INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Noise robust acoustic modeling for single-channel speech recognition based on a stream-wise transformer architecture

Masakiyo Fujimoto and Hisashi Kawai

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

Abstract

This paper addresses a noise-robust automatic speech recognition (ASR) method under the constraints of real-time, one-pass, and single-channel processing. Under such strong constraints, single-channel speech enhancement becomes a key technology because methods with multiple-passes or batch processing, such as acoustic model adaptation, are not suitable for use. However, single-channel speech enhancement often degrades ASR performance due to speech distortion. To overcome this problem, we propose a noise robust acoustic modeling method based on the stream-wise transformer model. The proposed method accepts multi-stream features obtained by multiple single-channel speech enhancement methods as input and selectively uses an appropriate feature stream according to the noise environment by paying attention to the noteworthy stream on the basis of multi-head attention. The proposed method considers the attention for the stream direction instead of the time series direction, and it is thus capable of real-time and low-latency processing. Comparative evaluations reveal that the proposed method successfully improves the accuracy of ASR in noisy environments and reduces the number of model parameters even under strong constraints.

Index Terms: noise robust ASR, one-pass single-channel processing, speech enhancement, stream-wise transformer

1. Introduction

Ensuring the noise robustness of automatic speech recognition (ASR) in daily environments is a crucial problem for high-quality ASR applications. We have become keenly aware of this problem through our research and development of a speech-to-speech multilingual translation application for mobile devices and an automatic subtitling system for broadcasting programs.

To overcome this crucial problem, various noise robust ASR methods have been studied for many years. The simplest way to ensure noise robustness is speech enhancement in the front-end processing of ASR. As representative legacy methods, 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] have been widely used. Recent deep-learning-based approaches, such as the use of a de-noising auto-encoder (DAE) [4] and binary masking [5], have demonstrated higher performance than legacy methods. In addition, deep-learning-based approaches with direct input–output of the raw waveform have been proposed and shown excellent performance [6, 7, 8]. Speech enhancement is the simplest way to achieve noise robust ASR; however, the distortion caused by speech enhancement often degrades the performance of ASR. This performance degradation is notable in recent deep neural network-based ASR frameworks, especially in single-channel processing. In contrast, multi-channel processing provides a notable improvement in noisy speech ASR, and various methods have been proposed; e.g., methods based on microphone-array signal processing [9, 10, 11, 12] and methods based on deep-learning [13, 14, 15]. Although multi-channel processing has remarkable performance, it requires special hardware equipments, such as a microphone-array and a multi-channel microphone amplifier. These hardware requirements pose a difficult problem for implementing speech applications in a mobile environment and for deploying applications in a typical office environment.

As alternatives to front-end processing, advanced acoustic models (AMs) with complicated network architectures, e.g., the convolutional neural network (CNN) [16, 17], recurrent neural network (RNN) with long short-term memory (LSTM) [18], and convolutional LSTM (CLSTM) [19, 20] have attracted attention as alternatives to simple fully-connected feed-forward networks. However, these advanced models may not be fully effective if they are simply applied to noise-robust ASR. A single-channel speech input and one-pass framework in real-time (low-latency) processing are mandatory for our research and development targets in speech applications. Therefore, it is necessary to avoid methods that require multi-pass and iterative batch processing, such as off-line AM adaptation and rescoring with RNN language models.

To address the aforementioned problems, we have proposed a multi-stream input model [21] that takes multiple enhanced speech features as input. This model adopts the gating mechanism used in LSTM to focus on appropriate enhanced speech features. However, this model does not consider the relationship between feature streams. As a method of focusing on a specific feature, the use of attention models has been proposed in the research fields of machine translation and natural language processing. Representative attention models are an LSTM-based seq-to-seq model [22] and a transformer model [23]. These attention models have also been used for AMs in ASR [24, 25, 26]. In addition, a CNN-based transformer model, known as the conformer model [27] and multi-stream models [28, 29], has been proposed for a state-of-the-art end-to-end ASR. Inspired by the recent excellent success of the attention model, this paper proposes a stream-wise transformer model that pays attention to the stream direction rather than to the time series direction. Unlike the conventional transformer model, the proposed method requires no long-term observation of the feature sequence and can achieve real-time and low-latency processing because it pays attention to an appropriate speech feature stream among multiple enhanced speech streams in a certain time frame.

We evaluate the proposed method on the CHiME4 1-channel track [30] and the corpus of spontaneous Japanese (CSJ) [31, 32]. Results reveal that the proposed method improves the accuracy of ASR in noisy environments and reduces the number of model parameters, even under the constraints of real-time, one-pass, and single-channel processing.

http://dx.doi.org/10.21437/Interspeech.2021-225

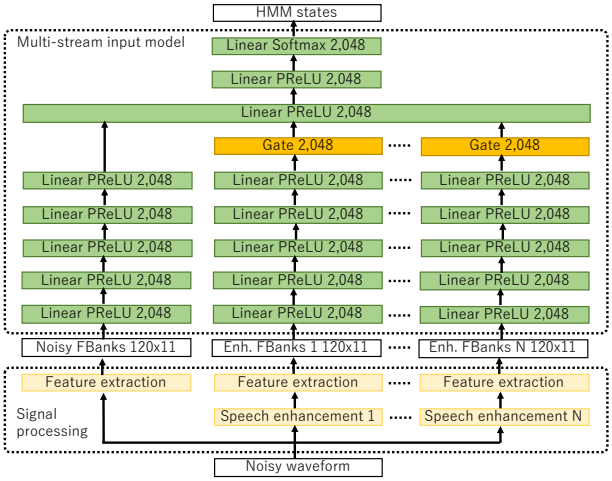

Figure 1: *Network structures for a multi-stream input model*

2. Multi-stream input acoustic modeling

This section briefly reviews our previously proposed multi-stream input model [21]. The structure of the multi-stream input model is shown in Fig. 1. In the figure, the signal processing unit extracts noisy speech features and N types of enhanced speech features from the input single-channel noisy waveform and uses them as input to the model. Therefore, the multi-stream input model allows not only noisy features but also additional N types of enhanced feature to be input. Here, each feature parameter consists of 40 log mel-filter bank features (FBanks) and their first and second derivatives, which are extracted using a Hamming window with a 25-ms frame length and 10-ms frame shift. A context window with 11 (± 5) frames is also applied to each feature. Since each speech enhancement method has its own limitations, this model accepts multiple inputs of enhanced FBanks in order to achieve robustness against various noise environments. In addition, since the input of enhanced FBanks alone may degrade the ASR accuracy due to the influence of speech distortion caused by speech enhancement, noisy FBanks are also input to reduce influence of speech distortion.

As shown in Fig. 1, all features are propagated to individual sub-networks, and the outputs of each sub-network are merged into a single data flow. Using these sub-networks, advanced feature extraction can be performed in contrast with the case that multiple features are simply input to the model. In addition, we introduce a gating mechanism used in LSTM to selectively use noteworthy input features. The multi-stream input model does not require multiple-passes, multiple-channels, batch processing, or high-latency processing.

3. Stream-wise transformer modeling

The aforementioned multi-stream input model uses gate layers to independently select noteworthy input features for each stream; however, the relative relationship between streams is not considered. We therefore introduce an attention mechanism that considers the relationship between streams and clarifies which streams should receive attention. Various attention-based model have been proposed. In this paper, we adopt the transformer model [23], which has been used in the research fields of machine translation and natural language processing,

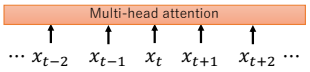

Figure 2: *Multi-head attention unit*

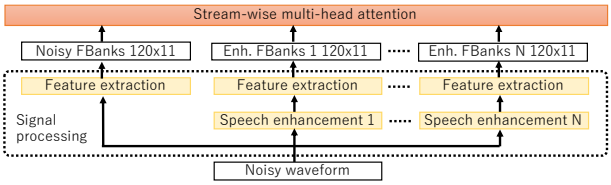

Figure 3: *Stream-wise multi-head attention unit*

as the AM.

The conventional transformer model has a multi-head attention unit. As shown in Fig. 2, the input to the conventional multi-head attention unit is a time series of feature $\boldsymbol{x}_t$, and the multi-head attention unit investigates which time of the feature sequence to pay attention to. In contrast, the proposed method uses the multi-stream feature, which is the output of the signal processing unit, as the input to the multi-head attention unit as shown in Fig. 3. The proposed method thus considers the attention for the stream direction instead of the time series direction. We call this method the stream-wise multi-head attention method and call the transformer model with the stream-wise multi-head attention unit the stream-wise transformer model.

The structure of the stream-wise transformer model is shown in Fig. 4. The input to the encoder of the stream-wise transformer model is multi-stream features in the t-th frame, which is the output of the signal processing unit. Meanwhile, the input to the decoder is only a noisy feature in the t-th frame. Here, since the input to the decoder is only a noisy feature, the self-attention in the decoder may be meaningless. Therefore, in the proposed method, we also consider the structure without self-attention in the decoder as shown in Fig. 5.

The features input to the conventional transformer model may have different time lengths and it is necessary to prohibit references to future information. Therefore, feature masking is required before inputting the features to the multi-head attention unit. In contrast, the proposed method requires no feature masking because all the information is observable at the time of input to the stream-wise multi-head attention unit.

Moreover, for additional advanced feature extraction and dimension reduction, we investigate the insertion of a CNN-based sub-sampling unit as shown in Fig. 6 between the signal processing unit and stream-wise transformer model.

4. Experiments with the CHiME4 corpus

The proposed method was evaluated on the CHiME4 1-channel track [30].

4.1. Experimental setup

The CHiME4 corpus was recorded using a tablet device equipped with six microphones in four types of noise environment: a public transportation platform, cafeteria, pedestrian area, and street intersection. The training set consists of 1,600 real and 7,138 simulated (Simu) utterances spoken by four and 83 speakers, respectively. The amount of training data is about 18.0 hours and the vocabulary size is 5k words. The development and evaluation sets consist of 3,280 and 2,640 utterances,

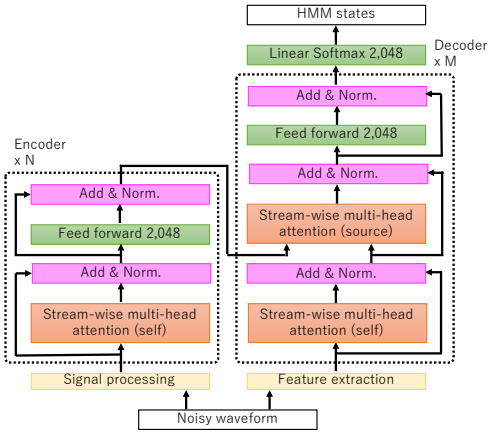

Figure 4: *Stream-wise transformer model*

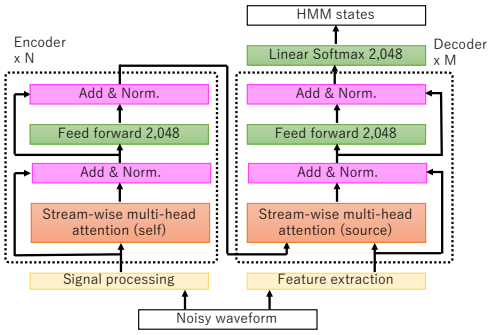

Figure 5: *Stream-wise transformer model without self-attention in the decoder layer*

respectively, each containing equal quantities of real and simulated data. Both the real and simulated sets were spoken by four speakers, respectively. The development set was used for cross validation during AM training and parameter tuning. In the CHiME4 1-channel track, the speech data recorded by the fifth microphone was used for an evaluation.

The speech enhancement methods used in the evaluation were SS, MMSE-STSA, GMM-FE, DAE, and Conv-TasNet [7]. All these enhancement methods are capable of frame-by-frame real-time processing. The GMM for GMM-FE and DAE were trained using clean simulation data and real data recorded using a headset microphone. The Gaussian mixture model consisted of 512 Gaussian distributions. The DAE was trained using unidirectional CLSTM [20] which consists of 32 filters with 3×3 shape and a hyperbolic tangent activation function. The structure of Conv-TasNet follows that in the literature [7].

All AMs were trained using PyTorch [33], and ASR decoding with trained AMs was conducted using the Kaldi [34]. The AM of the baseline system consisted of a fully connected feed-forward network with seven hidden layers. Each hidden layer has 2048 units, and the activation function was parametric rectified linear unit (PReLU) [35]. The structure of the multi-stream input model is shown in Fig. 1, and five different speech enhancement methods were used. Therefore, the number of input feature streams was six, including noisy features. The features input to the encoder of the proposed stream-wise transformer were the same as those of the multi-stream input model, and features input to the decoder were only noisy features . The number of heads for multi-head attention and the numbers of layers for the encoder and decoder were adjusted

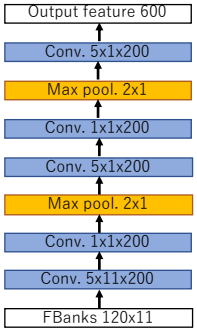

Figure 6: *CNN-based sub-sampling unit*

using the developement data. The target labels, which consisted of 1,967 context dependent-hidden Markov model (CD-HMM) states, were obtained using the Kaldi CHiME4 recipe [36]. In the training phase, the parameters of each AM were randomly initialized and optimized using momentum stochastic gradient descent with a cross entropy criterion. A mini-batch of 256 frames and an initial learning rate of 0.02 were used for optimization.

The training of the tri-gram language model also followed the Kaldi CHiME4 recipe. The ASR experiments were performed using fully composed weighted finite state transducers with the AMs. The evaluation criterion was the word error rate (WER).

4.2. Experimental results

Table 1 indicates the WERs obtained for each speech enhancement method[1]. In the table, Baseline refers to the results obtained without speech enhancement. The table shows that the results of each speech enhancement were either worse or only slightly better than those of the baseline. With Conv-TasNet, an absolute WER improvement of approximately 2.7% was obtained for real evaluation data; however, this is not a sufficient performance improvement.

Meanwhile, Tables 2 and 3 indicate the model structure, the number of model parameters, and WERs obtained with the multi-stream input model (MSI) and proposed methods. In the proposed method, the following four types of model structures were evaluated.

STF: Stream-wise transformer model
STF-NoDS: STF without decoder self-attention
STF-C: STF with the CNN-based sub-sampling unit
STF-NoDS-C: STF-C without decoder self-attention

The table shows that the proposed stream-wise transformer models outperformed the previously proposed multi-stream input model. It also confirms that the decoder self-attention does not notably affect the WER performance. Therefore, the proposed method without decoder self-attention has the advantage that almost the same results are obtained even if the number of model parameters is reduced. In addition, further WER improvement and model parameter size reduction are achieved using the CNN-based sub-sampling units. Finally, the proposed method, STF-NoDS-C, achieved a 6% absolute WER reduction with real evaluation data and 50% reduction in the number of model parameters compared with the baseline results, demonstrating the effectiveness of the proposed method.

[1]To reduce influence of speech distortion, the original noisy speech was added to the enhanced speech before training and evaluation.

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

Enhancement method	Development set		Evaluation set	
	Simu	Real	Simu	Real
Baseline	13.85	15.13	15.89	23.31
SS	14.51	15.23	16.12	23.66
MMSE-STSA	13.95	15.04	15.94	23.06
GMM-FE	13.45	14.62	15.49	22.99
DAE	13.42	14.35	15.36	22.52
Conv-TasNet	**11.94**	**12.76**	**13.82**	**21.03**

Table 2: *Structure and the number of parameters for each model of the CHiME4 1-channel track*

Model architecture	#Heads	#Encoder layers	#Decoder layers	#Params.
Baseline	—	—	—	30.5M
MSI	—	—	—	167.6M
STF	8	3	3	93.4M
STF-NoDS	4	1	4	61.6M
STF-C	8	1	2	18.0M
STF-NoDS-C	8	1	2	**15.3M**

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

Model architecture	Development set		Evaluation set	
	Simu	Real	Simu	Real
Baseline	13.85	15.13	15.89	23.31
MSI	10.90	11.20	12.87	18.54
STF	**10.47**	10.51	**12.23**	18.14
STF-NoDS	10.53	**10.38**	12.24	**17.87**
STF-C	**9.97**	**10.12**	**11.64**	**17.09**
STF-NoDS-C	10.31	10.32	11.72	17.17

5. Evaluation on the CSJ corpus

We also evaluated the proposed method on the CSJ [31, 32], a large-scale corpus of the Japanese language, to demonstrate the 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. During training, ten lectures (2.0 hours) were selected as the development (Dev) set. 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 under clean conditions using headset microphones, and we thus added four noise data (for an airport lobby, exhibition hall, shopping mall, and train station) artificially to each data set with randomly selected signal-to-noise (SNR) ranges of 0 to 10 dB. The noise data were taken from the ATR ambient noise sound database [37]. We designed closed-domain sets and open-domain sets for evaluation data. The details of the noise conditions for each dataset are given 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 [38]. 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 14.04% for E01, 11.12% for E02, and 15.1% for E03.

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

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

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

Enhancement method	Closed domain			Open domain		
	E01	E02	E03	E01	E02	E03
Baseline	15.78	13.51	16.74	20.59	21.20	21.11
SS	16.26	13.74	17.00	21.08	21.36	21.76
MMSE-STSA	15.89	13.67	16.99	20.86	21.67	21.35
GMM-FE	15.71	13.65	16.70	20.72	21.27	20.89
DAE	**15.51**	13.60	16.63	20.69	21.27	**20.82**
Conv-TasNet	15.61	**13.32**	**16.59**	**20.44**	**21.23**	21.01

Table 6: *Structure and the number of parameters for each model on the CSJ*

Model architecture	#Heads	#Encoder layers	#Decoder layers	#Params.
Baseline	—	—	—	45.2M
MSI	—	—	—	182.3M
STF	4	3	4	121.3M
STF-NoDS	4	3	4	94.7M
STF-C	8	2	2	26.1M
STF-NoDS-C	4	1	2	**17.0M**

Table 7: *ASR results for the proposed method on the CSJ in terms of the WER (%).*

Model architecture	Closed domain			Open domain		
	E01	E02	E03	E01	E02	E03
Baseline	15.78	13.51	16.74	20.59	21.20	21.11
MSI	13.81	11.56	14.77	18.47	18.24	18.60
STF	13.38	11.19	**14.03**	18.33	**18.51**	**17.74**
STF-NoDS	**13.34**	**11.15**	**14.03**	**18.14**	18.59	18.04
STF-C	**12.87**	**10.74**	**13.71**	**17.80**	**17.23**	**17.22**
STF-NoDS-C	12.96	10.86	14.08	17.88	17.50	17.45

5.2. Experimental results

Table 5 indicates the WERs obtained for each speech enhancement method. Similar to the case for the CHiME4 results, it is seen that speech enhancement alone is not effective. Tables 6 and 7 show that the proposed method both improved the WER significantly and reduced the number of model parameters even for the CSJ corpus. These results demonstrate that the proposed method is effective regardless of the ASR task.

6. Conclusions

This paper described a noise robust ASR method that effectively works under the constraints of real-time, one-pass, and single-channel processing. The proposed stream-wise transformer model, which accepts multi-stream features obtained by multiple single-channel speech enhancement methods as input and selectively uses an appropriate feature stream according to the noise environment by paying attention to the noteworthy stream, improved the ASR accuracy in noisy environments and reduced the number of model parameters. In the near future, we will consider incorporating various state-of-the-art model structures, such as a conformer. Additionally, we will attempt to build the signal processing unit as a trainable module and to construct an end-to-end model that can be jointly optimized.

7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 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 Transactions on Acoustics, Speech, and Signal Processing*, 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] S. W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. of APSIPA '17*, December 2017, pp. 6–12.

[7] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time – frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, May 2019.

[8] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *Proc. of ICASSP '20*, May 2020, pp. 7004–7008.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. of ICASSP '16*, March 2016, pp. 5745–5749.

[14] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, vol. 46, pp. 374–385, November 2017.

[15] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in *Proc. of ICASSP '20*, May 2020, pp. 6384–6388.

[16] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, October 2014.

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

[18] 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.

[19] 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.

[20] 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.

[21] ——, "One-pass single-channel noisy speech recognition using a combination of noisy and enhanced features," in *Proc. of Interspeech '19*, September 2019, pp. 486–490.

[22] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. of EMNLP '15*, September 2015, pp. 1412–1421.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NeurIPS '17*, December 2017.

[24] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 11, no. 8, pp. 1240–1253, September 2017.

[25] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. of ICASSP '18*, April 2018, pp. 5884–5888.

[26] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang, C. Fuegen, G. Zweig, and M. L. Seltzer, "Transformer-based acoustic modeling for hybrid speech recognition," in *Proc. of ICASSP '20*, May 2020, pp. 6874–6878.

[27] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. of Interspeech '20*, October 2020, pp. 5036–5040.

[28] R. Li, X. Wang, S. H. Mallidi, S. Watanabe, T. Hori, and H. Hermansky, "Multi-stream end-to-end speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 646–655, December 2019.

[29] J. Pan, J. Shapiro, J. Wohlwend, K. J. Han, T. Lei, and T. Ma, "ASAPP-ASR: Multistream CNN and self-attentive SRU for SOTA speech recognition," in *Proc. of Interspeech '20*, October 2020, pp. 16–20.

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

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

[32] 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.

[33] "PyTorch," https://pytorch.org/.

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

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. of ICCV '15*, December 2015, pp. 1026–1034.

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

[37] 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).

[38] 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.