ACOUSTIC SCENE AWARE DEREVERBERATION USING 2-CHANNEL SPECTRAL ENHANCEMENT FOR REVERB CHALLENGE

Xiaofei WANG, Yanmeng GUO, Xi YANG, Qiang FU and Yonghong YAN

Key Laboratory of Speech Acoustics and Content Understanding,
Institutes of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

ABSTRACT

A 2-channel dereverberation method contributed to the REVERB challenge is proposed in this paper. It aims at achieving robust dereverberation under different reverberant conditions. 2-channel spectral enhancement method is used where the gain of each frequency bin is controlled by acoustic scene, which is detected based on the analysis of full-band coherent property. A preprocessing module is supplied to satisfy the requirement of enhancement. In addition, we describe the back-end re-trained acoustic model to match the front-end signal processing. Results, including enhancement indexes and improvement on recognition rate, are evaluated on the simulated data and really recorded data provided by REVERB organizers.

Index Terms— REVERB Challenge, Dereverberation, Spectral Enhancement, Acoustic Scene Awareness

1. INTRODUCTION

The reverberation is known to degrade severely the audible quality of speech and performance of automatic speech recognition (ASR) [1]. Lots of researches have proved that audio processing is helpful in improving the quality of the reverberant speech. Meanwhile, the combination of the front-end audio processing with the back-end speech recognition techniques is also effective to improve the ASR performance in reverberant conditions [2][3]. Among the front-end signal processing technologies, three categories of dereverberation methods are generally applied: 1) beamforming using microphone arrays, 2) spectral enhancement, 3) blind system identification and inversion [4]. Spectral enhancement based dereverberation shows superiority due to its robustness in both reverberant and noisy environment [5]. However, under certain enhancement method, there is usually a tradeoff between the reverberation or noise suppressing amount and the target signal distortion [6]. There are two reasons that cause the tradeoff. First, the models established for reverberation and target speech signal are inaccurate and applicable in limited ranges. Second, the spectral enhancement method has the side-effect of music noise.

According to REVERB challenge, the reverberant data is simulated or recorded in various rooms with different distances between source and microphones [7][8][9], and three kinds of utterance are provided: 1-channel, 2-channel and 8-channel. We choose the 2-channel dereverberation method for both Speech Enhancement (SE) task and Automatic Speech Recognition (ASR) task for three reasons. First, the 2-channel structure is the basic topology of all microphone arrays, so the research can be generalized to any other microphone arrays conveniently. Second, among all the techniques of microphone array, the 2-channel algorithm has the lowest requirement for both hardware and software, which is important for the application of microphone array techniques. Finally, it is ideal to fulfill the dereverberation task based on 2 sensors, just like what the human auditory system is doing everyday, though there is still a long way to go.

Fractional time delay alignment filter is applied to the reverberant signal, and the acoustic scene is classified by analyzing the coherent component. Based on the acoustic scene, an appropriate spectral enhancing scheme is selected to eliminate the interference as much as possible while keeping the speech distortion always in a low level. In the ASR task, late reverberation is paid more attention to because early reflections are beneficial to recognition rate [10][11]. The back-end ASR is based on a triphone HMM architecture as in [1]. The performance of the dereveberation method is tested by the recognizers with acoustic models of "clean-condition" and "multi-condition" HMMs [12], with and without unsupervised Constrained Maximum Likelihood Linear Regression (CMLLR) model adaptation. What's more, we additionally re-train the acoustic model to adapt to the potential distortion in the front-end enhanced signals, and the recognition result is also given.

This paper is organized as follows. In Section 2, system description for REVERB Challenge is introduced and front-end processing is mainly focused on. Section 3 gives the back-end characterization. In Section 4, performance on ob-

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) and the CAS Priority Deployment Project (No. KGZD-EW-103-2).

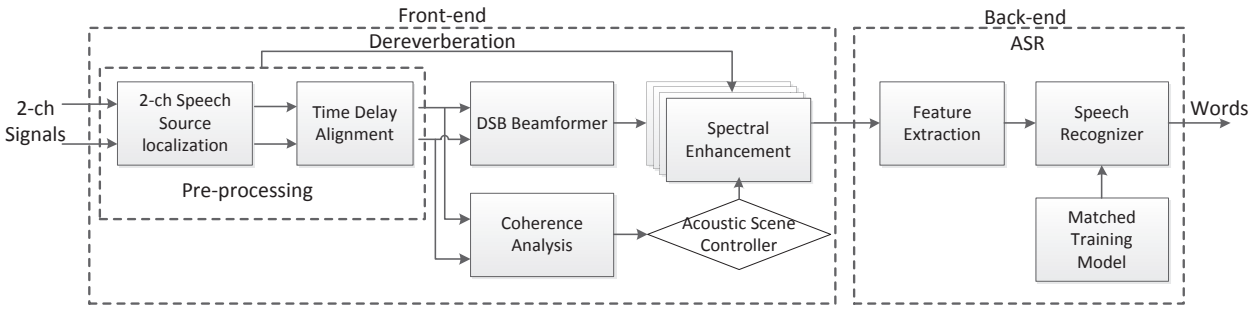

Fig. 1. Block Diagram of the proposed system for REVERB challenge.

jective evaluation is supplied recommended by organizers of challenge. At last, the conclusion is given in Section 5.

2. SYSTEM OVERVIEW

Fig.1 shows the overall system proposed for REVERB challenge. It contains two basic modules, front-end signal processing module and back-end evaluation module. In the front-end module, we handle a pre-processing module before dereverberation.

2.1. Database

In the REVERB challenge, both simulated and really recorded data are provided. The simulated data (SimData) is convolved by clean utterance from WSJCAM0 corpus [7] with the recorded room impulse response (RIR) in different rooms. The reverberation time (T60) of the rooms are 250ms, 500ms and 700ms respectively. Recorded background noise is added to the reverberant data at a fixed signal-to-noise ratio (SNR) of 20dB. The really recorded data (RealData), utterances from the MC-WSJ-AV corpus [8], consists of utterances recorded in a noisy and reverberant room with reverberation time of 700ms. Both SimData and RealData include two types of distances between the speaker and microphone array (near=50cm and far=200cm). The develop and test data are all from the SimData and RealData databases under the following important assumptions. First, there is no drastic change in RIR within an utterance. Second, relative speaker-microphone position changes from utterance to utterance, which means the direction of arrival (DOA) of the target speech signal is uncertain, and this is essential to our dereverberation method. The recording 8-channel cicular array has diameter of 20cm and the 2-channel microphone distance, denoted by d_{mic}, can be calculated.

2.2. Pre-processing

2.2.1. Full-utterance 2-ch speech source localization

As mentioned above, DOA of target speech signal is unknown. So a speech source localization (SSL) procedure is under requirement for further purpose. The main difficulties in SSL are caused by diffuse noise, reverberation and spatial aliasing[13][14]. But their affections are different in different frequency bins. Therefore, the main idea in the SSL is to extract the time difference of arrival (TDOA) only from the frequency bins with high Signal-to-Noise Ratio (SNR) and Direct-to-Reverberate Ratio (DRR), and the spatial aliasing is eliminated by a stepwise strategy based on the consistency of real TDOA. A DOA estimation scheme is proposed as follows.

Firstly, the observed signal is analyzed by short time Fourier transform (STFT), and the proportion of coherent component of each frequency bin is roughly calculated. Then the coherent component and non-coherent component are separately tracked in each frequency bin. Secondly, the frequency bins which are dominated by the direct sound are extracted, and a probability distribution of TDOA is obtained based on low frequency band. Then, the aliased TDOAs in high frequency bins are eliminated based on the TDOA distribution in lower frequency band. Finally, the TDOA is estimated based on the information of the whole frequency band and DOA is suggested. It is worth mentioning that only the probability distribution of TDOA is full-utterance based which indicates that if the DOA information is priori known, the following core procedures can be realized realtime.

2.2.2. Time-delay estimation and alignment

With the DOA information, a delay alignment filter should be applied to observed signals at sensors to satisfy the requirement of beamforming and other operations. Ideal time-delay alignment filter needs the amplitude response keep being 1 and the group delay be constant. Based on the DOA information of target speech signal derived from SSL module, group delay τ between the time delayed channel and the reference channel can be deducted by Eq.(1), where ψ is DOA, c is the

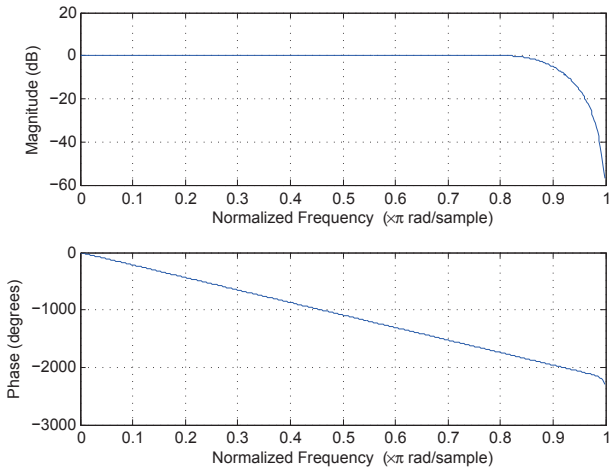

Fig. 2. Amplitude and Phase frequency response of alignment filter with $d_{mic} = 7.65cm$ and filter order equals to 32 where signal comes from 30° far field.

sound velocity in the air.

$$\tau = d_{mic} sin(\psi)/c \tag{1}$$

Accordingly, phase response of frequency f is shown in Eq.(2).

$$\phi(f) = 2\pi f \tau \tag{2}$$

Assuming that the the alignment filter is FIR and limiting the filter order, we truncate the filter with a Hanning window and apply IDFT to the ideal frequency response of alignment filter. By filtering the frame based observed signal with the filter, we can consider the target signal comes from straight front of broadside microphones. It is a significant procedure to perform beamforming later. Fig.2 gives an example of alignment filter. From the figure, we observe that only the high frequency part will have distortion. And this distortion will be considered in the dereverberation method later.

2.3. 2-channel Dereverberation

2.3.1. Signal model

Let $s(n)$ represent the target clean signal, and $x'_m(n)$ is the time-aligned signal at sensor $m(m = 1, 2)$ to keep distinction with observed signals. $n_m(n)$ is the noise of environment. $h_m(t)$ can be seen as the time-delayed RIR which is the convolution of real RIR from target signal to sensor $m(m = 1, 2)$ and alignment filter obtained above. The observed signal per channel can be expressed as follows.

$$x'_m(n) = h_m(t) * s(n) + n_m(n) \tag{3}$$

Applying STFT to the 16kHz time-aligned signal, we have sinal expression at lth frame and kth frequency bin in time-frequency domain.

$$X'_m(l,k) = H_m(l,k)S(l,k) + N_m(l,k) \tag{4}$$

2.3.2. Acoustic scene aware controller

The database to be recognized contains various conditions of rooms and speaker-mic distance. To our knowledge, it's difficult to tackle the reverberant signal well in all kinds of conditions with 2-channel data. Before enhancement, we should priori acknowledge the acoustic scene so that we can legitimately choose strategy of dereverberation.

Reverberation, especially late reverberation, shows isotropic property as well as environment diffuse noise, while the direct sound shows strong coherent property [15]. There are two main acoustic scenes of the room we should blindly aware. One is the reflection condition which can be interpreted by reverberation time (T60) and the other one is the speake-mic distance. By estimating the proportion of coherent component, the effects of two are synthesized. All the diffuse part can be seen as noise to be filtered. We follow the Coherent-to-Diffuse energy Ratio (CDR) estimation in [16], which is expressed as follows.

$$\epsilon(e^{j\Omega}) = \frac{|\mathrm{sinc}(\Omega f_s d_{mic}/c)|^2 - |\Gamma_{X_1X_2}(e^{j\Omega})|^2}{|\Gamma_{X_1X_2}(e^{j\Omega})|^2 - 1} \tag{5}$$

$\epsilon(e^{j\Omega})$ is CDR estimator of each frequency bin and $\Gamma_{X_1X_2}$ is the expression of complex coherence function [17]. To our knowledge, global CDR, denoted by $\hat{\epsilon}$, could reduce the estimation variance since the frame length often less than order of RIR. Through full-band averaging and recursively smoothing, global CDR acts as an indicator of coherent component strength, which is mainly direct sound. Though there exists bias of estimation especially in low DRR case, it's still a reasonable acoustic scene aware controller. And analysis of coherent property is the main reason we choose 2-ch framework rather than 1-ch one.

2.3.3. Spectral enhancement

Spectral Enhancement method has a generalized form. The estimate of the amplitude spectrum of the target signal can be expressed as follows.

$$|\hat{S}(l,k)| = G(l,k)|\hat{X}(l,k)| \tag{6}$$

$G(l,k)$ is the gain estimated on each frequency bin and $|\hat{X}(l,k)|$ is the amplitude spectrum of signal to be enhanced. Before overlap-and-add scheme, regardless of leakage between frequency bins causing by STFT, both $G(l,k)$ and $\hat{X}(l,k)$ should chosen cautiously versus distortion to achieve robustness.

2.3.4. Frame based processing

Up to now, lots of 1-channel and 2-channel dereverberation methods are proved efficiency under the framework of spectral enhancement and achieve robustness to noise compared with inverse filtering. Fixed beamforming, such as

Delay-and-Sum Beamformer (DSB), helps to suppress the reverberation based on priori DOA information though its suppression ability is limited. Late reverberaion suppressing method using generalized statistic model of reverberation [18] shows outstanding performance especially when reverberation is strong. Based on the controller mentioned above, we form different estimators of $G(l,k)$ and $\hat{X}(l,k)$. G_{late} is the gain of each frequency bin which is generated by spectral subtraction method based on late reverberation estimation [18]. Using Eq.(5), G_{cdr} is formed under Wiener solution which can be interpreted by $\epsilon(l,k)/(1+\epsilon(l,k))$ (where $\epsilon(l,k)=\epsilon(e^{j\Omega})$). X_0 and X_{DSB}, which are separately 1-ch extraction of observed reverberant signal and output of DSB beamformer shown by Fig.1, are two estimators of $\hat{X}(l,k)$.

Three parameters $\sigma_1, \sigma_2, \sigma_3$ (constant, from large to small) are introduced to control the selection of $G(l,k)$ and $\hat{X}(l,k)$. An integrated strategy of spectral enhancement method is proposed as follows.

Algorithm 1 Strategy of Spectral enhancement.

if $\hat{\epsilon} > \sigma_1$ **then**
  $G(l,k)(FFT\_bins/3 : FFT\_bins) = 1$
  $G(l,k)(1 : FFT\_bins/3-1) = max(G_{late}, G_{cdr})$
  $\hat{X}(l,k) = X_0$
else if $\hat{\epsilon} > \sigma_2$ **then**
  $G(l,k) = \max(G_{late}, G_{cdr})$
  $\hat{X}(l,k) = X_{DSB}$
else if $\hat{\epsilon} > \sigma_3$ **then**
  $G(l,k) = \min(G_{late}, G_{cdr})$
  $\hat{X}(l,k) = X_{DSB}$
else
  $G(l,k) = G_{late}G_{cdr}$
  $\hat{X}(l,k) = X_{DSB}$
end if

Spectral enhancement strategy suggested above separates the acoustic scene into four cases. In the first case, the acoustic scene is ideal so that the speech signal recorded is very close to clean speech. Even distortion of alignment in high frequency band would have brought negative effects. That's the reason we still save the original observed signal X_0. Lower frequency bins are more likely to be affected by reverberation. Therefore, we intend to keep high frequency part unprocessed. In the second and third cases, a moderate tradeoff between dereverberation and signal distortion is achieved. G_{late} or G_{cdr} respectively shows their superiority when reverberation is relatively strong or weak. In the last case, more reverberation reduction means better performance when reverberation is strong enough.

To avoid music noise, both time recursive and adjacent frequency gain smoothing are conducted. The recovered signal is obtained by inverse STFT and overlap-and-add scheme. The phase of recovered speech signal equals the noisy phase of X_{DSB}. All the processing is with windows of 512 points and step-size of 256 points which means the result of a DFT with length 512 (32 ms) at a shift of 16 ms. So the FFT_bins above equals half the number of frequency bins plus one.

3. BACK-END DESCRIPTION

Baseline models of ASR task as well as re-trained model are provided in this section. The "clean-condition" baseline system uses 39D mel-frequency cepstral coefficients (MFCCs) including Delta and Delta-Delta coefficients as features. As acoustic models, it employs tied-state HMMs with 10 Gaussian components per state trained according to the maximum-likelihood criterion [1]. All the training data for "clean-condition" HMMs is from WSJCAM0 corpus [7]. Further, the model is re-trained using the features of artificially distorted 7861 utterances to form the "multi-condition" HMMs. The utterances are in mixture with 24 kinds of RIRs and 6 kinds of noises. We test the enhanced signals on the two baseline systems both using and not using the unsupervised CMLLR model adaptation.

Under the framework of HTK based recognizer organizer provided, we re-train the acoustic model of "multi-condition" HMMs. The proper starting point is that the artificially distorted training signals are mismatch with the enhanced ones. Therefore, Substituting the 7861 reverberant noisy utterances by the enhanced signal and enlarging the re-training data with 7861$*$24 enhanced convolving utterances, we get the re-trained "multi-condition" HMMs. We also provide ASR result of this re-trained acoustic model. Then the five possible cases are:

Clean+noEnh: "clean-condition" HMMs without dereverberation;

Clean+Enh: "clean-condition" HMMs with dereverberation;

Multi+noEnh: "multi-condition" HMMs without dereverberation;

Multi+Enh: "multi-condition" HMMs with dereverberation;

ReTrn+Enh: re-trained "multi-condition" HMMs with dereverberation.

4. RESULT AND ANALYSIS

4.1. SE task

This section provides the results of SE task from the perspective of Cepstrum Distance (CD) (Table 1), Speech-to-Reverberation Modulation energy Ratio (SRMR) (Table 2 and 6), Log Likelihood Ratio (LLR) (Table 3), Frequency-weighted segmental SNR (FWSegSNR) (Table 4), and PESQ (Table 5) on the test set, where **org** means the original reverberant signals unprocessed and **enh** represents the enhanced speech signals.

It can be seen from the various objective indexes of SE task that the spectral enhancement based dereverberation method achieves improvement on speech quality. The improvement is shown by two aspects. First, Reverberation and noise reduction are achieved from the objective measures of average SRMR and FWSegSNR of different reverberant rooms. Second, signal distortion is lower after processing shown by CD. It is worth mentioning that at the same time introducing the reverberation suppression to reverberant signal with large reverberation time, CD of enhanced signals with small reverberation time keep small, which can be interpreted by the result of **room1_near** in Table 1. It owns to the first case we choose in the enhancement strategy above. However, all the objective indexes on average has tended to better expect LLR. This is because the distortion of spectral enhancement always exists which may counteract the improvement especially the reverberation is not strong. And because of distant-talking scene, the signal attenuates when propagating in the air. The attenuation can not be compensated using the proposed spectral enhancement method. PESQ scores shown by Table 5 also illustrate an improvement on speech signal quality from the perspective of perceptual evaluation.

Additionally, the complexion of calculation is supplied. The real-time-factors of SimData and RealData are 0.47 and 0.44, while real-time-factor of the reference enhancement code provided by organiser is 0.035 [1].

Room	Cepstral distance in dB			
	mean		median	
	org	enh	org	enh
room1_near	1.99	1.96	1.68	1.69
room1_far	2.67	2.78	2.38	2.65
room2_near	4.63	3.52	4.24	3.35
room2_far	5.21	4.51	5.04	4.25
room3_near	4.38	3.57	4.04	3.43
room3_far	4.96	4.42	4.73	4.18
average	**3.97**	**3.46**	**3.69**	**3.26**

Table 1. Cepstral distance of **test SimData** before and after dereverberation.

Room	SRMR (only mean used)			
	mean		median	
	org	enh	org	enh
room1_near	4.50	4.13	_	_
room1_far	4.58	4.53	_	_
room2_near	3.74	3.88	_	_
room2_far	2.97	4.25	_	_
room3_near	3.57	3.80	_	_
room3_far	2.73	3.84	_	_
average	**3.68**	**4.07**	_	_

Table 2. SRMR of **test SimData** before and after dereverberation.

Room	Log likelihood ratio			
	mean		median	
	org	enh	org	enh
room1_near	0.35	0.35	0.33	0.33
room1_far	0.38	0.45	0.35	0.42
room2_near	0.49	0.56	0.40	0.49
room2_far	0.75	0.78	0.63	0.71
room3_near	0.65	0.65	0.59	0.60
room3_far	0.84	0.80	0.76	0.75
average	**0.58**	**0.60**	**0.51**	**0.55**

Table 3. Log likelihood ratio of **test SimData** before and after dereverberation.

Room	FWSegSNR in dB			
	mean		median	
	org	enh	org	enh
room1_near	8.12	9.86	10.72	10.99
room1_far	6.68	8.56	9.24	8.72
room2_near	3.35	7.19	5.52	8.76
room2_far	1.04	4.29	1.77	6.43
room3_near	2.27	5.59	4.21	6.82
room3_far	0.24	3.04	0.89	4.78
average	**3.62**	**6.42**	**5.39**	**7.75**

Table 4. Frequency-weighted segmental SNR of **test SimData** before and after dereverberation.

Room	PESQ (only mean used)			
	mean		median	
	org	enh	org	enh
room1_near	2.14	2.09	_	_
room1_far	1.61	1.64	_	_
room2_near	1.40	1.69	_	_
room2_far	1.19	1.36	_	_
room3_near	1.37	1.53	_	_
room3_far	1.17	1.23	_	_
average	**1.48**	**1.59**	_	_

Table 5. PESQ of **test SimData** before and after dereverberation.

Room	SRMR (only mean used)			
	mean		median	
	org	enh	org	enh
room1_near	3.17	4.44	_	_
room1_far	3.19	4.67	_	_
average	**3.18**	**4.55**	_	_

Table 6. SRMR of **test RealData** before and after enhancement.

Test Data		Word error rate(%)													
		SimData											RealData		
		Room 1,2,3			Ave.	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
		Clean				Near	Far	Near	Far	Near	Far		Near	Far	
Clean+ noEnh	nocmllr	12.84	12.49	12.13	12.48	18.06	25.38	42.98	82.20	53.54	88.04	51.68	89.72	87.34	88.53
	cmllr	_	_	_	_	14.81	18.86	24.63	64.58	33.77	78.42	39.16	82.31	80.76	81.53
Clean+ Enh	nocmllr	_	_	_	_	17.43	25.25	27.85	49.48	36.51	65.94	37.06	73.91	71.34	72.62
	cmllr	_	_	_	_	14.47	19.47	21.19	34.86	27.16	50.50	27.93	62.66	61.58	62.12
Multi+ noEnh	nocmllr	30.29	30.07	30.11	30.15	20.60	21.15	23.70	38.72	28.08	44.86	29.51	58.45	55.44	56.94
	cmllr	15.99	15.52	15.70	15.73	16.23	18.71	20.50	32.47	24.76	38.88	25.25	50.14	47.57	48.85
Multi+ Enh	nocmllr	_	_	_	_	23.64	36.46	27.72	37.69	34.00	45.85	34.22	59.95	59.49	59.72
	cmllr	_	_	_	_	16.93	20.04	19.91	26.84	23.95	34.33	23.66	44.87	45.81	45.34
ReTrn+ Enh	nocmllr	16.59	15.84	16.41	16.27	15.64	18.76	19.79	28.56	24.01	35.15	23.64	49.50	49.49	49.49
	cmllr	13.76	13.43	13.62	13.60	**14.76**	**16.52**	**18.23**	**24.79**	**21.09**	**31.50**	**21.14**	**42.10**	**45.17**	**43.63**

Table 7. Word error rate for **test** data.

Develop Data		Word error rate(%)													
		SimData											RealData		
		Room 1,2,3			Ave.	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
		Clean				Near	Far	Near	Far	Near	Far		Near	Far	
Clean+ noEnh	nocmllr	10.50	11.51	10.81	10.93	15.29	25.29	43.90	85.80	51.95	88.90	51.81	88.71	88.31	88.51
	cmllr	_	_	_	_	12.93	17.72	24.11	72.57	30.44	79.65	39.53	83.16	84.48	83.81
Clean+ Enh	nocmllr	_	_	_	_	15.07	24.66	26.37	54.82	38.25	64.24	37.21	65.75	66.03	65.88
	cmllr	_	_	_	_	12.76	18.53	19.97	38.70	26.98	49.13	27.66	56.46	58.99	57.71
Multi+ noEnh	nocmllr	25.79	28.44	27.32	27.18	15.49	18.90	23.51	42.40	27.25	46.07	28.92	52.96	51.61	52.28
	cmllr	13.25	14.79	14.09	14.04	13.27	17.08	20.80	36.83	23.54	39.44	25.14	47.91	46.55	47.23
Multi+ Enh	nocmllr	_	_	_	_	18.02	32.45	27.88	39.59	32.12	46.27	32.71	51.72	50.92	51.32
	cmllr	_	_	_	_	13.91	18.04	19.20	29.01	23.66	33.88	22.94	42.11	41.01	41.56
ReTrn+ Enh	nocmllr	13.59	14.15	13.16	13.63	13.00	16.57	18.54	31.55	25.32	36.03	23.48	47.60	45.11	46.36
	cmllr	12.27	12.87	11.47	12.20	**12.59**	**15.54**	**17.03**	**27.06**	**21.12**	**30.46**	**20.62**	**43.36**	**41.56**	**42.46**

Table 8. Word error rate for **develop** data.

4.2. ASR task

For the ASR task, word error rate (WER) of test data and develop data is reported as Table 7 and Table 8 show. Each table includes the ASR results of "clean-condition" model, "multi-condition" model and re-trained "multi-condition" model. WER of clean, near, far data and their average are reported separately. The ASR result of enhanced signal is matched with the SE result previously.

Performance of dereverberation is examined using both "clean-condition" and "multi-condition" acoustic model. Recognising the test set with "clean-condition" model without CMLLR adaptation, the dereverberation method achieves decrease on WER from 51.68% to 37.06% on average of SimData and from 88.53% to 72.62% on average of RealData. Consistent improvement across all recording conditions is achieved by using CMLLR which results in WER 27.93% on SimData and 62.12% on RealData. However, recognising with "multi-condition" model without CMLLR adaptation, the performance (SimData: 34.22%, RealData: 56.94%) is worse than the "multi-condition" baseline (SimData: 29.51%, RealData: 59.72%) because the recognising enhanced data is mismatch with the reverberant data used to train the "multi-condition" acoustic model, though using CMLLR gives a little improvement.

To overcome the mismatch, the re-trained "multi-condition" HMMs gives a better result. The optimised one has a WER 23.64% on average of SimData and 49.49% on average of RealData without CMLLR and finally 21.14% of SimData and 43.63% of RealData with CMLLR. The relative decreasing rates of WER are 59.09% and 50.71% each. What's more, the clean test data recognized by the re-trained "multi-condition" model could also prove the improvement of dereverberation method from the perspective of the matching between training and recognising, where WER decreases from 30.15% to 16.27% on average. The develop set has the same trend. What draws our attention is that the average WER of RealData is higher than that of SimData. Two reasons may cause the observation. One is that the utterances of RealData are not included in training set and another is that the simulated

data can't imitate all the situations of real room environment. The main reasons will be further investigated.

5. CONCLUSION

We have presented out dereverberation approach to the REVERB challenge based on spectral enhancement. An acoustic scene aware technique is proposed to make dereverberation robust to different conditions. For SE task, objective indexes illustrate the improvement on speech signal quality. For ASR task, when it is combined with back-end ASR with matched training, it produces a significant decrease on WER.

6. REFERENCES

[1] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Emanuël Habets, Reinhold Haeb-Umbach, Volker Leutnant, Armin Sehr, Walter Kellermann, Roland Maas, et al., "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," .

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, November 2012.

[3] Armin Sehr and Walter Kellermann, "Towards robust distant-talking automatic speech recognition in reverberant environments," in *Speech and Audio Processing in Adverse Environments*, pp. 679–728. Springer, 2008.

[4] Patrick A Naylor and Nikolay D Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2005.

[5] Takuya Yoshioka, Tomohiro Nakatani, and Masato Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 2, pp. 231–246, 2009.

[6] Philipos C Loizou and Gibak Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 47–56, 2011.

[7] Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals, "Wsjcamo: A british english speech corpus for large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. IEEE, 1995, vol. 1, pp. 81–84.

[8] Mike Lincoln, Iain McCowan, Jithendra Vepa, and Hari Krishna Maganti, "The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 2005, pp. 357–362.

[9] Douglas B Paul and Janet M Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[10] Armin Sehr, Emanuël AP Habets, Roland Maas, and Walter Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[11] Roland Maas, Emanuël AP Habets, Armin Sehr, and Walter Kellermann, "On the application of reverberation suppression to robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 297–300.

[12] Steve J Young, Gunnar Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, V Valtchev, and PC Woodland, "The htk book version 3.4," 2006.

[13] Ryuichi Shimoyama and Ken Yamazaki, "Computational acoustic vision by solving phase ambiguity confusion," *Acoustical Science and Technology*, vol. 30, pp. 199–208, 2009.

[14] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180. Springer, 2001.

[15] Heinrich Kuttruff, *Room acoustics*, Taylor & Francis, 2000.

[16] Marco Jeub, Christoph Nelke, Christophe Beaugeant, and Peter Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *19th European Signal Processing Conference (EUSIPCO 2011)*, 2011, pp. 1347–1351.

[17] Iain A McCowan and Hervé Bourlard, "Microphone array post-filter based on noise field coherence," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 709–716, 2003.

[18] Emanuël AP Habets, Sharon Gannot, and Israel Cohen, "Late reverberant spectral variance estimation based on a statistical model," *Signal Processing Letters, IEEE*, vol. 16, no. 9, pp. 770–773, 2009.