Noise-robust Attention Learning for End-to-End Speech Recognition

Yosuke Higuchi*, Naohiro Tawara†, Atsunori Ogawa†, Tomoharu Iwata†, Tetsunori Kobayashi*, Tetsuji Ogawa*
* *Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan*
† *NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan*
higuchi@pcl.cs.waseda.ac.jp

***Abstract*—We propose a method for improving the noise robustness of an end-to-end automatic speech recognition (ASR) model using attention weights. Several studies have adopted a combination of recurrent neural networks and attention mechanisms to achieve direct speech-to-text translation. In the real-world environment, however, noisy conditions make it difficult for the attention mechanisms to estimate the accurate alignment between the input speech frames and output characters, leading to the degradation of the recognition performance of the end-to-end model. In this work, we propose noise-robust attention learning (NRAL) which explicitly tells the attention mechanism where to "listen at" in a sequence of noisy speech features. Specifically, we train the attention weights estimated from a noisy speech to approximate the weights estimated from a clean speech. The experimental results based on the CHiME-4 task indicate that the proposed NRAL approach effectively improves the noise robustness of the end-to-end ASR model.**

***Index Terms*—Attention mechanism, noise robustness, speech recognition, deep neural networks**

I. Introduction

In recent years, extensive research attention has been devoted to developing a single deep neural network (DNN)-based end-to-end automatic speech recognition (ASR) model [1]–[5]. Various models and approaches have been proposed for enhancing the performance of the end-to-end ASR model such as recurrent neural networks (RNNs) with encoder-decoder architectures [6] and attention mechanisms [7], subsampling techniques [4], [8], [9], joint training with connectionist temporal classification (CTC) [10], and transformer with self-attention networks [11], [12].

Although the end-to-end ASR model is almost as effective as the traditional hybrid system [13], the end-to-end ASR model is known to be vulnerable to noise [10]. Because the end-to-end model is trained solely on paired speech and text data, the alignment estimated by the attention mechanism can be easily corrupted due to variations in the structure of acoustic signals arising from various sources, e.g. noises, channels, and speakers. To absorb these variations, the end-to-end model requires a large amount of data based on various conditions.

Making the ASR system robust against these variations is a long-standing research topic [14]. Several attempts have been made to deal with noises. For instance, some approaches have trained DNN-based acoustic models to be robust against noises by introducing additional noisy data recorded in different conditions [15] in a multi-style training manner [16]. Augmenting training data by simulating noises sampled from noisy datasets [17] has also proven to be effective at improving the performance of ASR [18] and other fields related to speech processing [19]. Other approaches to improving the noise robustness of DNN-based acoustic models are making bottleneck features, the outputs of a hidden layer, invariant to noise [20], [21]; some studies have proposed that an adversarial training technique be adopted in DNNs [22]. Several works have focused on improving the noise robustness of the end-to-end ASR model. [10] trained an attention-based encoder-decoder model using an auxiliary task of CTC [23], with the aim of resolving the corruption of attention alignments caused by noisy speech. Using a pair of clean and simulated noisy speech, [24] made the model's hidden representations invariant to noise by making the representations from noisy speech approximate ones from clean speech.

Our work aims to train a noise-robust end-to-end ASR model by making the attention weights estimated from a noisy speech approximate the weights estimated from a clean speech. In the end-to-end ASR model, the attention mechanism works as to estimate an alignment between the encoded speech frames and output characters. The alignment is expected to be consistent, even if noise is superimposed on the input clean speech to artificially converted it into noisy speech. We introduce a constraint that makes corrupted attention weights from noisy speech approximate the aligned weights; then, we explicitly tell the model where to "listen at" in the encoded speech frames that are relevant to producing each output. [24] trains a model to map both clean inputs and their noisy counterparts onto the same point in the representation space, making the model to produce close hidden representations between the clean and noisy speech. Although of all the existing related study, this approach is the most similar to ours, ours differ in the sense that we attempt to produce the same attention alignments between the paired data. Rather than coercing noisy speech into being mapped onto the same representation space as clean speeches, the proposed training furnishes the model with more abstract guides, leaving its parameters sufficiently flexible to be trained on noisy speech.

The rest of the paper is organized as follows. The proposed noise-robust attention learning (NRAL) framework is introduced in Section 2. Integration of the previous work [24] into NRAL is described in Section 3. The effectiveness of NRAL is evaluated based on the CHiME-4 task in Section 4. Finally, the conclusions are presented in Section 5.

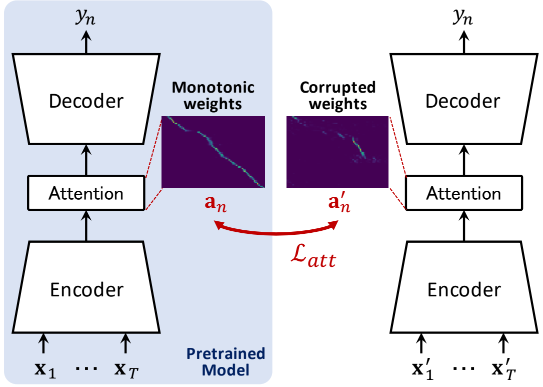


Fig. 1. Proposed noise-robust attention learning using pair of clean speech, $\boldsymbol{x}$, and noisy speech, $\boldsymbol{x}'$

II. Proposed Noise-robust Attention Learning Framework

Figure 1 is an overview of the proposed noise-robust attention learning (NRAL) framework. The proposed training framework explicitly teaches a model where to "listen at" in the encoded speech frames that are relevant to producing each output. The steps involved in training the model are as follows:

1) train a model using clean speech data, and obtain the aligned attention weights;
2) create a clean-noisy data pair by artificially adding noise to the original clean data; and
3) re-train the model using the simulated noisy data and the aligned attention weights.

The model trained via this proposed framework, owing to its ability to accurately estimate alignments under noisy conditions, is expected to be noise-robust. In the following subsections, we describe an RNN-based end-to-end ASR model that uses the attention mechanism and loss designed based on the attention weights in the proposed training framework.

A. End-to-End ASR model

1) Attention-based encoder-decoder: We use an end-to-end ASR model based on an attention-based encoder-decoder framework [3]. Unlike conventional ASR frameworks, this framework does not make any conditional independence assumptions; rather, it autoregressively estimates the posterior probabilities at each time-step conditioning on previous labels:

$$P(Y|X) = \prod_{n=1}^{N} P(y_n|X, y_{1:n-1}), \tag{1}$$

where $X = \{\mathbf{x}_t\}_{t=1}^{T}$ denotes the input speech frames with T length, $Y = \{y_n | y_n \in \{1, ..., K\}\}_{n=1}^{N}$; the output characters with N length; and K distinct labels, including special the start-of-sentence (sos) and end-of-sentence (eos) tokens.

The model consists of three networks: encoder, attention, and decoder.

$$\begin{aligned} H &= \mathrm{Encoder}(X), && (2)\\ \mathbf{c}_n &= \mathrm{Attention}(\mathbf{s}_{n-1}, H), && (3)\\ y_n &\sim \mathrm{Decoder}(\mathbf{c}_n, y_{1:n-1}). && (4) \end{aligned}$$

The encoder converts the input speech frames X into a sequence of high-dimensional representations, $H = \{\mathbf{h}_l\}_{l=1}^{L}$, as in Eq. (2,) where L is the number of frames of the encoder output. Then, the attention network calculates a context vector, $\mathbf{c}_n$, as shown in Eq. (3,) based on the encoder output H that are relevant to producing outputs in a decoder state, $\mathbf{s}_{n-1}$. Finally, given the context vector, $\mathbf{c}_n$, and previous character outputs, $y_{1:n-1}$, the decoder generates a character, y_n. The network architecture consists of a VGG layer incorporated into the bi-directional long short-term memory to model the encoder [9] and an LSTM [25] to model the decoder.

Based on the cross-entropy criterion, the objective function of the model is calculated as follows:

$$\mathcal{L}_{\mathrm{char}}(X) = -\log P(Y^*|X) = -\sum_{n=1}^{N} \log P(y_n^*|y_{1:n-1}^*), \tag{5}$$

where y_n^* denotes the ground truth of the character at n step.

2) Location-based attention mechanism: For the attention mechanism of the end-to-end model, a location-based attention mechanism [3] is adopted. The location-based attention mechanism utilizes the alignment produced at the previous time step, $\mathbf{a}_{n-1}$; Eq. (3) is revised as follows:

$$\mathbf{c}_n = \mathrm{Attention}(\mathbf{a}_{n-1}, \mathbf{s}_{n-1}, H). \tag{6}$$

The following equations represent the process of computing the attention weight, $\mathbf{a}_n$, and the context vector, $\mathbf{c}_n$, at time step n:

$$\begin{aligned} \mathbf{f}_n &= F * \mathbf{a}_{n-1}, && (7)\\ e_{n,l} &= w^T \tanh(W\mathbf{s}_{n-1} + V\mathbf{h}_l + U\mathbf{f}_n + b), && (8)\\ a_{n,l} &= \frac{\exp(\alpha e_{n,l})}{\sum_l \exp(\alpha e_{n,l})}, \quad \mathbf{c}_n = \sum_{l=1}^{L} a_{n,l}\mathbf{h}_l, && (9) \end{aligned}$$

where F is a trainable convolutional filter; $*$ denotes convolution; w, W, V, and U are trainable matrices; b is a trainable bias parameter; and α is a sharpening factor [3].

B. Proposed noise-robust attention learning using attention weights

Based on the assumption that the input speech frames and output characters align consistently, the attention weights learned from the clean speech can be used as meaningful information in improving noisy speech training. Given clean speech frames, $X = \{\mathbf{x}_i\}_{t=1}^{T}$, and their noisy counterparts, $X' = \{\mathbf{x}'_i\}_{t=1}^{T}$, the corresponding attention weights at time step n are represented as follows: $\mathbf{a}_n = \{a_{n,l}\}_{l=1}^{L}$ and $\mathbf{a}'_n = \{a'_{n,l}\}_{l=1}^{L}$. Here, because each attention weight is calculated using Eq. (9) based on softmax, $\mathbf{a}_n$ and $\mathbf{a}'_n$ can be regarded as multinomial distributions. Based on this assumption, we define a loss between the attention weights using the Kullback-Leibler divergence, and the attention loss is calculated as follows:

$$\begin{aligned}\mathcal{L}_{\text{att}}(X, X') &= \sum_{n=1}^{N} D_{\text{KL}}(\mathbf{a}_n || \mathbf{a}'_n), \qquad (10)\\ &= -\sum_{n=1}^{N}\sum_{l=1}^{L} a_{n,l} \log \frac{a_{n,l}}{a'_{n,l}}. \qquad (11)\end{aligned}$$

The proposed training framework aims to optimize the function that consists of the character classification loss in Eq. (5) and the attention loss in Eq. (10) as follows:

$$\mathcal{L}_{\text{NRAL}}(X, X') = \mathcal{L}_{\text{char}}(X') + \lambda \mathcal{L}_{\text{att}}(X, X'), \tag{12}$$

where λ denotes the tunable parameter that controls the weight for $\mathcal{L}_{\text{att}}$.

III. Integration of invariant representation learning

Similar to the training method proposed in this study, [24] proposes the invariant representation learning (IRL) whereby a model is trained to map both clean inputs and their noisy counterparts onto the same point in the representation space. Given clean speech frames, X, and their noisy counterparts, X', the encoder output H' is expected to approximate H by optimizing the following objective (IRL-E):

$$\mathcal{L}_{\text{IRL-E}}(X, X') = \beta d_{\text{L}_2}(H, H') - \gamma d_{\cos}(H, H'), \tag{13}$$

where d_{L_2} is the L_2 distance, $d_{\cos}$ is the cosine distance, and β and γ are the tunable parameters that control the weights for d_{L_2} and $d_{\cos}$, respectively.

In addition to the encoder layer, the hidden states of the decoder layers, $\mathbf{s}'_n$, are also penalized based on $\mathbf{s}_n$, using the same criteria as in Eq. (13); the outputs are defined as (IRL-C) as follows:

$$\mathcal{L}_{\text{IRL-C}}(X, X') = \mathcal{L}_{\text{IRL-E}} + \sum_{n=1}^{N} \Big[\beta d_{\text{L}_2}(\mathbf{s}_n, \mathbf{s}'_n) - \gamma d_{\cos}(\mathbf{s}_n, \mathbf{s}'_n)\Big]. \tag{14}$$

The only difference between the IRL and NRAL lies in the choice of where the difference between clean speech and its noisy counterpart is penalized in the model. Rather than coercing the hidden representations map onto the same representation space, we make the attention mechanism produce the same alignment. The proposed attention-based loss is expected to provide the model with more abstract guides, leaving it with sufficient flexibility to train on noisy speech. As the attention weights work on aligning the inputs and outputs, they provide explicit information to the model on which input speech frames are strongly related to the output characters, thereby enabling the model to learn the relations more effectively. We also attempt to combine the NRAL and IRL losses to verify if the model became more noise-robust as a result. To this end, the objective of the combined losses is defined as follows:

$$\mathcal{L}(X, X') = \mathcal{L}_{\text{NRAL}}(X, X'; \lambda) + \mathcal{L}_{\text{IRL}}(X, X'; \beta, \gamma). \tag{15}$$

Note that λ, β, and γ are the hyper-parameters for each loss function.

TABLE I
Experimental data

WSJ [26], [27]	**#speakers**	**#utt.**	**Total dur.**
WSJ1	283	37,416	80 h
WSJ0	83	7,138	15 h
CHiME-4 [28]	**#speakers**	**#utt.**	**Total dur.**
tr05_simu	83	7,138	15 h
dt05_{simu, real}	4	503	5.6 h
et05_{simu, real}	4	333	0.4 h

IV. Experiments

To evaluate the effectiveness of the proposed noise-robust attention learning framework, we conducted speech recognition experiments to compare different end-to-end models. All the models were evaluated based on the character error rates (CERs). The effectiveness of the model itself when there was no external knowledge, such as the ones provided by language models, was evaluated.

A. Datasets

The comparisons were conducted using WSJ1 and WSJ0 [26], [27] as the clean speech corpora, and CHiME-4 [28] as a noisy speech corpus. The details of each corpus are presented in Table I. CHiME-4 was recorded using a tablet device in everyday environments: a cafe, a street junction, public transport, and a pedestrian area. It was composed of two types of data: real data (tr05_real), utterances of actual speakers recorded in real noisy environments, and simulated data (tr05_simu), generated by artificially adding noise to WSJ0. tr05_simu was used as the noisy data in all the experiments. WSJ0 and tr05_simu were used as the data pair in the proposed training framework. The evaluation was performed using the CHiME-4 evaluation sets, including "et05_real_isolated_1ch track" and "et05_simu_isolated_1ch track". The hyperparameters were tuned using the CHiME-4 development sets, including "dt05_real_isolated_1ch track" and "dt05_simu_isolated_1ch track."

For the network inputs, 80-mel-scale filterbank coefficients with a 25-ms analysis window and a 10-ms window shift were extracted from all the raw speech data.

B. Experimental Setup

All the experiments were conducted using the ESPnet [29] CHiME-4 recipe. In the end-to-end ASR model, the encoder was a three-layer biLSTM with 1024 units; its input was downsampled using the VGG layers. The decoder was a one-layer LSTM with 1024 units; the attention mechanism utilized was a location-based mechanism [3]. The network was trained using AdaDelta, and early stopping was applied, based on the

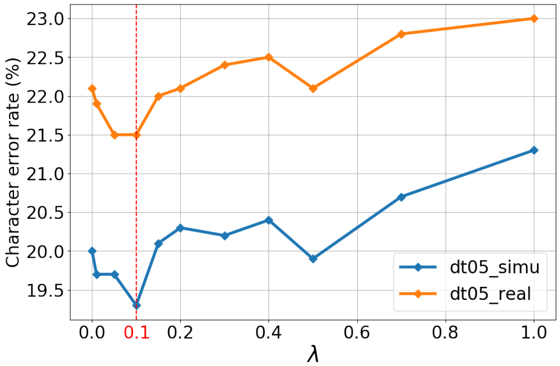


Fig. 2. Character error rates for multi-task $\mathcal{L}_{\text{char}}$ and $\mathcal{L}_{\text{att}}$. The hyperparameter λ was used to control the weight of $\mathcal{L}_{\text{att}}$.

loss on a development set. All the hyperparameters, including λ, β, and γ, were tuned using the development set. To decode the models, the beam search algorithm with a beam size of 30 was used.

C. Models Evaluated

The following models were evaluated:

- **Clean**: A model trained on clean speech;
- **Multistyle**: A model trained on clean and noisy speech;
- **NRAL**: A model trained on noisy speech using the attention loss in Eq. (12); (When the model was trained using $\lambda = 0.0$, it was simply fine-tuned on noisy speech; the proposed attention loss was not used.)
- **IRL-E** [24]: A model trained on noisy speech using the encoder loss in Eq. (13).
- **IRL-C** [24]: In addition to **IRL-E**, a model trained on noisy speech using the decoder loss in Eq. (14).

We also evaluated models trained using a combination of the NRAL and IRL losses in Eq. (15).

D. Results

Table II lists the CER for the CHiME-4 development and evaluation sets. For **Clean**, because noisy data were not included during training, the error rates were obviously worse, compared to those of the other models. By incorporating noisy speech into the training of the architecture and objective function of the same model, the **Multistyle** approach significantly outperformed **Clean**. Figure 2 shows the result for the multi-task learning framework for L_{char} and L_{att}, defined in Eq.(12). For **NRAL** with $\lambda = 0.0$, the pretrained model (**Clean**) was simply adapted to the noisy speech, as a result of which it outperforms **Multistyle**. The proposed attention loss with $\lambda = 0.1$, **NRAL**, outperformed the same model with $\lambda = 0.0$. Furthermore, incorporating **IRL-E** and **IRL-C** into our proposed attention loss significantly improved the performance of the models. **IRL-C+NRAL** outperformed the other models. The values for the parameters, β and γ, were set to 1.0 and 1.0, respectively.

Figure 3 shows the impact of the attention weights between the encoded speech frames and output characters on the evaluation data with two different simulation noises. Note that the **Ideal** weights were extracted from the **Clean** model with the clean speech. Considering the weights in Figure 3(a), **NRAL** appeared to be effective at estimating reasonable alignments; the results for the **Clean** and **Multistyle** on the other hand had partially corrupted alignments. Considering the weights in Figure 3(b), estimating the alignments for the audio was rather challenging by looking at the weights for **Clean** which are completely corrupted. Although the weights for **NRAL** were partially corrupted, compared to **Ideal**, its estimated weights were much clearer than those of **Multistyle**.

V. Conclusions

This study presented a learning framework for improving the noise robustness of the end-to-end speech recognition network. Toward this end, we made the attention weights estimated by the clean speech approximate that of the noisy speech. The experimental comparisons based on CER revealed that the proposed training framework effectively improved the noise robustness of the model. The errors could be further improved by combining the loss that penalizes the hidden representations.

In the future, we plan to investigate the effectiveness of the proposed training framework on a multi-tasking model with connectionist temporal classification.

References

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the International conference on machine learning (ICML)*, 2014.

[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proceedings of the advances in neural information processing systems (NeurIPS)*, 2015.

[4] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[5] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," in *Proceedings of the INTERSPEECH*, 2018.

[6] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the advances in neural information processing systems (NeurIPS)*, 2014.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

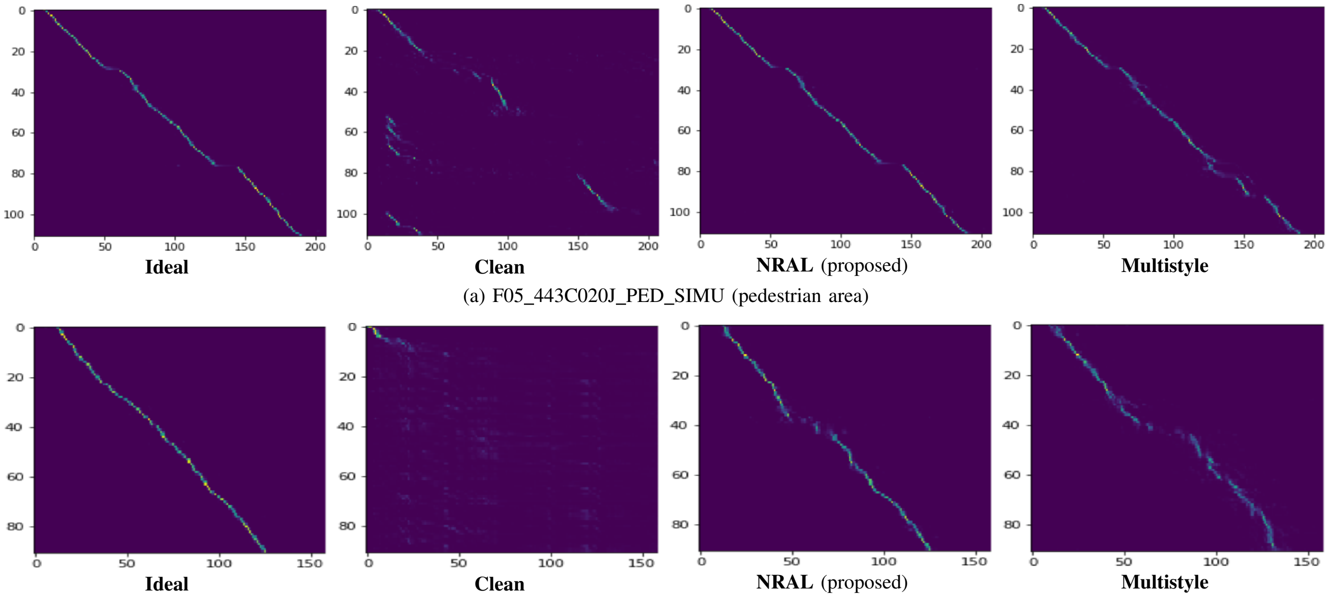
[9] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *Proceedings of the INTERSPEECH*, 2017.

[10] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[11] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

TABLE II
CHARACTER ERROR RATES (%) FOR CHiME-4 DEVELOPMENT AND EVALUATION SETS

	dt05_simu	dt05_real	et05_simu	et05_real
Clean	73.4	72.6	79.1	82.3
Multistyle	21.3	23.2	29.9	35.5
NRAL ($\lambda = 0.0$)	20.0	22.1	29.4	34.3
NRAL ($\lambda = 0.1$)	19.3	21.5	28.8	33.1
IRL-E [24]	19.8	21.8	29.0	33.7
IRL-E [24] + NRAL ($\lambda = 0.1$)	19.0	20.9	28.2	33.0
IRL-C [24]	19.1	21.0	28.5	32.8
IRL-C [24] + NRAL ($\lambda = 0.01$)	**18.5**	**20.2**	**27.8**	**32.5**



(a) F05_443C020J_PED_SIMU (pedestrian area)

(b) F05_442C020Z_STR_SIMU (street junction)

Fig. 3. Attention weights for noisy speech with various simulated noise

[12] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, “Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration,” in *Proceedings of the INTERSPEECH*, 2019.

[13] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, “RWTH ASR systems for librispeech: Hybrid vs attention-w/o data augmentation,” *arXiv preprint arXiv:1905.03072*, 2019.

[14] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[15] M. L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[16] R. Lippmann, E. Martin, and D. Paul, “Multi-style training for robust isolated-word speech recognition,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1987.

[17] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.

[18] S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T. F. Zheng, and Y. Li, “Noisy training for deep neural networks in speech recognition,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 2, 2015.

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[20] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, “Invariant representations for noisy speech recognition,” *arXiv preprint arXiv:1612.01928*, 2016.

[21] Y. Shinohara, “Adversarial multi-task learning of deep neural networks for robust speech recognition.” in *Proceedings of the INTERSPEECH*, 2016.

[22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of International Conference on Machine Learning (ICML)*, 2006.

[24] D. Liang, Z. Huang, and Z. C. Lipton, “Learning noise-invariant representations for robust speech recognition,” in *Proceedings of the IEEE Workshop on Spoken Language Technology Workshop (SLT)*, 2018.

[25] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] L. D. Consortium *et al.*, “CSR-II (WSJ1) complete,” *Linguistic Data Consortium, Philadelphia, vol. LDC94S13A*, 1994.

[27] J. Garofalo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) complete,” *Linguistic Data Consortium, Philadelphia, vol. LDC93S6A*, 2007.

[28] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[29] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: End-to-end speech processing toolkit,” in *Proceedings of the INTERSPEECH*, 2018.