INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Joint Training of Multi-channel-condition Dereverberation and Acoustic Modeling of Microphone Array Speech for Robust Distant Speech Recognition

Fengpei Ge [1*], *Kehuang Li*[2], *Bo Wu*[3], *Sabato Marco Siniscalchi*[4], *Yonghong Yan*[1], *Chin-Hui Lee*[2]

[1]The Key Laboratory of Speech Acoustic and Content Understanding, Institute of Acoustics, China
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, GA, USA
[3]National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China
[4]Faculty of Computer and Telecommunication Engineering, University of Enna Kore, Enna, Italy

gefengpei@qq.com, {kehlekernel, rambowu11, siniscalchi77.19}@gmail.com,
yonghong.yan@hccl.ioa.ac.cn, chl@ece.gatech.edu

Abstract

We propose a novel data utilization strategy, called multi-channel-condition learning, leveraging upon complementary information captured in microphone array speech to jointly train dereverberation and acoustic deep neural network (DNN) models for robust distant speech recognition. Experimental results, with a single automatic speech recognition (ASR) system, on the REVERB2014 simulated evaluation data show that, on 1-channel testing, the baseline joint training scheme attains a word error rate (WER) of 7.47%, reduced from 8.72% for separate training. The proposed multi-channel-condition learning scheme has been experimented on different channel data combinations and usage showing many interesting implications. Finally, training on all 8-channel data and with DNN-based language model rescoring, a state-of-the-art WER of 4.05% is achieved. We anticipate an even lower WER when combining more top ASR systems.

Index Terms: distant speech recognition, reverberant speech recognition, multi-condition, joint training

1. Introduction

Recently the technology of automatic speech recognition (ASR) is being used increasingly in everyday life. However, progress is still needed if we are to handle more challenging situations such as distant ASR in the presence of reverberation [1][2][3]. In an enclosed space, the signal received by a microphone placed at a distance from a speaker contains not only the direct-path signal, but also attenuated and delayed copies of the original speech signals, caused by the reflections from walls, ceilings, and floors [4]. Such a multi-path propagation phenomenon is referred to as reverberation, and it introduces echoes and spectral distortions into the observation signal [5]. The problem of reverberation has long been noted to be critical for distant ASR [2]. The convolutive nature of reverberation induces a long-term correlation between a current observation and past observation of reverberant speech [6]. This correlation has been exploited to mitigate the effect of reverberation directly on the speech signal or on the acoustic model used for ASR. As a result, reverberation often seriously degrades speech quality and intelligibility and causes decreased accuracies for distant ASR.

The reverberation usually causes the mismatch between training and testing conditions, which can be viewed in the signal space, the feature space, or the model space [7]. In the past, it has been repeatedly proven that feature compensation and model adaptation are effective for robust ASR [8][9]. In this study we focus our attention on signal space robustness in terms of speech enhancement (SE), which can be achieved by a pre-processing stage aiming at removing the reverberation effect. Indeed, several speech dereverberation solutions based on conventional signal pre-processing techniques are available in the literature. For example, an inverse filter of room impulse response (RIR) was used to deconvolve the reverberant signals in [10]. However, the RIR may be varying in time and hard to estimate [4]. Several algorithms were also proposed to handle the negative reverberation effects [11][12]. In [13][14], late reverberations were estimated with multi-step linear prediction, and followed by spectral subtraction to reduce the effects. In [15], non-negative matrix factorization (NMF) was used to factorize spectrograms into nonnegative speech and noise dictionaries and their non-negative activations [16][17].

Nonetheless, improved speech perceptions gained through signal-space enhancement techniques do not always directly deliver top ASR accuracies, as demonstrated in the recent Reverberant Voice Enhancement and Recognition Benchmark (REVERB) Challenge [18][19]. Deep neural networks (DNNs) have proven to be a reliable vehicle to attain state-of-the-art enhancement results due to their strong regression capabilities [20][21]. Du et al. [22][37] has shown that speech robustness can be improved leveraging upon DNN based speech enhancement, and a word error rate (WER) reduction of 50% from a baseline system with clean-condition training was demonstrated on the Aurora-4 task. In [38], several parallel and front-back joint-training architectures were devised to enhance far-field speech recognition. There are several specific implementation details that make our work differ from [37], and [38]; moreover, we directly address a multi-channel scenario and work on reverberant ASR. In our previous work [23] a reverberation-time-aware DNN (RTA-DNN) based dereverberation framework was proposed to handle a wide range of reverberant conditions. In [24] a multi-condition concept was proposed to address the challenges of reverberant speech scenarios in real life. In fact, multi-channel speech recorded/simulated by microphone array can be viewed as one special multi-condition situation.

In this paper we adopt DNN-based speech enhancement and propose a novel data utilization strategy, named *multi-channel-condition* learning, to improve DNN-based dereverberation and acoustic models in a joint training setting. We use stereo data to train our models. With a single ASR

This work was done during the first author's visiting stay at Georgia Institute of Technology in 2016-2017.

http://dx.doi.org/10.21437/Interspeech.2017-579

system, the experimental results show that joint training delivers a WER of 7.47% on the 1-channel REVERB2014 simulated evaluation data. With multi-channel-condition learning of all 8-channel data, and by leveraging upon rescoring with DNN based language models, a state-of-the-art WER of 4.05% is achieved. A lower WER is expected with more ASR systems.

2. Baseline System Overview

The framework of our baseline system is shown in Figure 1. First, reverberant speech was fed into the speech enhancement module and decoded with the aid of acoustic model (AM), dictionary and language model (LM). Often a DNN-based LM rescore module was used to retune the lattice confidence to obtain the ultimate recognition results.

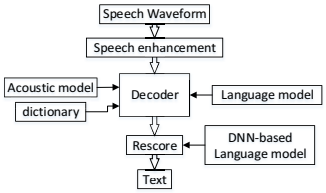

Figure 1 : *System overview for distant speech recognition*

2.1. Speech enhancement front-end

In speech enhancement (SE) front-end, we adopted a reverberation-time-aware DNN-based speech dereverberation system [23] when the back-end was separately trained. Moreover, log-power spectra (LPS) features were adopted. In the training stage, a regression DNN [25] was trained by a set of reverberant and anechoic speech pairs represented by the enhanced LPS. We adapted two key design parameters, frame shift, *R*, and context expansion, *N*, into both training and dereverberation. An utterance-based RT60 estimator followed by a lookup table was required in the dereverberation stage. The required phase was directly extracted from reverberant speech [26]. Finally the dereverberated waveform was reconstructed from the estimated spectral magnitude and the reverberant speech phase with an overlap-add method [27].

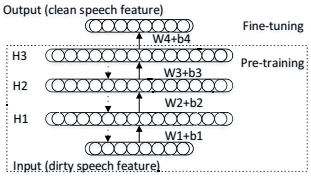

Figure 2 : *DNN for speech enhancement*

The architecture of the proposed dereverberation DNN is illustrated in Figure 2. The 257-dimension LPS feature vectors were used as the input and target features of DNNs, which had 3 hidden layers, 2048 nodes per layer and a linear output. The number of pre-training epochs for each RBM layer was 1, and the learning rate was 0.4. As for fine-tuning, the learning rate and the maximum number of epochs were 0.00008 and 30, respectively. The mini-batch size was set to 128. The configuration parameters were found in [25]. The input and target features of DNN were globally normalized to zero mean and unit variance.

2.2. ASR back-end

In this study, we implemented the baseline ASR back-end as provided by the REVERB challenge and a little extra effort was done. We used the Kaldi toolkit [28] to build all ASR systems presented in this paper.

2.2.1. Acoustic model

We employed a conventional context-dependent DNN hidden Markov model (CD-DNN-HMM) based acoustic model. The input consisted of log mel filter bank (LMFB) features that were processed with global mean and variance normalization. The DNN was initialized using layer-wise restricted Boltzmann machine (RBM) pre-training, then fine-tuned using stochastic gradient descent (SGD) to optimize the sequential training (sMBR) criterion. We used the back propagation algorithm with labels obtained by performing an HMM state alignment of the clean speech training data with a GMM based recognizer trained with the maximum likelihood criterion.

We trained the DNN using multi-condition training data, which covers several reverberant conditions. The REVERB challenge provided a baseline multi-condition training data set that consists of simulated reverberant speech with additional noise. For matching with our SE front-end better, the enhanced speech and the clean speech were also used in training acoustic model. Consequently, the extended training set consists of the same utterances, and any variation originates solely from different acoustic conditions.

2.2.2. Language model

The standard 3-gram language model (LM) provided in the REVERB Challenge was utilized in decoding. Additionally, RNN-based LM [29] and LSTM-based LM [30] were used to rescore the decoding lattices.

2.3. Joint training of dereverberation and acoustic models

For better matching between the speech enhancement (SE) front-end and the ASR back-end, joint training [21] was adopted in this study. Firstly, how to prepare the initial model was an important part. We utilized the single-channel dereverberation approach mentioned in Section 2.1 to obtain a dereverberation DNN regression model as one part of the initial model for joint training. However, we achieved a seamless connection with the acoustic model by replacing the LPS with the LMFB at the input and target layers in the dereverberation DNN regression model. As the other part of the initial model for joint training, the multi-condition acoustic model mentioned in section 2.2.1 was used. The diagram of joint training was shown in Figure 3. The initial/seed model for joint training was obtained through concatenating the speech enhancement model and acoustic model directly. Then several iterations of fine-tuning were carried out on the one-channel, multi-condition training data in order to yield the optimized joint training deep neural network (JT-DNN) model to be simultaneously used as the dereverberation and acoustic module in the final ASR system.

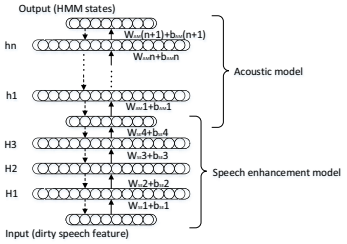

Figure 3 : *Joint training of enhancement and acoustic modules*

3. Multi-channel-condition Joint Learning

Speech obtained in different channels of a microphone array contains subtle diversity that provides complementary information, which can be exploited using spatial filtering, or channel selection. In spatial filtering, the signals acquired from the individual microphone are properly combined, so that the array can act as a spatial filter for suppressing noise and reverberation. In beam-forming, for example, the signals are filtered and weighted in order to obtain a single beam with enhanced sensitivity in the direction of the desired source while sounds from other directions gets attenuated, e.g., [31]. In channel selection, e.g., [32][33], the idea is to select reliable channels to improve system performance. As shown in Section 4.3.2.2, the specific selection has a meaningful effect on the final recognition accuracy; therefore, a selection criterion guided by expert knowledge has to be devised, e.g., [33].

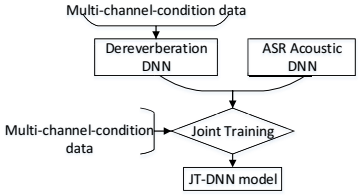

Figure 4 : *A block diagram of the multi-channel-condition learning approach*

A major issue with the two above-mentioned multi-channel techniques is that expert knowledge is involved in order to obtain the desired result. Moreover, beam-forming achieves its maximum performance when the geometry of the array is known and fixed; otherwise, its effectiveness can be seriously reduced. In contrast, a chief promise of deep modelling is to discover rich and complex interactions in the signals without any ad-hoc pre-processing but simply leveraging data. Data augmentation has actually been empirical proved useful to boost DNN generalization and regression capabilities [25] provided that different conditions are covered in training. End-to-end approaches, such as the proposed joint training, allow avoiding intrinsic inconsistencies in traditional technology based on specialized blocks forming a complex pipeline, e.g., [34]. We therefore propose a novel multi-channel-condition joint training scheme, M-JT-DNN, which is illustrated in Figure 4. The key idea is to use speech from all selected channels as additional data, which can be pooled together to form an updated training data set for improving the DNN model robustness. The DNN will eventually boost the signal in the direction of the desired source and possibly ignore/deemphasize some of the available channels using only what available in the data.

4. Experiments and Result Analysis

We evaluated our proposed methods on the official datasets of the 2014 REVERB Challenge [18][19] in which the utterances were taken from the WSJCAM0 corpus [35] and corrupted with different levels of reverberation. Here we focus only on simulated evaluation reverberant data (SimData Eval.).

4.1. Data sets

SimData contains a set of reverberant speech signals that are artificially simulated by convolving clean speech signals with measured RIRs and subsequently adding measured noise signals. It simulates 6 different reverberation conditions: 3 rooms with different volumes (small, medium and large size), 2 types of distances between a speaker and a microphone array (near=50cm and far=200cm). RIRs are measured in 3 different rooms with an 8-ch circular array with diameter of 20 cm. The array is equipped with omnidirectional microphones. Stationary background noise, which is caused mainly by air conditioning systems in a room, is measured under the same conditions with the same arrays as used for RIR measurement. Further details of the data summarized can be found in the official REVERB Challenge site [18].

4.2. Experimental setting

The dereverberation DNN regression model used for initializing the JT-DNN model was trained with 72-dimensional LMFB features with an 11-frame context window. It has 3 hidden layers, 2500 nodes per layer and a linear output. The number of pre-training epochs for each RBM layer was 1, and the learning rate was 0.4. As for fine-tuning, the learning rate and the maximum number of epochs were 0.00008 and 30, respectively. The mini-batch size was set to 128. The input and target features of DNN were globally normalized to zero mean and unit variance. The training data sets were different for each experiment and specified separately in the section of experiment results. The acoustic model in this study was based on the framework of CD-DNN-HMMs, using 72-dimensional LMFB features with an 11-frame context window. The DNN has 6 hidden layers, and each layer has 2048 sigmoid units. The output softmax layer has 2085 senone units. The DNN was initialized with the stacked RBM-based pre-training, and the sequential training (sMBR) was used for frame-level fine-tuning. Note that we no longer need to reconstruct enhanced waveforms. Therefore, the LPS features are replaced with the LMFB features in joint training to deliver better ASR performances in all systems discussed below.

4.3. Experimental results

Here we compared ASR performances with different system configurations. Note that in all tables, "×" signals that a 3-gram language model is used in decoding. The label "√a" indicates that a rescoring step with an RNN-based LM (RNN-LM) was accomplished; whereas, the symbol "√b" represents rescoring with LSTM-based LM (LSTM-LM).

4.3.1. Separate versus joint training

First, we compared the proposed joint training (JT-DNN) approach against the separate training (ST) strategy of SE front-end and ASR back-end using one-channel (ch-1) data only. Results are reported in in Table 1. It was learned that by leveraging upon joint training with more discriminative ASR features and decoding with the 3-gram LM, a WER equal to 7.92%, shown in the rightmost column of the first row in the lower part of Table 1, was obtained, representing the best one-channel ASR performance with no extra speech training data. Compared with the WER of 8.75%, shown in the rightmost column of the first row in the upper part of Table 1, obtained for our baseline single system with separate dereverberation for waveform reconstruction, a WER reduction of 9.5% was observed. If we rescore the lattices obtained in the joint training baseline with RNN-LM, a WER equal to 5.88% was achieved. LSMT-LM based rescoring allowed us to deliver a WER of 4.46%, which sets a state-of-the-art performance previously reported. Detailed results and system configurations of separate and joint training can be found in our recent paper [36]. In Table 1, we also reported ASR

performance with LMFB features, a 3-gram LM, and the separated training (ST) strategy. By comparing WERs in the 1st, 4th and 5th rows, we can conclude that LMFB features improve ASR performance, yet the proposed joint training is what actually improves the robustness of our recognizer.

Table 1 . *WERs (%) on the SimData task, ch-1, with separate and joint training*

	rescore	Room1		Room2		Room3		Average
		near	far	near	far	near	far	
LPS/ST	×	5.70	6.00	7.10	11.50	8.80	13.40	8.75
	√ a	3.74	4.40	5.48	9.11	6.19	10.10	6.50
	√ b	2.71	3.15	3.95	6.99	4.71	8.38	4.98
LMFB/ST	×	5.18	5.95	6.54	11.78	8.67	13.75	8.65
LMBF/ JT-DNN	×	5.05	5.86	6.00	10.63	7.73	12.27	7.92
	√ a	3.11	4.01	4.42	7.75	5.82	9.74	5.88
	√ b	2.49	2.80	3.18	6.43	4.32	7.52	4.46

4.3.2. Multi-channel-condition joint learning

Next, we viewed speech from different channels as one special type of multi-condition, named multi-channel-condition. Three improved systems over baseline joint training with different multi-channel-condition training configurations were evaluated. Information coming from different channels was employed for refining both the front-end and back-end modules within out joint training scheme. All evaluation results are reported on ch-1.

4.3.2.1 Effect of the number of channels

In Table 2, we list the ASR performances with two and eight channels. In two-channel-condition, a WER equal to 4.27% was attained with LSTM-LM rescoring, reduced from 4.46% in the lower part of Table 1 with one-channel-condition. To extend to the eight-channel-condition scenario, we pooled all speech from the eight channels to obtain the M-JT-DNN model. The other elements of the acoustic model remained unchanged. A WER of 4.05% was achieved, substantially dropped from the one-channel-condition system with 4.46%.

Table 2 . *WER (%) with the M-JT-DNN scheme*

	rescore	Room1		Room2		Room3		Average
		near	far	near	far	near	far	
Two-channel (ch-1+2)	×	4.91	5.64	5.75	10.08	7.54	11.09	7.50
	√ a	3.30	4.15	4.37	7.83	5.78	8.36	5.63
	√ b	2.27	2.91	3.30	6.14	4.28	6.72	4.27
Eight-channel	×	4.90	5.56	6.03	10.11	7.76	10.99	7.56
	√ a	3.39	3.73	4.00	7.54	5.63	8.03	5.39
	√ b	2.35	2.57	2.97	5.82	4.37	6.23	4.05

4.3.2.2 Channel selection in two-channel-condition

Table 3 . *WER (%) comparison using different channel speech*

	rescore	Room1		Room2		Room3		Average
		near	far	near	far	near	far	
ch-1+ch-5	×	4.91	5.39	5.90	9.97	7.93	11.86	7.66
	√ a	3.25	4.08	4.00	7.84	5.60	8.43	5.53
	√ b	2.17	2.63	3.01	5.95	4.08	6.94	4.13
ch-1+ch-8	×	5.22	5.74	6.03	11.08	8.70	12.23	8.17
	√ a	3.54	4.42	4.32	8.52	6.55	9.33	6.11
	√ b	2.42	2.96	3.12	6.17	4.98	7.20	4.48

Table 3 list the WERs with two other channel combinations, ch-1+ch-5 and ch-1+ch-8. When compared to the ch-1+ch-2 two-channel-condition system in the upper part of Table 2, with WERs of 7.50%, 5.63% and 4.27%, we observe that WERs were 7.66%, 5.53% and 4.13%, with ch-1 and ch-5 speech. They were both obviously better than the one-channel system in Table 1. By using ch-1 and ch-8 data, WERs were only comparable to the one-channel system. The performance difference may be related to the relative positions of ch-8.

4.3.2.3 Effect of clean data in two-channel-condition

Based on the two-channel-condition joint training scheme, the clean data, which was used as the target in the previously trained dereverberation DNN model, was added into the data pools, as ch-1+ch-2+clean, ch-1+ch-5+clean and ch-1+ch-8+clean, respectively, to enhance the self-learning ability and prevent over-correcting. Compared with the performance of the systems as shown in Table 3, almost all WERs were reduced. Using the configuration with ch-1+ch-2+clean and LSTM-LM based rescoring allowed us to deliver a WER of 4.05. Using the other two-channel-condition configurations, a significant WER drop was also observed. The result is remarkable considering that the top 1-, 2-, and 8-ch WERs were 5.20%, 4.40%, and 4.20% with lots of extra training speech data and system combination in [20].

Table 4 . *ASR Performances with clean speech joined in the multi-channel-condition training scheme in terms of WER (%)*

	rescore	Room1		Room2		Room3		Average
		near	far	near	far	near	far	
ch-1+ch-2	×	4.86	5.62	5.79	10.37	7.59	10.58	7.47
	√ a	3.51	4.00	4.28	7.62	5.41	7.98	5.47
	√ b	2.39	2.80	2.94	5.98	4.13	6.04	4.05
ch-1+ch-5	×	4.86	5.39	5.75	10.13	7.86	11.80	7.63
	√ a	3.27	3.98	4.07	7.92	5.73	8.46	5.57
	√ b	2.10	2.68	3.01	6.22	4.18	6.74	4.16
ch-1+ch-8	×	5.73	6.23	6.51	10.96	8.27	12.59	8.38
	√ a	3.44	3.95	4.20	7.62	5.60	7.85	5.44
	√ b	2.39	2.57	3.26	6.16	4.30	6.19	4.15

5. Conclusions

We have proposed a novel multi-channel-condition data utilization strategy to improve the robustness of dereverberation and acoustic models for distant ASR of microphone array speech. As another useful measure, joint training of speech enhancement model and acoustic model was adopted to fine-tune their matching. The multi-channel-condition learning scheme was also used in the joint training process. Through comparing the ASR performances with the different channel selection patterns, it was learned that the speech from different channels of a microphone array was helpful to perform the target channel evaluation. With the experimental results on the1-channel REVERB2014 simulated evaluation data, it was known that by leveraging upon rescoring with neural network based language models, the state-of-the-art WER of 4.05% was achieved with a single ASR system. Additionally, the anechoic/clean data picked up by the near microphone can assist to enhance the robustness of the dereverberation model by using as the input feature. In the next study, we will further experiment to merge the anechoic/clean data into the eight-channel-condition training scheme and investigate the system combination strategy.

6. Acknowledgements

The first author was supported by a grant from the China Scholarship Council and this work was partially supported by the National Natural Science Foundation of China (Nos. 11461141004, 61271426, U153611, 711504406, 115907 74) and National 863 Program (No. 2015AA016306).

7. References

[1] J. Li, L Deng, Y Gong, R Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE Trans. Audio Speech Lang. Process*., Vol. 22, pp. 745-777, 2014.

[2] R Haeb-Umbach, A Krueger, "Reverberant speech recognition," Chap. 10 in *Techniques for Noise Robustness in Automatic Speech Recognition*, ed. by T Virtanen, R Singh, and B Raj., Wiley, 2012.

[3] T Yoshioka, A Sehr, M Delcroix, K Kinoshita, R Maas, T Nakatani, W Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag*. Vol. 29, pp. 114 – 126, 2012.

[4] P. A. Naylor and N. D. Gaubitch, Eds., "*Speech Dereverberation*," London, UK: Springer, 2010.

[5] J. Benesty, S. Makino, and J. D. Chen, Eds, "*Speech Enhancement*," Berlin, Germany: Springer, 2005.

[6] Delcroix M, Yoshioka T, Ogawa A, et al. "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv Signal Process.*, No. 1, pp. 1-15, 2015.

[7] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Processing*, Vol. 4, pp. 190–202, 1996.

[8] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Communication*, Vol. 25, pp. 29–47, 1998.

[9] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, Vol. 88, pp. 1241–1269, 2000.

[10] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, Vol. 66, No. 1, 1979.

[11] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 14, 2006.

[12] S. Mosayyebpour, M. Esmaeili, and T. Aaron Gulliver, "Single-microphone early and late reverberation suppression in noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 21, pp. 322–335, 2013.

[13] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 17, No. 4, pp. 534–545, 2009.

[14] M. Delcroix et al., "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proc. REVERB Challenge Workshop*, 2014.

[15] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Inf. Process. Syst.*, 2001.

[16] T. Viratnen, J. F. Gemmeke and A. Hurmalainen, "Exemplar based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio Speech Lang. Process*., Vol. 19, pp. 2067–2080, 2011.

[17] H. Kallasjoki, J. Gemmeke, K. J. Palomäki, A. Beeston, and G. Brown, "Recognition of reverberant speech by missing data imputation and NMF feature enhancement," in *Proc. REVERB Challenge Workshop*, 2014.

[18] K. Kinoshita et.al, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. WASPAA*, 2013, pp. 1–4.

[19] K. Kinoshita et.al, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, No. 1, pp. 1–19, 2016.

[20] X. Xiao et al., "The NTU-ADSC systems for reverberation challenge 2014," in *Proc. REVERB Challenge Workshop*, 2014.

[21] M. Mimura, S. Sakai, and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders augmented with a phone-class feature," *EURASIP J. Adv Signal Process.*, no. 1, 2015.

[22] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Proc. Interspeech*, 2014.

[23] B. Wu, K. Li, M. Yang, C.-H. Lee, "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks HMM," *IEEE Trans. on Audio, Speech and Lang. Process.,* Vol. 25, pp. 98-107, 2017.

[24] K. Katayama, and Y. Miyanaga , "A Study of Reverberation Robust Speech Recognition Using Multi-Condition HMM," *IEICE, Vol.110, pp.83-88, 2010*.

[25] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process*., Vol. 23, pp. 7–19, 2015.

[26] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 30, No. 4, pp. 679–681, 1982.

[27] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, 2008.

[28] D. Povey et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[29] M. Thomas, "Statistical Language Models Based on Neural Networks," *Ph.D. thesis*, Brno Univ. of Technology, 2012.

[30] M. Sundermeyer, R. Schluter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. Interspeech*, 2012.

[31] M. Brandstein and D. Ward, "Microphone Arrays: Signal Processing Techniques and Applications." Berlin: Springer, 2001.

[32] M. Wolf, and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, Vol. 57, pp. 170-180, 2014.

[33] I. Himawan, P. Motlicek, S. Sridharan, D. Dean, D. Tjondronegoro, "Channel selection in the short-time modulation domain for distant speech recognition," In Proc. INTERSPEECH, pp. 741-745, 2015.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Image net classification with deep convolutional neural networks," in *Proc. NIPS*, 2012

[35] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: A british english speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1995.

[36] B. Wu, K. Li, Z. Huang, S. M. Siniscalchi, M. Yang and C.-H. Lee, "A Unified Deep Modeling Approach to Simultaneous Speech Dereverberation and Recognition for the REVERB Challenge," *Proc. HSCMA Workshop,* San Francisco, March 2017.

[37] T. Gao, J. Du, L. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2015.

[38] Y. Qian, T. Tan, D. Yu, "An investigation into using parallel data for far-field speech recognition, *Proc. ICASSP*, 2016.