ROBUSTNESS TO SPEAKER POSITION IN DISTANT-TALKING AUTOMATIC SPEECH RECOGNITION

Randy Gomez, Keisuke Nakamura, Kazuhiro Nakadai

Honda Research Institute Japan Co., Ltd.

ABSTRACT

In this paper, we show a method that significantly improved our previous work in single-channel dereverberation. The proposed method is more robust to changes in speaker position in distant-talking ASR. First, we update the room transfer function (RTF) and weighting parameters for dereverberation to the target speaker position. This scheme corrects speech power variation as a function of position in the waveform level. Consequently, its impact to the acoustic model is verified. Then, we implement a fast acoustic model update reflective of the speech power level of the target speaker position. Furthermore, the scheme in updating the model is simple and precludes time-consuming model re-estimation. As a result, the proposed method can be executed online. The synergy of these corrective measures significantly minimizes the mismatch between training and testing conditions. We test our method using real reverberant data with different locations inside the room. Experimental results show that the proposed method outperforms the conventional methods in terms of ASR performance. Moreover, our fast acoustic model update scheme is at par in terms of recognition performance against time-consuming model re-estimation.

Index Terms— Speech Enhancement, Dereverberation, Robustness, Automatic Speech Recognition

1. INTRODUCTION

Distant-talking ASR in an enclosed reverberant environment is a difficult task. Smearing effect of reverberation causes mismatch to the clean speech. Moreover, the dynamic change of reverberation due to a change in speaker position causes a mismatch in speech power level. The combined effects of mismatch is detrimental to the ASR performance. Although the effects of reverberation can be mitigated through signal processing (i.e., dereverberation), the latter is mostly addressed by matching the acoustic model with the actual condition of the processed data (e.g. same location) through re-training or adaptation [1]. Due to multiple location points, it is not practical to re-train corresponding models. Blind dereverberation method [2][3] is a good candidate for robust dereverberation in the waveform level. However, it is not immune to model mismatch when used in ASR.

In general, distant-talking ASR application involves microphone array processing. And when the array architecture precludes the assumption of free space (i.e. circular arrays mounted on robot heads), room transfer functions (RTFs) are needed for effective localization and separation of speech. In such applications, RTFs are readily available through measurements and estimation. Thus, it is practical to utilize these RTFs for improved speech enhancement (i.e., additional input for dereverberation).

In our previous works [4][5] we addressed the problem of dereverberation by removing the effects of late reflection through spectral subtraction (SS) [4][5] and expanded to Wiener filtering [6]. Previously, dereverberation parameters are frozen during the design phase which may be different during testing. The RTF estimation scheme [7] is previously based on crude assumptions and cannot handle rooms with occlusions. In addition, [4][5] is designed for single channel application. Thus, it is very sensitive to changes in speaker location due to the volatility in speech power. Moreover, the previous work focuses only in the waveform level, void of acoustic model compensation scheme.

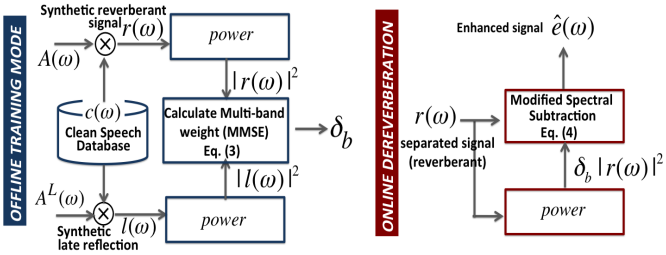

Fig. 1. Dereverberation based on spectral subtraction.

In this paper, we will show a method that takes advantage of the RTF estimation to improve dereverberation performance for distant-talking ASR. The RTF contains information concerning room properties (i.e. reflection of sounds, reverberation time, etc.). In short, it captures most of the acoustical dynamics in an enclosed environment. First, the RTF is updated online and then utilized for weighting parameters update, reflective of the target speaker's position. When used in conjunction with the dereverberation scheme in [4][5], the power level mismatch caused by the change in speaker-location is minimized in the waveform level. Second, we show a scheme for acoustic model update that links the waveform-level processing to the acoustic model, which is the heart of the ASR. Thus, reciprocating the minimization of mismatch in the acoustic model without retraining or collecting adaptation data. The organization of the paper is as follows; We introduce the background in Section 2. In Section 3, we show the method of RTF and weighting parameters update for effective dereverberation, followed by the fast acoustic model update scheme in Section 4. In Section 5, we discuss the experimental setup, together with the recognition results using real reverberant data in Section 6. We will conclude this paper in Section 7.

2. BACKGROUND

Reverberant speech model adopted in [8][9] can be expressed as

$$\begin{aligned} r(\omega) &= A^E(\omega)c(\omega) + A^L(\omega)c(\omega) \\ &= \quad e(\omega) \quad + \quad l(\omega) \end{aligned} \tag{1}$$

where $r(\omega)$ is the observed reverberant signal, $c(\omega)$ is the clean speech, $A^E(\omega)$ and $A^L(\omega)$ are the early and late reflection components of the full RTF $A(\omega)$. The boundary is experimentally predetermined in [4][5]. $r(\omega)$ can be treated as the superposition of

$e(\omega)$ and $l(\omega)$, known as the early and late reflection, respectively. In [4][5] we treat $l(\omega)$ as long-period noise which harms the ASR performance. Dereverberation is performed by suppressing $l(\omega)$ and recovering $e(\omega)$ estimate. The latter is further processed with Cepstrum Mean Normalization (CMN) during ASR. Eq. (1) simplifies dereverberation into a denoising problem, and through spectral subtraction (SS) [10], the estimate $\hat{e}(\omega)$ in frame-wise manner t is given as

$$|e(\omega,t)|^2 = \begin{cases} |r(\omega,t)|^2 - |l(\omega,t)|^2 \\ \quad \text{if } |r(\omega,t)|^2 - |l(\omega,t)|^2 > 0 \\ \\ \beta|r(\omega,t)|^2 \quad \text{otherwise.} \end{cases} \tag{2}$$

where β is the flooring coefficient. In real condition, $l(\omega,t)$ is unavailable, precluding the power estimate $|l(\omega,t)|^2$. A scheme in[4][5] shows a workaround to this problem, approximating $l(\omega,t)$ directly from the observed reverberant signal $r(\omega,t)$ through the error

$$E_m = \frac{1}{T}\sum_t \sum_{\delta_b \in B_m} |l(\omega,t) - \delta_b(\omega,t) r(\omega,t)|^2. \tag{3}$$

For the given set of bands $\boldsymbol{B} = \{B_1, \ldots, B_M\}$, the weighting parameter δ_b is determined through minimum mean square error criterion in Eq. (3) via offline training. The new estimate $\hat{e}(\omega)$ through the modified SS becomes

$$|e(\omega,t)|^2 = \begin{cases} |r(\omega,t)|^2 - \delta_b |r(\omega,t)|^2 \\ \quad \text{if } |r(\omega,t)|^2 - \delta_b |r(\omega,t)|^2 > 0 \\ \\ \beta|r(\omega,t)|^2 \quad \text{otherwise.} \end{cases} \tag{4}$$

The multi-band treatment improves error minimization as opposed to single-band [4][5]. Fig. 1 shows the adopted dereverberation scheme [4][5] (frame-wise processing is assumed, hence variable t is dropped from the figure). In the offline mode, $r(\omega)$ and the late reflection $l(\omega)$ are generated. The scheme in obtaining the RTF $A(\omega)$ and its late reflection components $A^L(\omega)$ is described in [4][5]. Then, multi-band weighting parameters δ_b are obtained through MMSE criterion (see Eq. (3)). In the actual dereverberation, δ_b obtained during offline training is used in conjunction with the modified SS in Eq. (4).

The method in [4][5] uses a pre-defined $A(\omega)$ for all positions inside the room and it suffers from robustness issues during actual run-time when speaker changes position away from the mic-array. In our proposed method, this is mitigated through a series of parameter updates for effective dereverberation in the waveform level. Furthermore, the acoustic model is also updated in conjunction with the updated waveform processing.

3. DEREVERBERATION PARAMETERS UPDATE

In Fig. 2, mic array processing using a hybrid algorithm of beamforming and blind separation called *Geometrically constrained High-order Decorrelation based Source Separation (GHDSS)* [11] [12] is employed resulting to $r(\omega)$. The rest of the processes are discussed below

3.1. RTF Curve Fitting

To simplify the RTF update through curve fitting, we assume that (a) the phase of the RTF does not change as a function of the possible speaker locations inside the room; and (b) the amplitude of the updated RTF decays exponentially as a function of distance. Let $A(\omega,d)$ denotes an arbitrary pre-measured RTF of known distance d between the mic-array and sound sources. The updated RTF $\hat{A}(\omega,\hat{d})$ at target location $\hat{d}$ is given as

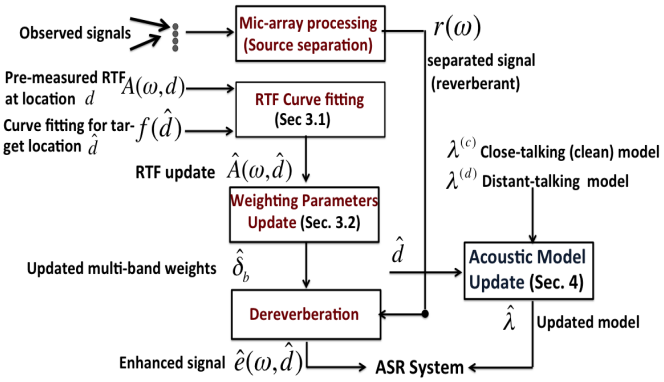

Fig. 2. Block diagram of parameters update for effective dereverberation

$$\hat{A}(\omega,\hat{d}) = f(\hat{d}) A(\omega,d)\ , \tag{5}$$

where $f(\hat{d}) \in \mathbb{R}$ is the exponential gain function of $\hat{d}$ and is obtained as a priori information based on a nonlinear curve fitting using limited number of measured RTFs. Specifically,

$$f(\hat{d}) = \frac{\alpha_1}{\hat{d}} + \alpha_2, \tag{6}$$

where α_1 and α_2 are the estimated fitting parameters. The steps for the radial distance update are as follows:

1) Pre-measurement of a limited number i_d RTFs $A(\omega, d_{[i]})$ from different $d_{[i]}$ points. These are readily available during mic-array processing (i.e. sound separation) in [11].
2) Obtain mean amplitude of RTFs over frequency bins by

$$\bar{A}(d_{[i]}) = \frac{1}{p_h - p_l + 1} \sum_{p=p_l}^{p_h} \left| A(\omega_{[p]}, d_{[i]}) \right| \ , \tag{7}$$

where p_h and p_l are the indices for maximum and minimum frequency, respectively.

3) Obtain α_1 and α_2 through nonlinear curve fitting:

$$\boldsymbol{F}_x = \begin{bmatrix} \frac{1}{d_{[1]}} & 1 \\ \vdots & \vdots \\ \frac{1}{d_{[i_d]}} & 1 \end{bmatrix}, \boldsymbol{F}_y = \begin{bmatrix} \bar{A}(d_{[1]}) \\ \vdots \\ \bar{A}(d_{[i_d]}) \end{bmatrix},$$

$$[\alpha_1, \alpha_2]^T = \left(\boldsymbol{F}_y^T \boldsymbol{F}_y \right)^{-1} \boldsymbol{F}_y^T \boldsymbol{F}_x \tag{8}$$

4) Select an arbitrary $A(\omega,d)$ from step (1). Substitute values of α_1 and α_2 in step (3) to Eq. (6) and proceed to Eq. (5).

Fig. 3 shows that both assumptions (a) and (b) above are satisfied. On the top figure, it is observed that distance update (for 9 different points d=0.5-d=2.5) only scales the amplitude of the mean power spectrum. More importantly, the phase remains unchanged (i.e. responses becomes zero at around 100th frame). In addition, the bottom figure shows an exponential decay of $f(\hat{d})$ which fits the amplitude of the measured points, confirming the validity of the nonlinear fitting.

3.2. Weighting Parameters Update

Prior to the actual weighting parameter update, several $\hat{A}(\omega,\hat{d})$ are synthetically generated for different values of $\hat{d}$ as described in Sec 3.1. Then, weighting parameters δ_b for $\boldsymbol{B} = \{B_1, \ldots, B_M\}$ are computed for each of the $\hat{A}(\omega,\hat{d})$ in the same manner as discussed in Sec. 2 (i.e., left of Fig. 1), and these values are kept in the database.

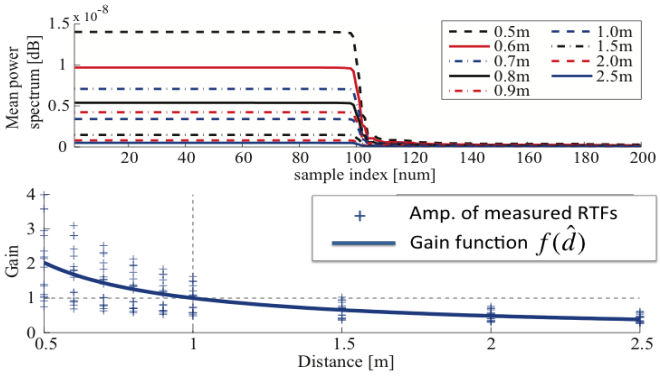

Fig. 3. Validity of the assumptions for effective RTF update.

During online testing, the RTF is updated for a given test speaker's target position $\hat{d}$. A simple waveform matching of the RTF update against the RTFs in the database is performed and the corresponding weighting parameter $\hat{\delta}_b$ of the best matched RTF is selected.

4. ASR ACOUSTIC MODEL UPDATE

Mismatch exists when using a model trained with a close-talking clean speech data $\lambda^{(c)}$ against the enhanced (dereverberated) speech for testing. Thus, model re-training or adaptation using the enhanced speech is usually implemented. Due to many possible location points, time and data issues, re-training becomes impractical. We note that mismatch caused by the change in speech power is different from the mismatch caused by speaker variation or environment conditions. The former maybe simpler to address without the need for re-estimation or adaptation. We investigate the model parameters as a function of distance between the speaker and the mic-array. Consequently, we verify the impact of the change in speech power to the model parameters.

Distant-talking training data from an arbitrary location d is enhanced (with CMN processing) and used to train the acoustic model $\lambda^{(d)}$. Upon completion of the Expectation Maximization cycle, $\lambda^{(d)}$ has the following parameters :

$$C_{im}^{(d)} = \frac{L_{im}^{(d)}}{\sum_{m=1}^{M} L_{im}^{(d)}}, \tag{9}$$

$$\boldsymbol{\mu}_{im}^{(d)} = \frac{\boldsymbol{m}_{im}^{(d)}}{L_{im}^{(d)}}, \tag{10}$$

$$\boldsymbol{\Sigma}_{im}^{(d)} = \frac{\boldsymbol{v}_{im}^{(d)}}{L_{im}^{(d)}} - \boldsymbol{\mu}_{im}^{(d)} \boldsymbol{\mu}_{im}^{(d)^T}, \tag{11}$$

$$a_{ij}^{(d)} = \frac{L_{ij}^{(d)}}{\sum_{j=1}^{J} L_{ij}^{(d)}}, \tag{12}$$

where $C_{im}^{(d)}$, $\boldsymbol{\mu}_{im}^{(d)}$, $\boldsymbol{\Sigma}_{im}^{(d)}$, and $a_{ij}^{(d)}$ are the mixture weight, mean, covariance matrix and updated transition probability respectively. m denotes the mixture while i and j signify the state (i is the current state). The statistics $L_{im}^{(d)}, L_{ij}^{(d)}, \boldsymbol{m}_{im}^{(d)}, \boldsymbol{v}_{im}^{(d)}$ are the accumulated mixture occupancy, state transition occupancy, mean statistics and variance statistics, respectively. In our investigation, we have verified that the values of $L_{im}^{(d)}, L_{ij}^{(d)}, \boldsymbol{m}_{im}^{(d)}, \boldsymbol{v}_{im}^{(d)}$ are affected by the distance d. Specifically, these parameters increase in value as the power level increases (speaker is close to the mic-array). To control these values,

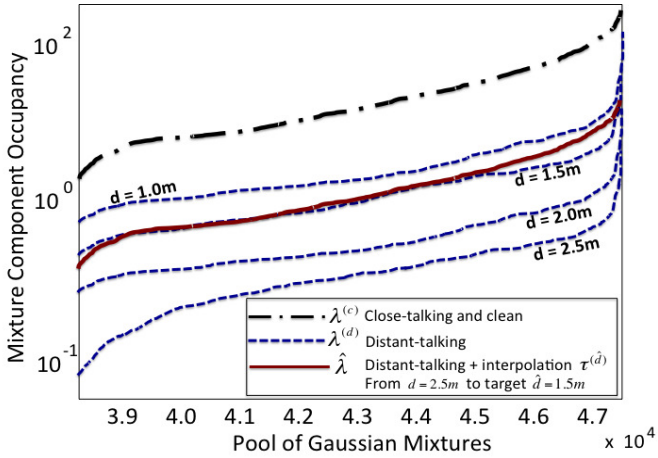

Fig. 4. Effects of distance (i.e., speech power variation) to the model.

we propose the interpolation of the statistics in $\lambda^{(c)}$, a model trained from the close-talking clean data with the statistics from $\lambda^{(d)}$. The effects of this are two-folds, first, it shifts the original values of the $L_{im}^{(d)}, L_{ij}^{(d)}, m_{im}^{(d)}, v_{im}^{(d)}$ to the desired values of the target distance $\hat{d}$ through interpolation. Second, the close-distant talking model $\lambda^{(c)}$, due to its "completeness" may fill the gaps of the incomplete data for $\lambda^{(d)}$. Note that due to power level issues, training may not be robust in $\lambda^{(d)}$. The updated parameters of $\hat{\lambda}$ are

$$\hat{C}_{im} = \frac{L_{im}^{(d)} + \tau^{(\hat{d})} L_{im}^{(c)}}{\sum_{m=1}^{M} L_{im}^{(d)} + \tau^{(\hat{d})} L_{im}^{(c)}}, \tag{13}$$

$$\hat{\boldsymbol{\mu}}_{im} = \frac{\boldsymbol{m}_{im}^{(d)} + \tau^{(\hat{d})} \boldsymbol{m}_{im}^{(c)}}{L_{im}^{(d)} + \tau^{(\hat{d})} L_{im}^{(c)}}, \tag{14}$$

$$\hat{\boldsymbol{\Sigma}}_{im} = \frac{\boldsymbol{v}_{im}^{(d)} + \tau^{(\hat{d})} \boldsymbol{v}_{im}^{(c)}}{L_{im}^{(d)} + \tau^{(\hat{d})} L_{im}^{(c)}} - \boldsymbol{\mu}_{im}^{(d)} \boldsymbol{\mu}_{im}^{(d)^T}, \tag{15}$$

$$\hat{a}_{ij} = \frac{L_{ij}^{(d)} + \tau^{(\hat{d})} L_{ij}^{(c)}}{\sum_{j=1}^{J} L_{ij}^{(d)} + \tau^{(\hat{d})} L_{ij}^{(c)}}, \tag{16}$$

where $L_{im}^{(c)}, L_{ij}^{(c)}, \boldsymbol{m}_{im}^{(c)}, \boldsymbol{v}_{im}^{(c)}$ are the statistics of $\lambda^{(c)}$. $\tau^{(\hat{d})}$ is the interpolation constant for target distance $\hat{d}$ used to control the value.

In Fig. 4, we pool all the model mixtures (horizontal axis) and plot the mixture component occupancy (vertical axis) for the close-talking clean speech and the distant-talking speech (d= 1.0m, 1.5m, 2.0m and 2.5m) enhanced in the same manner in Fig.1(right) with CMN. The plot is distinct of the different envelope amplitudes that manifest the impact of the mismatch as a function of the speech power (i.e. due to distance). By inspection, mismatch can be minimized by shifting an envelope amplitude to a target level. Suppose that d=2.5m has the corresponding model $\lambda^{(2.5m)}$ and the target speaker is located at $\hat{d}$=1.5m. In the same figure, we show that the interpolation of the clean model using $\tau^{(1.5m)}$ shifts the original envelope level from d=2.5m closer to d=1.5m.

5. EXPERIMENTAL SET-UP

The training database is from the Japanese Newspaper Article Sentence (JNAS) corpus. The open test set is composed of 200 utterances coming from 24 speakers. Recognition experiments are carried out on the Japanese dictation task with 20K-word vocabulary. The language model is a standard word trigram model. The acoustic

Table 1. Word accuracy (%) averaged over 10 mic-array positions in each room (**A**-**E** use close-talking clean speech model).

	Room 1: T_{60}= 240 ms.				Room 2: T_{60}= 640 ms.			
Methods	1.0 m	1.5 m	2.0m	2.5 m	1.0 m	1.5 m	2.0 m	2.5 m
A. Unprocessed	72.3%	56.5%	34.3%	19.1%	14.2%	-1.0%	-10.0%	-22.3%
B. Blind Dereverberation [2]	74.1%	63.3%	47.0%	35.4%	31.6%	17.8%	8.0%	1.0%
C. Spectral Subtraction SS (Previous work [4][5])	**76.1%**	**69.3%**	**53.4%**	**44.6%**	**36.0%**	**24.1%**	**14.5%**	**5.7%**
D. SS + RTF Upd: Sec 3	77.2%	71.0%	55.9%	47.7%	39.3%	28.2%	19.8%	11.5%
E. SS + RTF Upd: Meas. RTF	77.5%	71.4%	56.2%	47.9%	39.5%	28.4%	19.9%	11.8%
F. SS + RTF Upd: Sec 3 + Model Upd: Sec 4 (Proposed)	**79.0%**	**74.1%**	**60.5%**	**54.0%**	**46.2%**	**38.3%**	**34.8%**	**27.1%**
G. SS + RTF Upd: Sec 3 + Model Upd: Re-train (Upper limit)	79.6%	74.9%	61.3%	55.2%	46.8%	39.1%	35.7%	28.2%

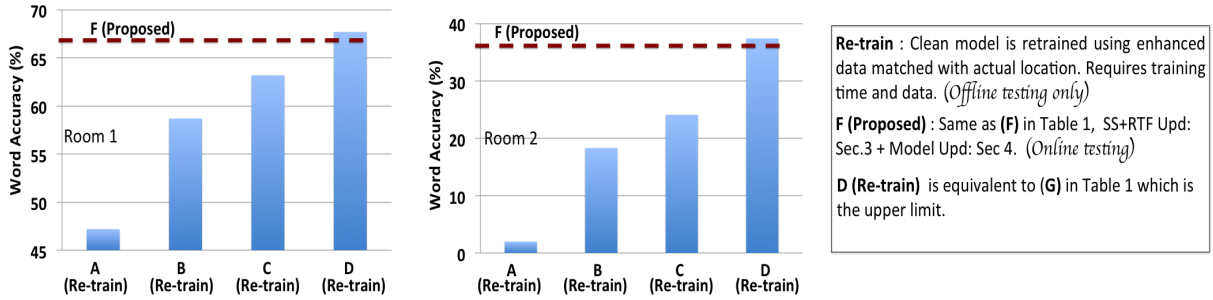

Fig. 5. Performance comparison between re-training and model update (Averaged for all $d_{[i]} = \{1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$).

model is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. The microphone array is embedded on the head of the robot. In the experiment, we used different occlusions such as table, chairs, etc. (real environment setting).

Real reverberant data are recorded inside two different reverberant rooms (Room 1 and Room 2) with reverberation time T_{60}=240 ms. and T_{60}=640 ms., respectively. The mic-array is set-up into ten different positions in each room. For each mic-array position, four location points $d_{[i]} = \{1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$, $1 \leq i \leq i_d = 4$ are designated for testing. Each test location point consists of 200 test utterances. In this experiment i_d=3 is sufficient for step (1) in Sec. 3.1 (i.e. we used 0.5m, 1.3m and 3.0m which are different from the ones used for testing). We also note that a single $f(\hat{d})$ in Eq. (6) is sufficient for the ten different mic-array set-up inside the room used to collect test data.

6. ASR RECOGNITION RESULTS

In Table I, (A)-(E) methods use close-talking clean speech acoustic model for ASR. (A) is the performance when the reverberant test data are not processed (no enhancement). (B) is the performance using a blind dereverberation method that does not require any RTF estimation to carry out dereverberation [2]. (C) is the performance when using the SS-based dereverberation [4][5] (no RTF or model updates to compensate for the change in speaker position). (D) is the result when RTF is updated using the scheme in Sec 3. (E) is the same as (D) except that the actual measured RTF is used instead of the scheme in Sec. 3. (F) is when both the RTF in Sec. 3 and model update in Sec. 4 are employed in the SS-based dereverberation (proposed). Lastly, (G) is the result when model re-training is used instead of the scheme in Sec. 4 (close-talking clean model is re-trained with the processed data matched at location d).

Blind dereverberation (B) outperforms the unprocessed data (A). We note that (B) operates blindly (no RTF is required). (C) performs better than (B) due to the RTF information and some optimization [5]. It is confirmed in (D) that the RTF update in Sec. 3 outperforms (C). We also note that there is no significant benefit in using the actual measured RTF in (E) when compared to (D). This means that RTF update discussed in Sec 3 is sufficient for ASR application. This is also validated in [13], where it is claimed that accurate RTF estimation for speech modelled by HMM is not necessary due to the loose representation of speech in the HMM (i.e., mixtures of gaussians). In (F) where both RTF and model updates are in effect, the recognition performance is superior than the existing methods (B) and (C). Furthermore, in this table, it is shown that there is insignificant change in performance when using model retraining (G) as opposed to the fast model update in Sec. 4 (F).

In Fig. 5 we show the performance of the methods in Table 1 (i.e., (A), (B), (C) and (D)) when re-training is employed (i.e., the close-talking clean speech model used in Table 1 is re-trained using the processed speech database matching the correct d). It is apparent that even with re-training: A (Re-train), B (Re-train), and C (Re-train), these methods cannot beat the proposed method in (F). D (Re-train) in Fig. 5 is equivalent to (G) in Table 1. D (Re-train) is only slightly better than the proposed method in (F) but it requires time-consuming re-estimation and cannot be executed online. In our experiment, we only show the results of model re-training and excluded the results for model adaptation since we have access to the training data. Model re-training is better than adaptation. Thus, its result serves as the baseline.

It is important to stress that ASR in very reverberant condition is a very difficult task. Consequently, distant-talking ASR with test speakers located more than 1 meter distance away from the mic-array is proven to be a herculian task. Experimental results under these severe conditions provided in Table 1 are for scientific evaluation only.

7. CONCLUSION

We have presented a series of corrective schemes that update the parameters used in dereverberation and in the acoustic model. In effect, we have improved our previous work in dereverberation, robust to the variation in speaker location in distant-talking ASR. By addressing both the effects in the waveform and acoustic model, we have shown that robustness is better achieved when both are combined. In our future works, we will further investigate the synergy between these two for improved performance in distant-talking ASR.

8. REFERENCES

[1] C.J.Leggeter and P.C. Woodland,"Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models" *In Proceedings of Computer Speech and Language*, 1995

[2] B. Yegnanarayana and P. Satyaranyarana, "Enhancement of Reverberant Speech Using LP Residual Signals", *In Proceedings of IEEE Trans. on Audio, Speech and Lang. Proc.*, 2000.

[3] S. Griebel and M. Brandstein, "Wavelet Transform Extrema Clustering for Multi-channel Speech Dereverberation" *IEEE Workshop on Acoustic Echo and Noise Control*, 1999

[4] R. Gomez, T. Kawahara, "Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer" *In Proceedings of Interspeech*, 2009.

[5] R. Gomez and T. Kawahara, "Robust Speech Recognition based on Dereverberation Parameter Optimization using Acoustic Model Likelihood" *In Proceedings IEEE Transactions Speech and Acoustics Processing*, 2010

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

[7] H. Kuttruff, "Room Acoustics" *Spon Press*, 2000

[8] Single and Multi-microphone Speech Dereverberation Using Spectral Enhancement *Ph.D. Thesis*, June 2007.

[9] P. Naylor and N. Gaubitch, "Speech Dereverberation" *In Proceedings IWAENC*, 2005

[10] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* ,1979.

[11] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "Adaptive Step-size Parameter Control for real World Blind Source Separation" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008.

[12] H. Sawada *et al.*,"Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. of ICASSP 2002*, 2002.

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