INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

One-pass single-channel noisy speech recognition using a combination of noisy and enhanced features

Masakiyo Fujimoto, Hisashi Kawai

National Institute of Information and Communications Technology, Japan
{masakiyo.fujimoto, hisashi.kawai}@nict.go.jp

Abstract

This paper introduces a method of noise-robust automatic speech recognition (ASR) that remains effective under one-pass single-channel processing. Under these constraints, the use of single-channel speech enhancement seems to be a reasonable noise-robust approach to ASR, because complicated techniques requiring multi-pass processing cannot be used. However, in many cases, single-channel speech enhancement seriously deteriorates the accuracy of ASR because of speech distortion. In addition, the advanced acoustic modeling framework (joint training) is relatively ineffective in the case of single-channel processing. To overcome these problems, we propose a noise-robust acoustic modeling framework based on a feature-level combination of noisy speech and enhanced speech. To obtain further improvements, we also adopt a sub-network-level combination of noisy and enhanced speech, and a gating mechanism that can dynamically select appropriate speech features. Through comparative evaluations, we confirm that the proposed method successfully improves the accuracy of ASR in noisy environments under strong constraints.

Index Terms: noise robust ASR, one-pass single-channel processing, speech enhancement, feature/sub-network-level combination, gating mechanism

1. Introduction

As speech applications on mobile devices and smart speakers on home devices become increasingly widespread, ensuring the noise robustness of automatic speech recognition (ASR) in our daily environment is becoming extremely important. Through the development of a speech-to-speech multilingual translation application for mobile devices, we have come to recognize the importance of this issue.

Various noise-robust ASR methods have been developed recently to overcome this problem. The simplest approach to ensuring noise robustness is speech enhancement during front-end processing of ASR. Traditional methods of speech enhancement such as spectral subtraction (SS) [1], minimum mean squared error-short term spectral amplitude estimation (MMSE-STSA) [2], and Gaussian mixture model-based feature enhancement (GMM-FE) [3] are well known. Deep learning-based approaches, such as a denoising autoencoder (DAE) [4] and binary masking [5], are also currently being used for noise-robust ASR instead of the more traditional methods. Although speech enhancement is the simplest approach to noise-robust ASR, the distortion caused by speech enhancement often deteriorates ASR performance. This performance degradation is noteworthy in recent deep neural network (DNN)-based ASR frameworks, and is particularly noticeable in single-channel processing. In contrast, the minimum variance distortionless response beamformer (MVDR-BF) method [6], which is a multi-channel processing technique designed within a distortionless framework, provides a significant improvement in noisy speech ASR [7].

As alternatives to front-end processing, advanced DNN-based acoustic modeling frameworks have also been proposed, including noise-aware training (NAT) [8], joint training (JT) [9], and multi-task learning (MTL) [10]. NAT uses an estimated statistics (mean vector) of noise as an auxiliary input feature, and trains the acoustic model (AM) by considering the characteristics of noise. JT concatenates DAE and the AM, and jointly optimizes the entire parameter set. MTL often uses DAE (feature mapping from noisy to clean speech) as an auxiliary task to eliminate the influence of noise. Adaptations of DNN-based approaches have also been proposed [11][12]. In contrast to the above methods, precise acoustic modeling with complicated network architectures has attracted attention as an alternative to simple fully-connected feed-forward networks. In particular, convolutional neural networks (CNNs) [13][14][15], recurrent neural networks with long short-term memory (LSTM) [16], and convolutional LSTM (CLSTM) [17][18] are well-known and powerful tools for accurate acoustic modeling.

The effectiveness of MVDR-BF is well known, and several excellent studies related to this approach have been reported [7][15][19][20]. However, multi-channel processing requires special hardware, such as a microphone array and a multi-channel microphone amplifier. These hardware requirements are quite unsuitable for speech applications used in mobile environments. As our main target platform is mobile devices, specifically smartphones, processing single-channel speech input within a one-pass framework in real time (low latency) is vital for maintaining a usable form of the application. Adopting an AM adaptation with auxiliary feature parameters, such as i-vector [21], requires iterative batch-based processing, which is difficult under these constraints. As already mentioned, most existing approaches to single-channel speech enhancement seriously deteriorate the accuracy of ASR because of speech distortion. In single-channel processing, advanced acoustic modeling frameworks are also relatively ineffective.

In this paper, we tackle the above problem head-on, and attempt to reduce the influence of speech distortion by building a precise AM with a feature-level combination of noisy speech and enhanced speech. In addition, we adopt a sub-network-level combination of noisy and enhanced speech, and a gating mechanism that can dynamically select appropriate speech features, resulting in more precise acoustic modeling of single-channel noisy speech. We evaluate the proposed method on the CHiME4 1-channel track [22] and the corpus of spontaneous Japanese (CSJ) [23][24]. The evaluation results reveal that the proposed method successfully improves the accuracy of ASR in noisy environments, even under the constraints of one-pass single-channel processing.

http://dx.doi.org/10.21437/Interspeech.2019-1270

2. Reevaluation of conventional methods

We first reevaluate conventional noise robust ASR methods including speech enhancement and acoustic modeling frameworks using the CHiME4 1-channel track [22], and identify their limitations. Because our aim is to ensure the noise robustness of ASR with low latency and one-pass single-channel processing, we consider methods that satisfy these conditions.

2.1. Experimental setup

The CHiME4 corpus was recorded using a tablet device equipped with six microphones in various noise environments: a public transportation platform, cafeteria, pedestrian area, and at a street intersection. The training set consists of 1,600 real and 7,138 simulated (Simu) utterances spoken by four and 83 different people, respectively. The amount of training data is about 18.0 hours and the vocabulary size is 5k words. The development (Dev) and evaluation (Eval) sets consist of 3,280 and 2,640 utterances, respectively, each containing equal quantities of real and simulated data. Both the real and simulated sets were spoken by four speakers. The development set was used for cross validation during AM training and parameter tuning. In the CHiME4 corpus 1-channel track, the speech data recorded by the fifth microphone was used for an evaluation.

All AMs were trained using TensorFlow [25], and ASR decoding with trained AMs was conducted using the Kaldi toolkit [26]. The input feature parameters were 40 log mel-filter bank features (FBanks) and their first and second derivatives, which were extracted using a Hamming window with a 25-ms frame length and 10-ms frame shift. A context window with 11 (± 5) frames was applied to each utterance, so the dimension of the input feature parameters was 1,320. The target labels, which consisted of 1,967 context dependent-hidden Markov model (CD-HMM) states, were obtained using the Kaldi CHiME4 recipe [27]. Based on this setup, we trained an AM consisting of a fully connected feed-forward network with seven hidden layers. Each hidden layer consisted of 2,048 units and a rectified linear unit (ReLU) [28] activation function. In the training phase, weight decay (penalty: 0.0002) [29], batch normalization [30], and dropout (keep probability: 0.5) [31] were applied to prevent over-fitting. The parameters of the AM were randomly initialized and optimized using momentum stochastic gradient descent with a cross entropy criterion. A mini-batch of 128 frames and an initial learning rate of 0.01 were used for optimization.

Language modeling also followed the Kaldi CHiME4 recipe. The ASR experiments were performed using fully composed trigram weighted finite state transducers with the AMs. The evaluation criterion was the word error rate (WER).

2.2. Evaluation of enhanced speech

We evaluated four speech enhancement methods, namely SS, MMSE-STSA, GMM-FE, and DAE. All these methods are capable of frame by frame real-time processing. In this evaluation, the AMs were trained using the corresponding enhanced speech (FBanks). The GMMs for GMM-FE and DAE were trained using clean simulation data and real data recorded using a headset microphone. The GMM consisted of 512 Gaussian distributions. The DAE was trained using uni-directional CLSTM [18] as shown in Fig. 1.

Table 1 presents the WERs obtained for each speech enhancement method. In the table, Baseline and BeamformIt [32] are the results obtained without speech enhancement and with the beam-former in the Kaldi recipe for the CHiME4 6-channel track, respectively. As shown in the table, the WERs are worse than the baseline results in most environments when using speech enhancement. Because DAE trains a mapping function from noisy FBanks to clean FBanks, it outperforms the other methods. However, its improvement was slight. The crucial factor in these degradations is the aforementioned speech distortion caused by speech enhancement. Thus it is essential to reduce speech distortion as much as possible for noise robust ASR. However, it is difficult to reduce distortion when using single-channel processing, and there is a major gap between this and multi-channel processing based on a distortion-less scheme such as BeamformIt or MVDR-BF.

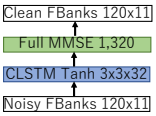

Figure 1: *Network structure of DAE. The CLSTM layer consists of 32 filters with 3×3 shape and a hyperbolic tangent activation function.*

Table 1: *ASR results for speech enhancement with the CHiME4 1-channel track in terms of WER (%)*

Enhancement method	Dev set		Eval set	
	Simu	Real	Simu	Real
Baseline	13.2	14.2	15.1	**22.6**
SS	13.5	14.6	15.2	22.9
MMSE-STSA	13.2	14.3	15.2	22.8
GMM-FE	13.2	14.2	15.2	23.0
DAE	**13.0**	**13.9**	**15.0**	23.3
BeamformIt [32]	*11.4*	*9.8*	*16.4*	*16.4*

2.3. Evaluation of acoustic modeling frameworks

Next, we evaluated several acoustic modeling frameworks, including NAT, JT, and MTL. All of these methods are capable of one-pass real-time processing. Figure 2 illustrates the AM structure for each method. NAT used the mean vector of noise as the auxiliary feature, which was estimated using the first ten frames of each utterance. JT concatenated the DAE shown in Fig. 1 and the AM trained with clean FBanks. All parameters were then jointly optimized. The AM structure for MTL is shown in Fig. 2(c). The loss function $\mathcal{L}$ for MTL was $\mathcal{L} = \mathcal{L}_{main} + \alpha \cdot \mathcal{L}_{aux}$, where $\mathcal{L}_{main}$ and $\mathcal{L}_{aux}$ denote the loss functions for the main and auxiliary tasks, respectively, and α denotes the weight for the auxiliary task. MTL used clean FBanks as the target signal for the auxiliary task. Therefore, $\mathcal{L}_{aux}$ was given by the MMSE criterion and α was set to 0.4. All AM structures and parameters were determined by preliminary experiments with development sets.

The WERs obtained from each acoustic modeling framework are presented in Table 2. As shown in the table, the improvement over the baseline results was negligible with the exception of JT. In JT, an overall average WER improvement of approximately 1.4% was achieved through the effects of both DAE speech enhancement and joint optimization. However, in the case of unknown noise condition, the ASR performance may seriously deteriorate due to the increase of speech distortion.

3. Details of the proposed method

From the ASR results presented in Sec. 2, although JT achieved some improvement, noise-robust ASR with one-pass single-channel processing remains a difficult problem. Therefore, we

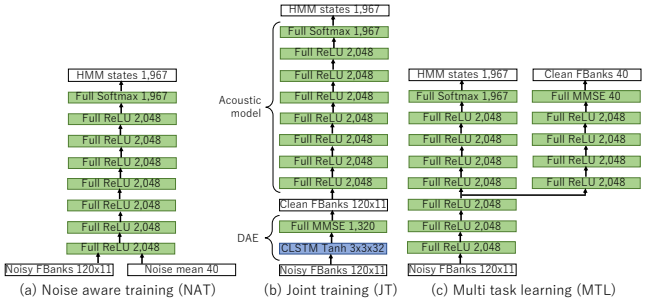

Figure 2: *Network structures for each advanced acoustic modeling framework*

Table 2: *ASR results for each acoustic modeling framework with the CHiME4 1-channel track in terms of WER (%)*

Network type	Dev set		Eval set	
	Simu	Real	Simu	Real
Baseline	13.2	14.2	15.1	22.6
NAT	13.3	14.4	15.2	22.3
JT	**12.3**	**12.5**	**14.1**	**20.7**
MTL	13.1	14.1	15.1	22.3

propose a new single-channel noise-robust ASR framework. The methods proposed do not use multiple passes, multiple channels, batch processing, or high-latency processing.

3.1. Combination of noisy and enhanced features

As mentioned in Sec. 2.2, the WERs become worse in most environments because of speech distortion caused by single-channel speech enhancement. However, when enhanced FBanks are analyzed over a short-term period, the signal-to-noise ratio (SNR) of speech-dominant periods such as vowel regions is generally high. In these periods, it can be assumed that the influence of speech distortion decreases even in enhanced FBanks. In low-SNR periods where the speech power is relatively weak, such as regions of speech onset/offset and unvoiced consonants, the influence of distortion may increase. Therefore, it is better to avoid using low-SNR periods of enhanced FBanks for ASR systems. Based on these observations, the proposed method uses both noisy FBanks and enhanced FBanks in the AM. We call this method feature-level combination, and note that it is possible to achieve both speech enhancement in high-SNR periods and low distortion in low-SNR periods.

3.2. Sub-network-level combination

Next, we investigate a sub-network-level combination. In this method, each feature is propagated to the individual sub-networks, and each output is then concatenated into a single data flow. This method realizes an advanced feature combination by stacking nonlinear feature transformations from individual sub-networks.

3.3. Gating mechanism

In both the feature-level and sub-network-level combinations, although the influence of speech distortion is expected to be small in the high-SNR periods, it may sometimes lead to a higher WER, even with slight speech distortion. To overcome this problem, we introduce a gating mechanism used for LSTM and highway networks [33].

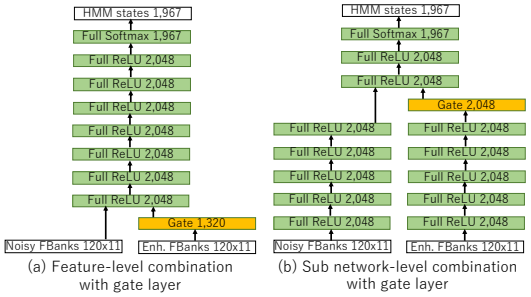

Figure 3: *Proposed AM structures with feature/sub-network-level combinations and gating mechanism*

In the t-th frame, when the input vector $\boldsymbol{x}_t$ is given, the output vector from the gate layer is $\boldsymbol{y}_t = \sigma\left(\boldsymbol{W}\boldsymbol{x}_t + \boldsymbol{b}\right) \odot \boldsymbol{x}_t$, where $\boldsymbol{W}$ and $\boldsymbol{b}$ denote the weight matrix and bias vector for the gate layer, respectively. The notation $\sigma(\cdot)$ and $\odot$ denote the sigmoid function and the Hadamard product, respectively. This mechanism realizes dynamic and appropriate information selection from enhanced FBanks.

Finally, we construct advanced AM structures with feature/sub-network-level combinations and a gate layer, as shown in Figs. 3(a) and 3(b). For the sub-network-level combination, the numbers of hidden layers in the sub-network and the subsequent network were determined to be five and two, respectively, based on preliminary experiments using the CHiME4 development set.

3.4. Use of various speech enhancement methods

It has been shown empirically that existing speech enhancement methods can either effective or ineffective in different noise environments depending on the algorithm used. Thus, solving the problem of single-channel noise-robust ASR using only one speech enhancement method is not a good strategy. Instead, the effective use of various speech enhancement methods is crucial for solving the problem. The AM structures of the proposed method, shown in Fig. 3, have multiple input schemes for both noisy FBanks and enhanced FBanks. The proposed method can handle AM structures that accept various types of enhanced FBanks inputs, as long as sufficient computational resources are available. Each enhanced FBanks is then propagated through the individual sub-networks and the gate layer.

4. Evaluation of proposed methods

Table 3 presents the ASR results for the CHiME4 1-channel track obtained using the proposed methods, where ALL denotes the results obtained with the method described in Sec. 3.4, and all enhanced FBanks obtained from SS, MMSE-STSA, GMM-FE, and DAE were used as input for the AM. As seen in the table, the feature-level combination in Fig. 3(a) improved the WERs compared with the results in Table 1, although little improvement over the baseline was achieved. In contrast, with each speech enhancement method, an average WER improvement of more than 1% over the baseline results was obtained using the sub-network-level combination in Fig. 3(b). This AM structure produced the best performance in almost all cases. These results reveal that the proposed framework incorporating sub-network-level combinations and a gating mechanism provides effective noise-robust ASR.

Table 3: *ASR results for the proposed methods with the CHiME4 1-channel track in terms of WER (%).*

Enhancement method	sub-network	Dev set		Eval set	
		Simu	Real	Simu	Real
Baseline		13.2	14.2	15.1	22.6
SS		13.2	13.8	15.0	22.4
	✓	12.3	13.0	14.1	21.2
MMSE-STSA		13.0	13.6	14.8	21.9
	✓	12.3	12.8	13.9	21.0
GMM-FE		13.0	13.5	15.0	22.2
	✓	12.4	12.8	14.1	21.1
DAE		12.6	13.7	14.3	22.4
	✓	11.8	12.4	13.5	20.5
ALL		12.8	13.7	14.4	21.9
	✓	**11.3**	**11.5**	**12.7**	**19.4**

The results for ALL indicate that using four speech enhancement methods significantly outperforms a single speech enhancement method. These results were obtained by selecting an appropriate speech enhancement method according to the target noise environment. In this paper, although four representative speech enhancement methods, namely SS, MMSE-STSA, GMM-FE, and DAE, were used to reveal the effectiveness of the proposed acoustic modeling framework, other approaches besides these four methods could also be implemented. Therefore, it is important to expand the applicable noise environments of the proposed method by selecting an appropriate speech enhancement method. In addition, it is necessary to improve the performance of each individual speech enhancement method.

5. Evaluation on the large scale corpus

We also evaluated the proposed method on a large scale corpus (the CSJ [23][24]) to justify effectiveness of the proposed method.

5.1. Experimental setup

The CSJ consists of recordings of academic lectures in Japanese. We used 957 lectures (240.0 hours) as the training (Train) set. Ten lectures (2.0 hours) were selected as a cross validation (CV) set during training. Three official evaluation (Eval) sets, E01 (2.0 hours), E02 (2.1 hours), and E03 (1.4 hours), were used for ASR evaluation. Each evaluation set consisted of ten lectures. The CSJ was recorded in clean conditions with a headset microphone, so we artificially added four noise environments (airport lobby, exhibition, shopping mall, and train station) to each data set in randomly selected SNR ranges of 0 dB to 10 dB. The noise data were taken from the ATR ambient noise sound database [34]. For evaluation data, we designed closed domain data and open domain data. Details of the noise conditions in each data set are indicated in Table 4.

The vocabulary size of the CSJ is approximately 75k words. The target labels, which consist of 9,512 CD-HMM states, were obtained using the Kaldi CSJ recipe [35]. The other conditions for acoustic modeling were the same as those in the aforementioned CHiME4 evaluations. The WERs for each evaluation set under the clean conditions were 11.1% for E01, 8.7% for E02, and 11.7% for E03.

5.2. Experimental results

Tables 5, 6, and 7 present the ASR results for the CSJ obtained using speech enhancement methods, acoustic modeling frameworks, and the proposed method with sub-network-level combinations and the gating mechanism shown in Fig. 3(b), respectively. As seen in each Table, the results obtained using conventional methods and the proposed method indicate similar tendencies to the evaluation results using the CHiME4 1-channel track presented in Tables 1, 2, and 3. These results demonstrate that the proposed method is effective regardless of the ASR task. In particular, they confirm that the improvements gained using the method described in Sec. 3.4 (ALL) are remarkable.

Table 4: *Noise conditions for evaluation using the CSJ*

Data set	Noise type	SNR ranges
Train	Airport lobby and exhibition	0–10 dB
CV	Airport lobby and exhibition	0–10 dB
Eval (closed)	Airport lobby and exhibition	0–10 dB
Eval (open)	Shopping mall and train station	0–10 dB

Table 5: *ASR results for speech enhancement with the CSJ in terms of WER (%)*

Enhancement method	Closed domain			Open domain		
	E01	E02	E03	E01	E02	E03
Baseline	**16.0**	14.1	17.3	**21.1**	21.9	21.7
SS	16.1	14.1	17.3	21.3	21.9	21.8
MMSE-STSA	16.1	**13.9**	**17.2**	**21.1**	22.0	21.7
GMM-FE	16.1	14.0	17.3	21.2	**21.8**	**21.5**
DAE	**16.0**	**13.9**	**17.2**	21.2	22.0	21.8

Table 6: *ASR results for each acoustic modeling framework with the CSJ in terms of WER (%)*

Network type	Closed domain			Open domain		
	E01	E02	E03	E01	E02	E03
NAT	16.0	14.1	17.3	21.1	21.8	21.6
JT	**15.6**	**13.4**	**16.9**	**20.8**	**20.2**	**21.1**
MTL	16.0	13.9	17.1	21.0	21.7	21.6

Table 7: *ASR results for the proposed method (sub-network-level combinations and gating mechanism) with the CSJ in terms of WER (%).*

Enhancement method	Closed domain			Open domain		
	E01	E02	E03	E01	E02	E03
SS	14.9	12.9	16.1	19.8	20.1	20.0
MMSE-STSA	14.9	12.8	16.0	19.8	20.1	20.0
GMM-FE	14.8	12.8	15.9	19.6	19.7	19.6
DAE	14.7	12.7	15.9	19.6	19.7	19.8
ALL	**13.9**	**11.8**	**14.9**	**18.6**	**18.3**	**18.6**

6. Conclusions

In this paper, we have proposed a noise-robust ASR method that is effective under one-pass single-channel processing conditions. The proposed method successfully reduces the influence of speech distortion caused by speech enhancement by using a combination of noisy speech and enhanced speech features as the input for speech recognition. We also examined the advanced structures of AMs, and obtained further improvements by introducing a sub-network-level combination of noisy and enhanced speech and dynamic selection of appropriate features with a gating mechanism. In this paper, we used fully connected feed-forward networks for the AMs. In the future, we will study the use of network structures suitable for analyzing spatiotemporal information, such as CNN, LSTM, and CLSTM. Additionally, we will attempt to improve the performance of each speech enhancement method, and reduce the amount of parameters introducing a bottleneck structure into the sub-networks.

7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on ASSP*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[3] J. C. Segura, A. d. l. Torre, M. C. Benítez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using AURORA II database and tasks," in *Proc. of Eurospeech '01*, vol. I, September 2001, pp. 221–224.

[4] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of Interspeech '13*, August 2013, pp. 436–440.

[5] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Proc. of ASRU '13*, December 2013, pp. 279–284.

[6] E. A. P. Habets, J. Benesty, S. Gannot, and I. Cohen, *Speech processing in modern communication–Challenges and perspectives, Chapter 9: The MVDR beamformer for speech enhancement*. Springer–Verlag, Decemver 2009.

[7] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. of ICASSP '16*, March 2015, pp. 5210–5214.

[8] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP '13*, May 2013, pp. 7398–7402.

[9] K. H. Lee, T. G. Kang, W. H. Kang, and N. S. Kim, "DNN-based feature enhancement using joint training framework for robust multichannel speech recognition," in *Proc. of Interspeech '16*, September 2016, pp. 3027–3031.

[10] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *Proc. of ICASSP '16*, March 2016, pp. 5075–5079.

[11] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. of ICASSP '13*, May 2013, pp. 7947–7951.

[12] J. Li, J. T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Proc. of ICASSP '14*, May 2014, pp. 5537–5541.

[13] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE Trans. on ASLP*, vol. 22, no. 10, pp. 1533–1545, October 2014.

[14] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[15] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. of ASRU '15*, December 2015, pp. 436–443.

[16] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. of Interspeech '14*, September 2014, pp. 338–342.

[17] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. of NIPS '15*, December 2015, pp. 802–810.

[18] M. Fujimoto and H. Kawai, "Comparative evaluations of various factored deep convolutional RNN architectures for noise robust speech recognition," in *Proc. of ICASSP '18*, April 2018, pp. 4829–4843.

[19] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single–channel mask prediction networks," in *Proc. of Interspeech '16*, September 2016, pp. 1981–1985.

[20] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *Proc. of ICASSP '17*, March 2017, pp. 3246–3250.

[21] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on ASLP*, vol. 19, no. 4, pp. 788–798, May 2011.

[22] "The 4th CHiME speech separation and recognition challenge," http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/.

[23] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, April 2003.

[24] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the corpus of spontaneous Japanese," in *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, April 2003.

[25] "TensorFlow," https://www.tensorflow.org/.

[26] "Kaldi ASR tool-kit," http://kaldi-asr.org/.

[27] "Kaldi CHiME4 recipe," https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4.

[28] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. of AISTATS '11*, April 2011, pp. 315–323.

[29] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *in Proc. of NIPS '91*, December 1991, pp. 950–957.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of ICML '15*, July 2015, pp. 448–456.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. Volume 15, no. Issue 1, pp. 1929–1958, January 2014.

[32] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on ASLP*, vol. 15, no. 7, pp. 2011–2023, September 2007.

[33] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. of NIPS '15*, December 2015, pp. 2377–2385.

[34] T. Endo, T. Horiuchi, T. Shimizu, and S. Nakamura, "Speech recognition experiments with ATR ambient noise sound database – ATRANS –," in *Prpc. of IPSJ SIG Technical Report*, no. 2005–SLP–57 (8), July 2005, pp. 43–48, (in Japanese).

[35] T. Moriya, T. Tanaka, T. Shinozaki, A. Watanabe, and K. Duh, "Automation of system building for state–of–the–art large vocabulary speech recognition using evolution strategy," in *Proc. of ASRU '15*, December 2015, pp. 610–616.