymbol{\delta}(M,\mathbf{n}))$
$\arg\max_n P(\mathbf{x}_E(\boldsymbol{\delta}(1,\mathbf{n}))|\lambda)$
$\arg\max_n P(\mathbf{x}_E(\boldsymbol{\delta}(2,\mathbf{n}))|\lambda)$
$\arg\max_n P(\mathbf{x}_E(\boldsymbol{\delta}(m,\mathbf{n}))|\lambda)$
$\arg\max_n P(\mathbf{x}_E(\boldsymbol{\delta}(M,\mathbf{n}))|\lambda)$
$\delta(1)_{opt}$
$\delta(2)_{opt}$
$\delta(m)_{opt}$
$\delta(M)_{opt}$
$x_E(\boldsymbol{\delta}(1))_{opt}$
$x_E(\boldsymbol{\delta}(2))_{opt}$
$x_E(\boldsymbol{\delta}(M-1))_{opt}$
Model Update
Model Update
Model Update
λ_1
λ_2
λ_{M-1}
Dereverberation
Likelihood evaluation
Optimal Coefficients
Incremental Optimization Method

Figure 1: Block diagram of the proposed optimization technique in the acoustic training phase which is composed of batch and incremental methods.

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. 1. 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 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 1. 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.

Table 1: System specification used in evaluating the system

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	8000 Gaussian pdfs
Training data	Adult by JNAS
Test data	Adult by JNAS

Table 2: Basic Recognition Results

Methods	200msec	600msec
(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**%

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 1, and used for the final recognition.

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. System specification is summarized in Table 1. Recognition experiments are carried out on the Japanese dictation task with 20K-word vocabulary. 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 [9]. 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

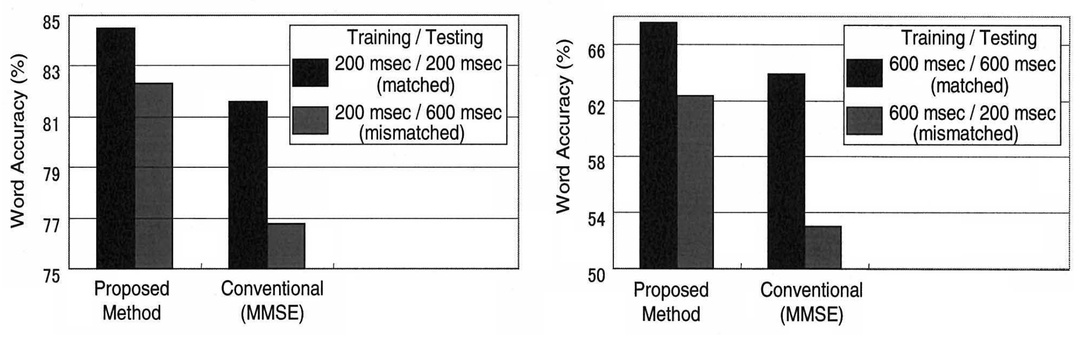

Figure 2: Test for robustness

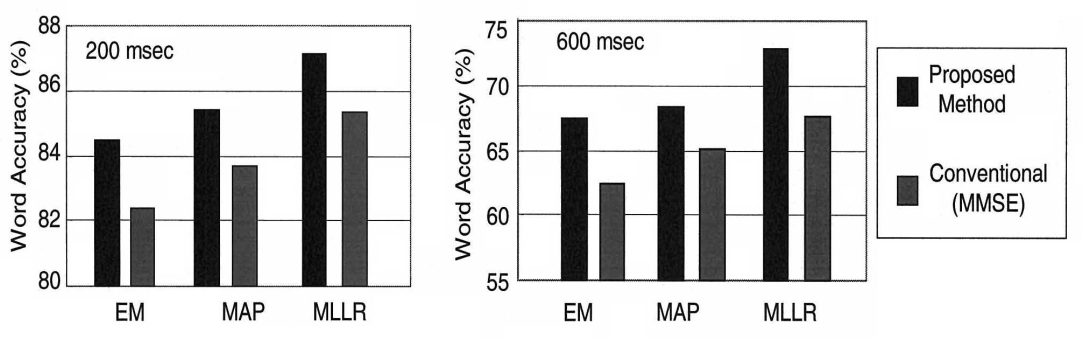

Figure 3: Performance when used in adaptation

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. By creating a mismatch of the reverberant condition between the training and testing data, we investigate the robustness of the proposed method as shown in Fig. 2. 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.

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 3 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 1 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

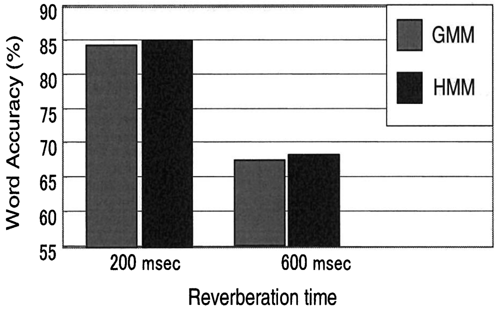

Figure 4: Performance comparison between GMM and HMM in optimizing the multi-band coefficients

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 [8].

In Fig. 4, 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.

6 Conclusion

We have presented the front-end dereverberation technique which is optimized based on the likelihood of the speech recognizer. The method is applied both in the acoustic model training phase and the actual decoding phase. In the acoustic training pahse, the dereverberation parameters are optimized using the training data. In the decoding phase, the system is able to update the dereverberation parameters based on the actual test data.. This is very important since it enables the system to adjust to the changes of the reverberant condition during the actual recognition. 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. 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. We will also attempt to remove the dependency of the current approach to room impulse response measurements.

References

[1] R. Gomez, J. Even, H. Saruwatari, and K. Shikano , "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *ICASSP*, 2008

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

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

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

[5] 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

[6] L. Lee and R. Rose, "Speaker Normalization using Efficient Frequency Warping Procedures" *ICASSP*, pp 353-356, 1996

[7] D.Pye and P.C.Woodland "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition" *ICASSP*, pp 1047-1050, 1997

[8] 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

[9] 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*