Speech Recognition-based Evaluation of a Noise Reduction Method in Known-Noise Environment

Chee Siang Leow[1,a) Hiromitsu Nishizaki[1,b) Akio Kobayashi[†1,c)
Takehito Utsuro[†2,d)

Abstract: This paper investigated a noise reduction method from speech which was recorded in the noisy environment in a factory, and the noise reduction method was evaluated in speech recognition experiment. In the proposed, first, noise is estimated from the speech which should be speech-recognized, and then, the noise sounds are superimpose to clean speeches. The noise-mixed speeches are used to train the noise reduction model. In the experiment of noisy speech recognition using the model, the WER of the noise-reduced speech reduced more than 10%.

1. Introduction

In recent years, many applications, such as Google Home®, Amazon Alexa® have shown the usefulness of automatic speech recognition (ASR) in real-life applications. The improvement of performance in ASR leads to these various applications to be able to be developed. Most of the state of the art ASR systems are trained with clean speech data for preventing the mismatch of data distribution. To maintain the performance of the ASR, frontend preprocessing such as voice activity detector (VAD) [1] and speech enhancement is often being to use to preprocess the input speech. In a noisy environment, a stable way to clean the speech is by using a multichannel microphone-array beamforming method [2]. However, for a single-channel ASR system, it still suffers from environmental noise, which causes a degradation of the performance.

Recently, many deep learning-based speech enhancement [3] has been proposed by directly estimating the clean log magnitude spectrum(LMS) [4], [5], [6] or estimating the ideal ratio mask (IRM) [7], [8], [9] from a noisy power or magnitude spectrum. However, most proposed methods were evaluated with Perception Speech Quality (PESQ) [10],short-time objective intelligibility (STOI)[11] metrics, or experiment on public datasets [12], [13], [14], [15], [16], [17] which could lead to the mismatch target noise to denoise in real noisy environments.

In our previous work [18] on a semi-automatic task manual creation from an instruction talk on machining operation, we found that ASR could help the factory engineers to make the task manual by transcribing the engineer's talk while they are explaining their specific skill with a single-channel microphone [*1]. However, plant engineers often work in a noisy environment while the various machine is running, such as milling machine, NC machine tool, or hammer hit sounds. These types of sounds can be considered as a known target noise that we would like to denoise these factory noises. For improving ASR performance in such a noisy environment, a frontend speech enhancement system is needed, and it should be trained with a low resource of target noise only from the known noise environment.

In this study, we evaluate methods of estimating the ideal ratio mask or directly estimating the clean LMS from a noisy sound. We tested with the various architectures of deep denoising neural networks (DDNN) to denoise speech with noise and evaluated the denoised speech by using an open-source ASR system Kaldi [19], in which an acoustic model was trained with only clean speech. In the experiments, our DDNN models improved the speech quality in the subjective perceptional experiment, and they also achieved 41.1% of the word error rate (WER) from 51.2% of the original speech without any denoising method.

The contributions of this paper are as follows:

- This paper first shows a way to use a very-low noise resource to train a DDNN model.
- This paper experimentally shows that even the very-low noise resource is sufficient to train a DDNN model because the trained model can generate the noise-reduced

[1] Integrated Graduate School of Medicine,Engineering and Agricultural Sciences,University of Yamanashi, Japan
[†1] Presently with Department of Industrial Information, Tsukuba University of Technology, Japan
[†2] Presently with Graduate School of Systems and Informaation Engineering, University of Tsukuba, Japan
[a)] cheesiang_leow@alps-lab.org
[b)] hnishi@yamanashi.ac.jp
[c)] a-kobayashi@a.tsukuba-tech.ac.jp
[d)] utsuro@iit.tsukuba.ac.jp

[*1] In a plant, an engineer cannot use any special microphone device for speech recording.

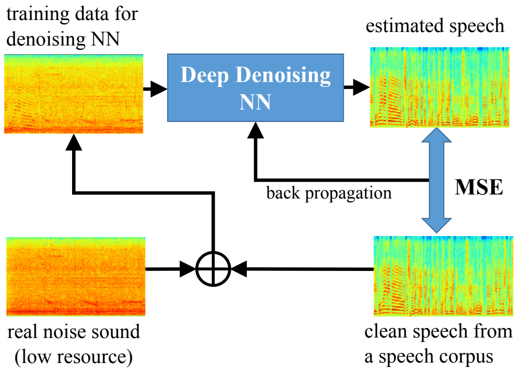


Fig. 1 Flow of making synthesize data and training a deep denoising neural network.

speech well against the noisy speech recorded in the real environment at the machining plant.

The remaining of this paper is organized as follows. In Section 2, we will show a synthesis method to make a training dataset to train a DDNN model. Then, Section 3 describes the architecture of the DDNN models, and Section 4 introduces denoising experiments. Finally, we will conclude our research and describes future works in Section 5.

2. Synthesize Noisy Speech for DDNN Training

2.1 Trainnig Pipeline

The flow of our training pipeline is shown in Fig.1. To synthesize the training data for DDNNs, we used an operation video, including an introduction speech recorded in a real plant, from the previous work [18]. From the video, we extracted the non-speech part of the video as noise; that is known noise. The total duration of noise extracted from the video is only 56 seconds. We use this low resource noise sound to create DDNN training data. The noise data is being augmented with clean speeches by mixing the factory noise with clean speeches to produce noisy-clean speech pairs.'' The short duration of noise extracted from the video is randomly concatenated to match the length of clean speech. We compute the log10 LMS with the following formula.

$$S = \log_{10} |\mathcal{F}[Y]| \tag{1}$$

2.2 Data Augmentation

The noisy speech training data is mixed in a range of -10 dB, -5 dB, and 0 dB to 20 dB. This allows us to increase the number of training data by 23 times to train the model. The negative dB range that we choose as -5 dB and -10 dB is because the noise level of our environment is about 0 dB to 5 dB. To prevent the model from overfitting on negative dB, we make the positive dB to be more data. We use the Corpus of Spontaneous Japanese (CSJ) [20] as a clean speech dataset. The clean speeches from CSJ are randomly chosen in 4,988 utterances; the total duration is about 7.6 hours. Finally, we can get 114,724 noisy-clean speech pairs (174.8 hours) for DDNN training by the data augmentation (speech and noise sound synthesis).

3. Speech Denoising Model

3.1 Model Structure

To denoise the frame-wise noisy log LMS, we used recurrent-based DDNNs which consist of a gated recurrent unit (GRU) [21] or long short-term memory (LSTM) unit [22]. Besides, we adopt convolutional neural network (CNN) layers for acoustic feature extraction. The architectures of DDNNs are shown in Fig.2. We prepare four types of DDNNs; (a) GRU-based DDNN ("GRU"), (b) the combination of CNN and Bi-directional LSTM layers ("CNN-BiLSTM"), (c) LSTM-based DDNN with a local attention mechanism ("LSTM w/ Att."), and (d) BiLSTM-based DDNN which estimates directly clean magnitute spectrum and a noise mask ("MT-BiLSTM"). "(a) GRU" in Fig. 2 consists of two layers of GRU with 1024 units, a batch normalization layer and 257 units of a dense layer with sigmoid function. In "(b) CNN-BiLSTM" we use three layers of CNN with filters of 32/64/128 whose kernels are 2x2/3x3/3x3 respectively, and two layers of BiLSTM with 256 units with feature concatenation. In addition, we adopted an attention mechanism-based speech enhancement[23] in "(c) LSTM w/ Att. " "LSTM w/ Att." has an local attention method in between the two layer of LSTM with a context length of five frames. Finally, "(d) MT-BiLSTM" is composed of simple two layers of BiLSTM with multi-task (MT) loss which adopted from [9]. All of the models are trained with mean squared error (MSE) loss. Note that the activation function of the output layer should use sigmoid function to estimate an ideal ratio mask in each DDNN except for ("LSTM w/ Att.") model, and we do not use any activation function at the output layer when a log LMS is directlry estimated.

3.2 Local Attention Layer

The local attention layer is shown in Fig. 3. The local attention layer is adopted by extending Bahdanau Attention[25] which first calculates the score by following formula:

$$\bar{x} = TANH(x) \tag{2}$$

$$EO, H = LSTM(\bar{x}) \tag{3}$$

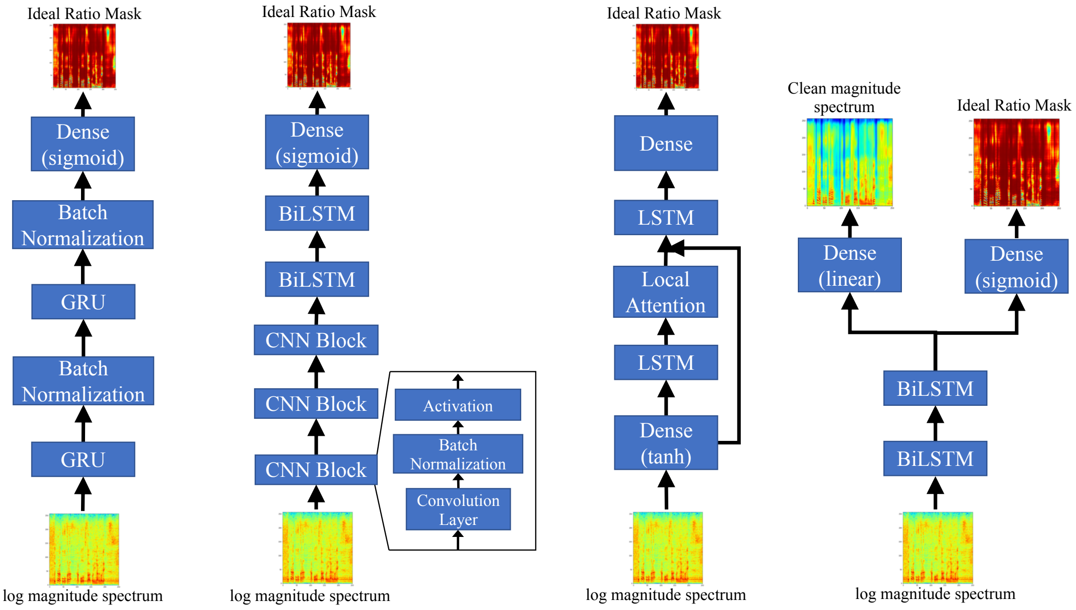
$$score = FC(TANH(FC(EO) + FC(H))) \tag{4}$$

where, the FC is a fully connected layer, TANH is a fully connected layer with hyperpolic tangent activation function, H is the LSTM layer states, EO is the outputs of a LSTM encoder layer. Then, we added the local attention with context width of 4 which uses the casual 4 frames time steps,$[x_{t-4}, ..., x_t]$ and multiplies to the current frames. Next, we compute the attention weights and the context vector by the following formulas:

$$attention\_weights = softmax(score) \tag{5}$$

$$context\_vector = EO \times attention\_weights \tag{6}$$

Furthermore, the inputs from $\bar{x}$ is multiply to the *context_vector* to get the *attention_weighted_vector* before



(a) GRU (b) CNN-BiLSTM (c) LSTM w/ Att. (d) MT-BiLSTM

Fig. 2 Various DDNN architectures.

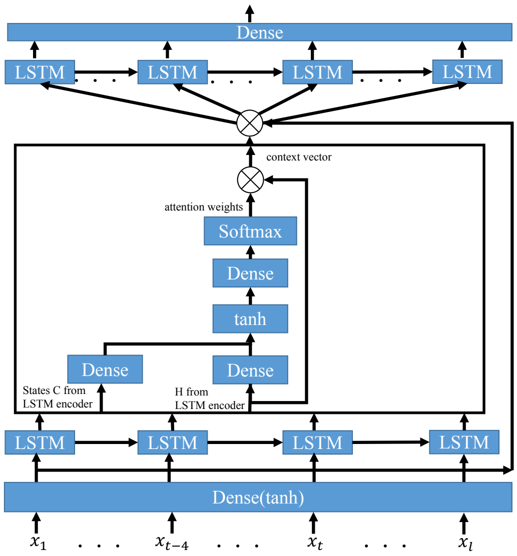


Fig. 3 Local attention layer

next LSTM layer, which is the following formula:

$$attention\_weighted\_vector = context\_vector \times \bar{x} \quad (7)$$

The attention_weighted vector is passed to the lstm and dense layer to estimate the IRM.

3.3 Training Losses

All the DDNN models estimate an ideal ratio mask [7] which is used to decide the ratio of target domain power. An ideal ratio masks $IRM(t)$ is computed with the following formula:

$$IRM(t) = (\frac{S_{Noise}(t)^2}{S_{Noise}(t)^2 + S_{Clean}(t)^2})^{\frac{1}{2}} \quad (8)$$

, where $S_{Noise}(t)$ is the noisy log10 LMS, $S_{Clean}(t)$ is the clean log10 LMS. Input features inputted to a DDNN are extracted as a log LMS with a hamming window size of 25 ms, shift 10 ms which produces each frame as 257 dim. of Discrete Fourier transform (DFT) bins. The log10 LMS is normalized on each bin to mean 0, global variance 1. Note that the clean speech data is used as a label and also for computing the ideal ratio masking labels. For the multi-task (MT) model (d), we defined the loss as the following formula (9):

Table 1 Hyper-parameters for DDNN training

Hyper-parameters	Condition
Mini-batchsize	100
Num. of epochs	20
Hidden layer Activation	ReLU
Dropout	0.3
Loss func.	Minimum Square Error
Optimizer	Adam
Init. learning rate	0.0001

$$loss_{total} = LMS\_loss_{MSE} + IRM\_loss_{MSE} \quad (9)$$

For the MT-BiLSTM model, it outputs the IRM and directly estimated LMS. Therefore, we can use the IRM to masked the directly estimated LMS to produce a clean speech.

4. Denoising Experiments

In the experiments, ewe trained each DDNN model with the hyper-parameters shown in Table 1. To evaluate WERs of ASR and speech quality, we should reconstruct the speech waveform from the denoised LMS. The denoised speech waveform Y can be calculated with the following formula (10) from the LMS using the estimated IRM:

$$Y = \mathcal{F}^{-1}[S_{Noise}(t) * (1 - IRM(t))] \quad (10)$$

Figure. 4 shows the estimated IRM and the LMS of clean speech with the respective DDNN model. From Fig. 4, we can see that most of the DDNNs can denoise the noisy LMS to the clean LMS well. However, although the ("LSTM w/

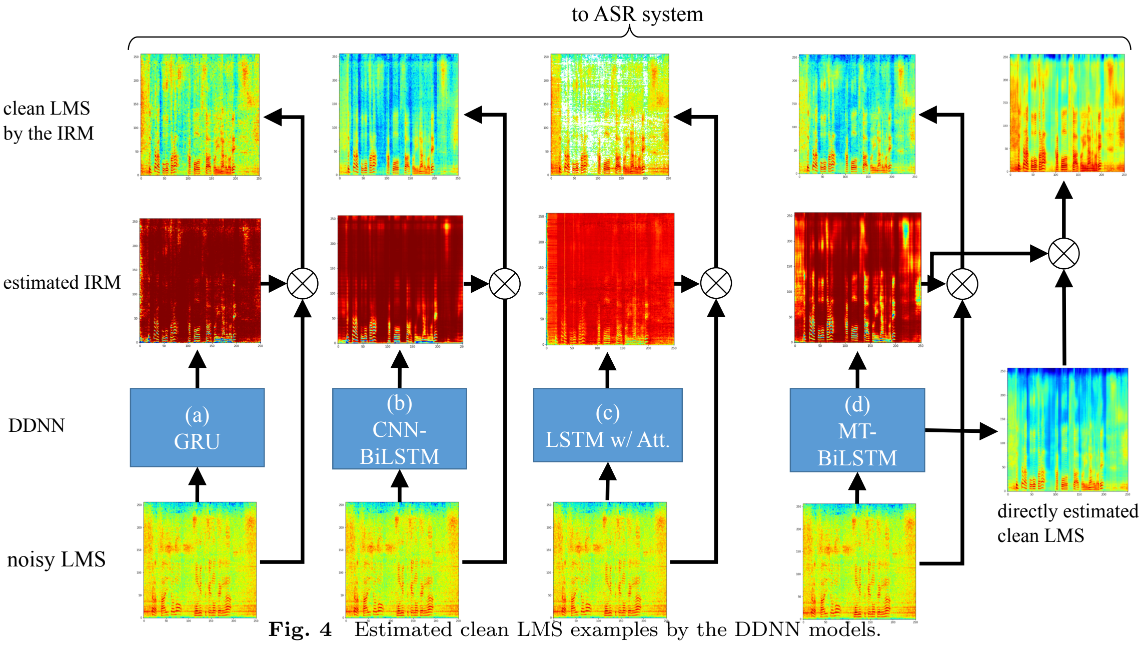


Fig. 4 Estimated clean LMS examples by the DDNN models.

Att.”) model can denoise the speech with including the white spectrogram area, this is due to the final layer which didn’t use the sigmoid function that made estimated mask value to be too big. We would like to investigate more with the (“LSTM w/ Att.”) model to find out the best way to use the attention for speeches denoising.

We conducted two experiments: perception experiment and ASR experiment for denoised speech. The sound waveforms are converted from the estimated clean LMSs based on the IRMs by the DDNN models. Then, we conducted a subjective evaluation of speech quality for the speech recorded in the machining plant, where various types of noise with high amplitude always occurred. Totally 14 subjects evaluated the denoised speech waveform and scored a range from 1 to 7 to calculate the mean opinion score (MOS). The subjective evaluation result is shown in Table2.

Table 2 shows that all the DDNN models can remove the noise from the noisy speech and the listenability of clean speech is better than the original noisy speech. We can see that “MT-BiLSTM” model of inputs-masked and directly estimated LMS had the score of 3.14, which performs the best to denoising the noisy speech, compared with the score 1.54 of the original noisy speech. This result can infer that the low resource noise is enough to train DDNNs to remove the specific noise from the noisy speech in perceptional experiment.

To evaluate our DDNN models from the point of view of ASR performance, first, we prepared the ASR system Kaldi, in which the acoustic and language models were trained with the “NNET1” setup of the Kaldi CSJ Recipe[24]. We conducted ASR experiments with two kinds of data. We used the 0 dB augmented noisy speech as closed data in which the data was used to train the DDNN models and the real noisy speech from the instruction video of machining operation recorded in the previous work[18].

Table 2 Perceptual evaluation of denoised speeches in MOS score (averaged for 14 subjects).

DDNN models	MOS
w/o denoising	1.54
(a) GRU	2.71
(b) CNN-BiLSTM	2.14
(c) LSTM-Attention	2.64
(d-1) MT-BiLSTM (inputs-masked)	3.14
(d-2) MT-BiLSTM (directly estimated LMS)	3.14
(d-3) MT-BiLSTM (LMS-masked)	2.64

Table 3 WERs [%] of each DDNN model.

DDNN model	closed	open
w/o denoising	94.9	51.2
(a) GRU	30.9	48.8
(b) CNN-BiLSTM	54.5	51.2
(c) LSTM w/ Att.	51.0	46.5
(d-1) MT-BiLSTM (input-masked)	45.5	**41.1**
(d-2) MT-BiLSTM (estimated LMS)	44.9	45.7
(d-3) MT-BiLSTM (LMS-masked)	75.9	55.0

Table 3 shows the ASR performance in WER metrics. The “MT-BiLSTM” model got the best performance in both the MOS and WER. From Fig.4, we can see the predicted IRM remains more in high-frequency bandwidth, which allows the ASR system which trained with clean only data to match the distribution more, comparing to the other denoising model. On the other hand, from Table 3, we found that most DDNN models have been able to remove the noise from the speech because the models achieved the improvement of the WERs. The best performance came from the “MT-BiLSTM” model against the IRM-adapted speech because it used the multi-task loss that could train the model robustly to be more generalized to denoise the speech.

5. Conclusion

We proposed the way to train the DDNN models with very-low noise resource extracted from the ASR target speech recorded in the noisy environment. We have tried four sorts of the DDNN models, which estimated the IRM to remove noise in the frequency domain. In the experiments, we showed that the DDNN model that trained with the very-low noise resource could easily remove the target noise from the speech recorded in the noisy-terrible environment because both perceptual evaluation by 14 subjects and the ASR performance were improved. In particular, the WER was reduced to 41.1% from 51.2% of without denoising the speech.

However, although the WER was improved, it is not enough to make the denoised speech to reach the state-of-the-art performance of ASR. The DDNN model with very-low noise resource in the speech enhancement research field should be done in other ways. Therefore, in future work, we are going to search a best neural network architecture for training denoising model especially model with attention layer, investigate the way to train a DDNN model and an acoustic model with connectionist temporal classification (CTC) loss simultaneously in the multi-task learning scheme for more improvement of the ASR performance.

Acknowledgments This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research, Grant Numbers 17H01977 and 19H01125.

References

[1] M. Fujimoto, K. Ishizuka, and T. Nakatani, *"A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in* Proc. of ICASSP 2018*, 2008, pp. 4441–4444.*

[2] Z. Zhang, X. Xiao, L. Wang, E. Chng, and H. Li, "Noise robust speech recognition using multi-channel based channel selection and channel weighting," reprint arXiv:1604.03276*, 2016.*

[3] Z. Zhang, J. Geiger, J. Pohjalainen, A. El-Desoky Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," ACM TIST, vol. 9, no. 5, 2018, pp.49:1–49:28.

[4] Xugang lu, T. Yu, M. Shigeki, and C. Hori, "Speech enhancement based on deep denoising auto-encoder," in Proc. of INTERSPEECH 2013*, 2013, pp. 436–440.*

[5] K. Nakazawa and K. Kondo, "De-reverberation using DNN for non-reference reverberant speech intelligibility estimation," in Proc. of GCCE 2018*, 2018, pp. 349–350.*

[6] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Trans. on ASLP*, vol. 23, no. 1, pp. 7–19, 2015.*

[7] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in Proc. of ICASSP 2013*, 2013, pp. 7092–7096.*

[8] S. Xia, H. Li, and X. Zhang, "Using optimal ratio mask as training target for supervised speech separation," in Proc. of APSIPA ASC 2017*, 2017, pp. 163–166.*

[9] Y. Tu, J. Du, N. Zhou, and C. Lee, "Online lstm-based iterative mask estimation for multi-channel speech enhancement and ASR," in Proc. of APSIPA ASC 2018*, 2018, pp.362–366.*

[10] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," in Proc. of ICASSP 2001 *, 2001, pp. 749–752,*

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in Proc. of ICASSP 2010*, 2010, pp. 4214–4217.*

[12] M. Kolbaek, Z. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," IEEE/ACM Trans. on ASLP*, vol. 25, no. 1, 2017, pp. 153–167.*

[13] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," Proc. of INTERSPEECH 2018*, 2018, pp.1561–1565.*

[14] S. Kim, B. Raj, and I. Lane, "Environmental noise embeddings for robust speech recognition," preprint arXiv:1601.02553*, 2016.*

[15] A. Varga and H. J.M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication*, vol. 12, no. 3, 1993, pp.247–251.*

[16] M. Zhao, D. Wang, Z. Zhang, and X. Zhang, "Music removal by convolutional denoising autoencoder in speech recognition," in Proc. of APSIPA ASC 2015*, 2015, pp. 338–341.*

[17] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in Proc. of ICASSP 2018*, 2018, pp.5024–5028.*

[18] C. S. Leow and H. Nishizaki, "A task manual creation support system using automatic speech recognition," in Proc. of GCCE 2018*, 2018, pp. 259–262.*

[19] P. Daniel, G. Arnab, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in Proc. of ASRU 2011*, 2011.*

[20] M. Kikuo, "Corpus of Spontaneous Japanese: Its design and evaluation," in Proc. of SSPR*, 2003.*

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

[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in Proc. of the NIPS 2014 Workshop on Deep Learning*, 2014.*

[23] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, "An attention-based neural network approach for single channel speech enhancement," in Proc. of ICASSP 2019*, 2019, pp.6895–6899.*

[24] T. Moriya, T. Tanaka, T. Shinozaki, S. Watanabe, and K. Duh, "Automation of system building for state-of-the-art large vocabulary speech recognition using evolution strategy," in Proc. of ASRU 2015*, 2015, pp. 610–616.*

[25] D. Bahdanau,K. Cho,Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in Proceedings of the International Conference on Learning Representations, (ICLR) *, 2015, pp.338–342.*