Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer

Randy Gomez and Tatsuya Kawahara

Kyoto University, ACCMS, 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 from that of the speech recognizer. In this paper, we improve the conventional Spectral Subtraction-based (SS) dereverberation technique. 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. We have evaluated using real reverberant data and experimental results show that the proposed method significantly improves the recognition performance over the conventional approach.

Index Terms: Dereverberation, Robust ASR

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][3]. This method employs SS similar to that steered by multi-step linear prediction [4] by removing only the late components of the reverberant speech signal. This approach [1][2][3] has two issues. First, the dereverberation parameters i.e. the multi-band coefficients are optimized using Minimum Mean Square Error (MMSE) criterion which 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. Secondly, it requires room impulse response (RIR) measurement which is constrained to the condition of the specific room. Although RIR measurement is effective, physical measurement is a lengthy and complicated process [5].

In this paper we address these problems by modifying the optimization criterion to directly optimize the likelihood of the recognizer instead of mere waveform enhancement. 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 in microphone arrays [6][7] and in Vocal Tract Length Normalization (VTLN) [8][9][10]. Moreover, we remove the dependency of the approach to the RIR measurement. A synthetic RIR generator which estimates the reverberation time T_{60} based on the likelihood is employed. Unlike the RIR measurement used in the conventional approach [1][2][3] requiring complicated procedures [5], the proposed RIR estimation only requires few arbitrary speech utterance spoken inside the reverberant room. This is used to estimate the RIR which is similar to that of [11]. Speech recognition experiments using real reverberant recording and synthetically generated reverberant data show improvement of recognition performance of the proposed method over the conventional MMSE approach.

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 proposed optimization of dereverberation parameters during acoustic model training followed by the fast optimization during decoding in section 4. In section 5, we discuss the experimental set-up which includes the automatic RIR generation. Experimental results are given in section 6, and we will conclude this paper in section 7.

2. Spectral Subtraction-based Dereverberation

In this section we outline the conventional dereverberation technique based on multi-band SS [1][2][3]. The reverberant speech signal is modeled as

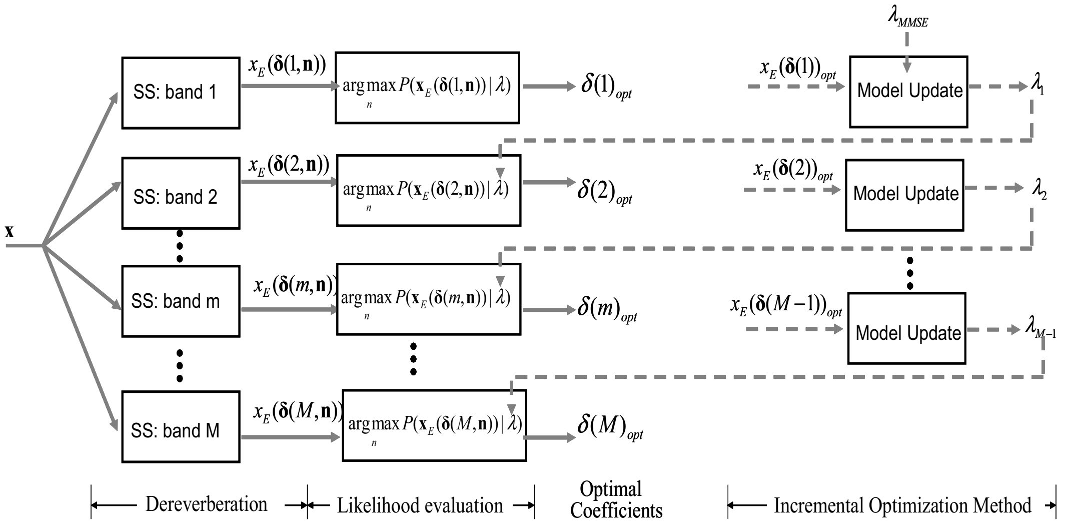
$$x(n) = x_E(n) + x_L(n), \tag{1}$$

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][3]. 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,

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



λ_{MMSE}
$\mathbf{x}$
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

for $f \in B_m$ where B_m 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].

3. Optimization of Dereverberation Parameters for Acoustic Modeling

We present two methods that optimize the dereverberation parameters jointly with acoustic modeling.

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 $n = \pm 1...N$ 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 λ_{opt} 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.

4. Fast Multi-band Dereverberation Parameter Selection during Decoding

Further optimization is implemented during actual recognition. In parallel with the acoustic model training in section 3, a Gaussian mixture model μ with 64 components is trained using the dereverberated data processed with the optimal multi-band weights. This is a text-independent model which only captures the statistical information pertaining to the optimized multi-band dereverberation parameters. The optimization starts with the dereverberation of the actual reverberant test utterance with the multi-band SS. The processed utterance is evaluated with each choice of scale parameters using μ_{rev}. Subsequently, the scale factor that leads to the best likelihood is selected and used to dereverberate the test utterance prior to input to ASR. We note that the GMM is only used for scale factor selection, and since this a very simple model as opposed to HMM, the decoding is fast and practical. As the reverberant condition is not guaranteed to be constant, the additional optimization during recognition helps minimize the mismatch between the reverberant conditions during training and testing.

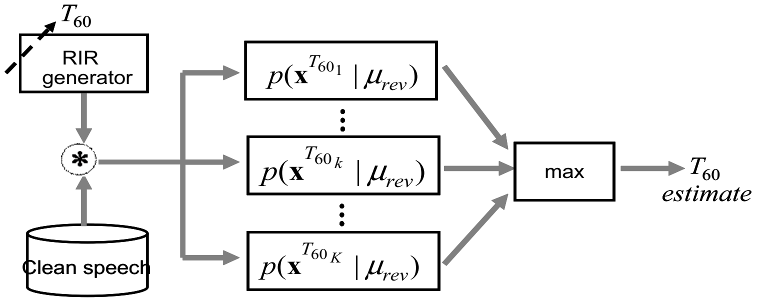

T_{60}
RIR generator
$*$
Clean speech
$p(\mathbf{x}^{T_{60_1}} \mid \mu_{rev})$
$p(\mathbf{x}^{T_{60_k}} \mid \mu_{rev})$
$p(\mathbf{x}^{T_{60_K}} \mid \mu_{rev})$
max
T_{60} estimate

Figure 2: Room impulse response approximation

Table 1: Recognition Results: **C1** Real recording, **C2** clean convolved with measured RIR, **C3** clean convolved with generated RIR

Methods	**C1** 200msec	**C1** 600msec	**C2** 200msec	**C2** 600msec	**C3** 200msec	**C3** 600msec
(A) No processing (clean model)	68.6 %	21.4 %	68.9 %	22.7 %	69.2%	23.1%
(B) No processing (reverb model)	75.4 %	32.1 %	76.4 %	35.0%	58.2%	36.5%
(C) Multi-LPC	79.2 %	51.6 %	79.8 %	52.8%	81.7%	53.2%
(D) MMSE (conventional)	**80.1** %	**54.2**%	**81.7** %	**55.9** %	**82.4** %	**64.5**%
(E) Batch (training only)	81.3 %	62.4%	82.2 %	63.2 %	82.4 %	64.1 %
(F) Incremental (training only)	82.4 %	63.7%	82.6 %	64.6 %	83.3 %	65.6 %
(G) Batch (training/decoding)	83.1 %	64.2%	83.4 %	65.8 %	84.1 %	66.2 %
(H) Incremental (training/decoding)	**84.5** %	**65.7**%	**85.0** %	**67.9** %	**85.2** %	**68.3** %

5. Experimental Set-up

5.1. Training and Testing Data

The training database is from the Japanese Newspaper Article Sentence (JNAS) corpus. The open test set is composed of 200 utterances. 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: T_{60}=200 msec and T_{60}=600 msec. Reverberant training data were made by convolving the clean database with the generated RIR discussed in this section. Two sets of reverberant test data were recorded in a room with known reverberation time : T_{60}=200 msec and T_{60}=600. Thus, we used actual reverberant data for recognition evaluation. In this experiment we use the total number of bands $M = 5$ which is consistent to that of the former work [1][2][3].

5.2. Estimating RIR Using Maximum Likelihood

The HMM represents a short speech segment with a duration of 30-100 msec. Each state captures information about a distribution of spectral parameters. With this perspective, the HMMs' description of a speech is of low resolution compared to the RIR with respect to time and frequency. Thus, for speech recognition application, it may be sufficient to use RIR estimate instead of the accurate RIR [11]. Existing studies suggest that ideally, the multiple reflections of sound can be described by a decaying acoustical energy, and the decay is best modeled by an exponential function [12]. Thus, the energy of the RIR is given as:

$$h^2(n) = e^{(6\ \ln(10)/T_{60})\ n}, \tag{4}$$

where n is the discrete time sample, and T_{60} is the reverberation time. Prior to RIR estimation, we trained a single clean GMM with 64 mixtures and adapted using arbitrary recorded reverberant utterances, resulting to the reverberant-adapted GMM μ_{rev}. In the GMM adaptation, transcriptions of the utterances are not needed. In our experiment, we used 10 utterances spoken inside the rooom. Fig. 2 shows the block diagram in approximating the RIR. First, the clean speech data are convolved with the generated RIR of variable T_{60} to generate reverberant data sets $x^{T_{60_1}}$ ...$x^{T_{60_K}}$. Then, the likelihood scores are evaluated against μ_{rev}, and the subsequent T_{60} that results to the highest likelihood score is selected.

6. Experimental Evaluation

6.1. Recognition Performance

The basic recognition performance of the proposed method is shown in Table 1. In this table, we compare the performance of the proposed method against the baselines and the conventional approach MMSE. T_{60} are 200 msec and 600 msec respectively. The test data are divided into three categories $\boldsymbol{C1}$,$\boldsymbol{C2}$ and $\boldsymbol{C3}$. In $\boldsymbol{C1}$ we use the real reverberant data, recorded in a room with known T_{60}. $\boldsymbol{C2}$ is a synthetically generated data, derived from filtering the clean utterance with measured RIR using the technique [5]. Lastly, $\boldsymbol{C3}$ is another synthetically generated reverberant test data, which is derived from filtering the clean utterance with the automatically generated RIR based on the likelihood as discussed in Section 5.

In Table 1, (A) is the performance when the reverberant test data is not processed at all (no dereverberation) using a clean acoustic model. (B) is the result when using a reverberant matched model. (C) is the performance of a different reverberation approach [4] where the processing of both the training and testing data are matched. (D) is the performance of the conventional approach when both the training and test data are dereverberated using the MMSE-based SS. In (D) the processing of both training and testing data requires actual RIR measurement. (E) and (F) are the results of the proposed optimization for the batch

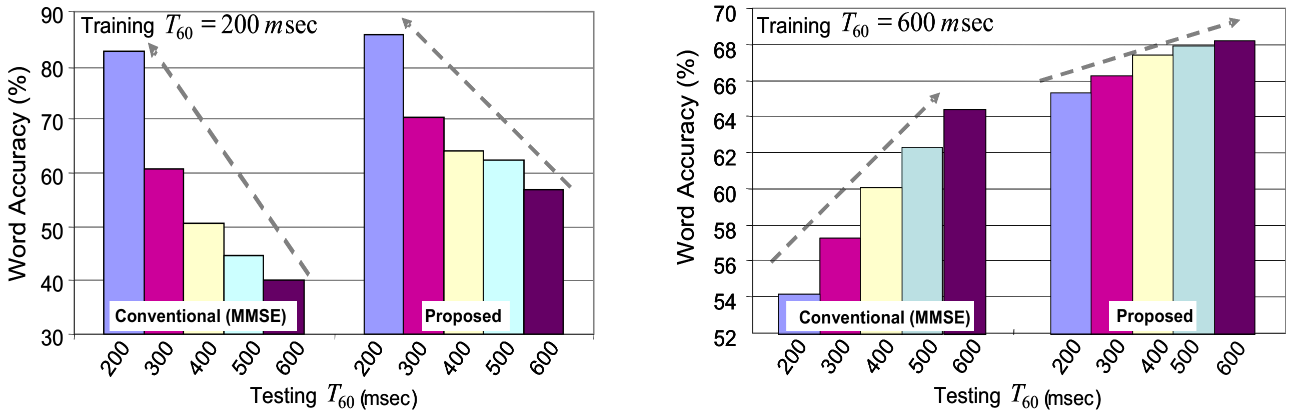


Figure 3: Comparison of the proposed method and conventional method (MMSE) in mis-matched conditions

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 (G) and (H), 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. The performance of the proposed method is consistent for both real recording and synthetic reverberant test data for all of the three categories $C1$-$C3$. We note that for the results (E)-(H) we used the automatic generation of RIR described in Section 5 as opposed to the conventional approach in which RIR was physically measured in the room [1] [2] [3].

6.2. Test for Robustness

A variation in physical arrangements inside the room can cause the reverberant condition to vary. As mentioned earlier, the reverberant condition cannot be assumed to be the same during training and testing. To investigate the robustness of the methods, we simulated a mismatch in reverberant conditions between the the training and testing data. Synthetically reverberant test sets with varying T_{60} are generated using the process described in section 5. 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 as shown in Fig. 3. Thus the proposed method is robust. Although we used a synthetic reverberant data in Fig. 3, we note that we achieved consistent performance of the proposed method when tested to both real and synthetic reverberant data in Table 1.

7. Conclusion

We have presented a 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. 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 removed the dependency of the RIR measurement used by the conventional approach. By using an arbitrary utterance spoken inside the room, we can generate the RIR based on likelihood. Moreover, the recognition experiments using both real and synthetic reverberant test data show consistent improvement in speech recognition performance.

Currently, the optimization during recognition is limited to selecting the dereverberation parameters using the test utterance. In our future works, we will expand the system to incorporate fast adaptation techniques to use the test utterance in adapting the model and not just for parameter selection.

8. References

[1] R. Gomez et.al. , "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *ICASSP*, 2008

[2] R. Gomez et.al., "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop HSCMA*, 2008

[3] R. Gomez, et.al., "Rapid Unsupervised Speaker Adaptation Robust in Reverberant Environment Conditions" *Interspeech*, 2008

[4] K. Kinoshita et.al. , "Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation" *ICASSP*, 2006

[5] Y. Suzuki, et.al., "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses"

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

[7] M. Seltzer et.al., "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments" *IEEE Trans. on Audio, Speech, and Lang. Proc.*, Vol. 14, No. 6, 2006

[8] L. Lee et.al., "Speaker Normalization using Efficient Frequency Warping Procedures" *ICASSP*, pp 353-356, 1996

[9] D.Pye et.al, "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition" *ICASSP*, pp 1047-1050, 1997

[10] L. Welling, et.al,, "Speaker Adaptive Modeling by Vocal Tract Normalization" *IEEE Trans. on Audio, Speech, and Lang. Proc.*, Vol. 10, No. 6, 2002

[11] H.-G. Hirsch et.al, "A new approach for the adaptation of HMMs to reverberation and background noise" *Speech Communication*, pp 244-263, 2008.

[12] H.Kuttruff, "Room Acoustics" *Elsevier Applied Science. Elsevier Science Publishers, 3rd edition.*, 1991