Interspeech 2018
2-6 September 2018, Hyderabad

Monaural Multi-Talker Speech Recognition with Attention Mechanism and Gated Convolutional Networks

Xuankai Chang[1], *Yanmin Qian*[1], *Dong Yu*[2]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Tencent AI Lab, Tencent, Bellevue, WA, USA

xuank@sjtu.edu.cn, yanminqian@sjtu.edu.cn, dyu@tencent.com

Abstract

To improve the speech recognition accuracy under the multi-talker scenario, we propose a novel model architecture that incorporates the attention mechanism and gated convolutional network (GCN) into our previously developed permutation invariant training based multi-talker speech recognition system (PIT-ASR). The new architecture has three components: an encoding transformer, an attention module and a frame-level senone predictor. The encoding transformer first transforms a mixed speech sequence into a sequence of embedding vectors. Then the attention mechanism extracts individual context vectors from this embedding sequence for different speaker sources. Finally the predictor generates the senone posteriors for all speaker sources independently with the knowledge from the context vectors. To get better embedding representations we explore gated convolutional networks in the encoding transformer. The experimental results on the artificially mixed two-talker WSJ0 corpus show that our proposed model can reduce the word error rate (WER) by more than 15% relatively compared to our previous PIT-ASR system.

Index Terms: permutation invariant training, attention model, gated convolutional network, multi-talker speech recognition

1. Introduction

The progress made in deep learning technology has led to significant improvements in single-talker speech recognition task in the past a few years [1, 2, 3, 4, 5, 6, 7, 8]. The state of the art system even achieved a human comparable performance on close-talk tasks [9] and some simple noisy tasks [10, 11]. Despite all these advancements, current methods still have a huge degradation when facing more complex noisy scenarios in reality, especially the far-field condition interfered with background noise, reverberation and speech from other talkers.

In this paper, we aim to attack the monaural multi-talker speech recognition problem, which aims to recognize the individual speech source from the overlapped speech mixed in one single channel. Recently, there have been several deep learning-based works focusing on the monaural multi-talker speech separation and recognition. In [12], a deep neural network was designed to directly recognize the phonemes and the label assignment during training was based on the speech energies. In [13, 14], a deep framework called deep clustering (DPCL) separates the speech by segregating the embedding vectors of all time-frequency bins on the spectrum. Another model called deep attractor network (DANet) [15] learns a high-dimensional embedding of the speech spectrum and clusters embeddings with attractor points. Permutation invariant training (PIT) [16, 17, 18, 19, 20, 21, 22, 23, 24] addresses this problem using a simple, yet effective training criterion by minimizing the average minimum error with the best output-target assignment.

Recently the attention mechanism becomes popular in the sequence to sequence framework used for, e.g., machine translation [25, 26] and end-to-end speech recognition [27, 28]. Taking the attention based end-to-end speech recognition as an example [27, 28], the input is the frame-level speech feature sequence, and the output is the character- or word-level sequence. An encoder-decoder framework is usually applied to learning the relationship between these two sequences. The basic encoder-decoder model cannot achieve a good performance because it is hard to accurately align the input and output sequences. The attention mechanism, however, enables the model to learn a better alignment between the input and output sequences with different lengths, extracts the related knowledge by focusing on the properly aligned input sequence, and thus significantly improves the system.

In this work, we explore a new model that integrates the attention mechanism into our previously developed PIT-ASR model for multi-talker speech recognition. The proposed architecture consists of three components: an encoding transformer, an attention module and a frame-level senone predictor. Different from the prior arts on the attention mechanism, which aims at obtaining a better alignment between the input and output sequences, this work exploits the attention mechanism to better trace speakers and eliminate interferences by attending on the embedding sequence segments. Motivated by the recent success on language modeling [29] and audio classification [30], we also utilize the gated convolutional network (GCN) in the encoding transformer to generate better embedding sequence. Compared with the conventional CNN, GCN improves the system by automatically controlling the information flow to the next layer via the learned gate.

The rest of the paper is organized as follows. In Section 2 we review the multi-talker speech recognition task and the permutation invariant training (PIT). In Section 3, we describe the proposed model that incorporates the attention mechanism and the gated convolutional networks. Experimental results are presented in Section 4. We conclude the paper in Section 5.

2. Permutation Invariant Training for Monaural Multi-Talker Speech Recognition

In monaural multi-talker speech recognition, it is given a linearly mixed single microphone signal $\mathbf{y}[n] = \sum_{s=1}^{S} \mathbf{x}_s[n]$, where $\mathbf{x}_s[n], s = 1, ..., S$ are S streams of speech sources from

Yanmin Qian and Dong Yu are the corresponding authors.

This work is supported by the China NSFC projects (No. U1736202 and No. 61603252), the Shanghai Sailing Program No. 16YF1405300, and the Tencent-Shanghai Jiao Tong University joint project.

10.21437/Interspeech.2018-1547

different speakers. The goal is to separate and recognize these streams. "label permutation" is a critical problem in the multi-talker speech separation and recognition tasks, and permutation invariant training (PIT) [16, 17, 18, 19, 20, 21, 22, 24] is an efficient and effective technique to address this problem. In the previous PIT-ASR [18] model, a deep bidirectional LSTM network takes the spectrum features $\mathbf{Y}$ of the mixed signal $\mathbf{y}$ as inputs, and outputs S individual senone posterior streams $\mathbf{O}^s, s = 1, \cdots, S$. The model is optimized with PIT to minimize the objective function

$$J = \frac{1}{S} \min_{s' \in permu(S)} \sum_s \sum_t CE(\ell_t^{s'}, \mathbf{O}_t^s), s = 1, \cdots, S \quad (1)$$

where $permu(S)$ is a permutation of $[1, \cdots, S]$. Note that PIT automatically finds the appropriate assignment no matter how the labels are ordered, and solves the label permutation problem and speaker tracing problem by computing the cross entropy (CE) over the whole sequence for each assignment. Although DPCL [13, 14] and DANet [15] techniques can also address the label permutation problem, PIT-ASR is much simpler and more compact since it allows direct multi-talker mixed speech recognition without explicit separation.

3. Attention-based Model for Multi-talker Speech Recognition with PIT

The newly proposed entire framework with attention mechanism is shown in Figure. 2. The architecture can be divided into three parts: 1) an encoding transformer, through which the mixed speech feature sequence is encoded into the embedding sequence $\mathbf{h}$; 2) an attention module, which learns an attention vector α to generate speaker i specific context vectors $\mathbf{c}_t^i$ at each frame t based on the embedding sequence; 3) a predictor, which takes the learned context vectors $\mathbf{c}_t^i$ as input and generates the senone posteriors for each speech source i as the output. The permutation invariant training is implemented on the outputs of the predictors. In order to obtain better mixed speech embedding representations through the encoding transformer, we replaced some recurrent layers (BLSTM-RNN) with convolutional layers (CNN) and the gated convolutional layers (GCN) as a further exploration.

3.1. Encoding Transformer and Gated Convolutional Neural Networks

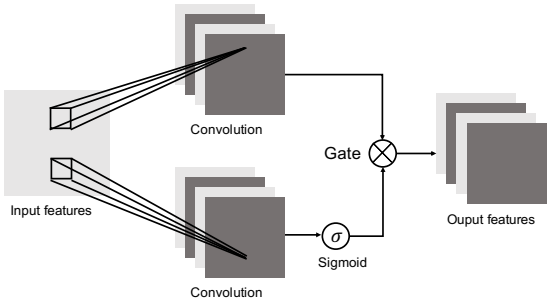

Figure 1: *Gated convolutional neural networks.*

The encoding transformer is used to process the input spectrum feature sequence into the embedding sequence. The input spectrum feature $\mathbf{y}_t$ is fed into a deep model and turned into the corresponding encoded embedding vector $\mathbf{h}_t$ at each time step t. Usually several stacked BLSTM-RNN layers can be used as the encoding network, in which the encoded embedding vector $\mathbf{h}_t$ is computed as

$$\mathbf{h}_t^f = LSTM^f(\mathbf{y}_t), t = 1, \ldots, T \quad (2)$$

$$\mathbf{h}_t^b = LSTM^b(\mathbf{y}_t), t = 1, \ldots, T \quad (3)$$

$$\mathbf{h}_t = Stack\left(\mathbf{h}_t^f, \mathbf{h}_t^b\right) \quad (4)$$

where $LSTM^f$ and $LSTM^b$ are the forward and backward LSTMs respectively.

Our previous work [20] indicated that using convolutional layers in the original PIT-ASR model can improve the performance of the system on the overlapped speech. For this reason we explored to replace some BLSTM-RNN layers of the encoding transformer with convolutional layers and gated convolutional networks (GCN) as shown in Figure.1. Given the input $\mathbf{X}$, the gated convolutional layer is defined as

$$h(\mathbf{X}) = (\mathbf{X} * \mathbf{W} + \mathbf{b}) \otimes \sigma(\mathbf{X} * \mathbf{V} + \mathbf{d}) \quad (5)$$

where $*$ is the convolution operator, σ is the sigmoid function, $\mathbf{W}, \mathbf{V}$ are the weight parameters, and $\mathbf{b}, \mathbf{d}$ are the biases.

GCN replaces the direct non-linear activation function in the conventional CNN with the gated linear units, and has shown promising ability on many machine learning tasks recently [29, 30]. With the gating mechanism, GCNs relieve the gradient vanishing problem in training deep models, as they provide a linear path for the gradients while retaining non-linear capabilities with the sigmoid operation. In our multi-talker mixed speech recognition task, GCNs in the encoding transformer also generate better embedding representation.

3.2. Attention Mechanism for Multi-talker Speech Recognition

Different from the previous attention works, which aim at obtaining a better alignment between the input and output sequences with different sequence lengths. In our model, however, alignment is unnecessary since the output, the frame-level senone posterior sequence, has the same length as the input sequence. Instead, the attention model here is used to more accurately trace the speakers and eliminate interferences by attending to the specific speech embedding sequence segment. The attention mechanism is shown in Figure.3. During the model training, all the parameters in the attention module are jointly optimized with those in other modules.

Local attention is used in this work which only selectively attends to a context window of input embedding sequence. To be more specific, at every time t, the context vector $\mathbf{c}_t^i$ is derived as a weighted average over a subsequence of the encoded embedding vectors within the window $[t - N, t + N]$. N is a hyper-parameter that represents the context window size. The attention weights for speaker i at time t is $\alpha_t^i \in \mathbb{R}^{2N+1}$.

At time step t, the senone posterior probability for speaker i is defined as

$$p(\mathbf{o}_t^i | \mathbf{o}_1^i, ..., \mathbf{o}_{t-1}^i, \mathbf{y}_t) = g(\mathbf{s}_t^i, \mathbf{c}_t^i) \quad (6)$$

where $\mathbf{c}_t^i$ and $\mathbf{s}_t^i$ are the generated context vectors from the attention model and the hidden state in the predictor for speaker i, and g is normally realized with an MLP. $\mathbf{s}_t^i$ is computed as

$$\mathbf{s}_t^i = LSTM(\mathbf{s}_{t-1}^i, \mathbf{c}_t^i) \quad (7)$$

where $LSTM$ is the uni-directional forward LSTM layer. And the attention vector α_t^i is obtained by comparing the previous

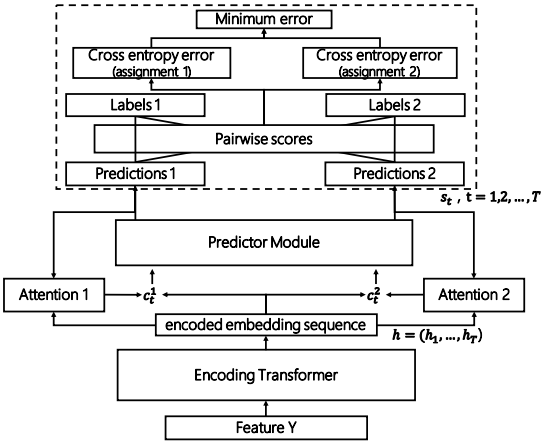

Figure 2: *PIT framework with attention mechanism for multi-talker speech recognition.*

hidden state $\mathbf{s}_{t-1}^i$ with each encoded embedding vector $\mathbf{h}_t$:

$$\begin{aligned}\alpha_t^i(k) &= Attend(\mathbf{s}_{t-1}^i, \mathbf{h}_k) \\ &= \frac{exp(score(\mathbf{s}_{t-1}^i, \mathbf{h}_k))}{\sum_j exp(score(\mathbf{s}_{t-1}^i, \mathbf{h}_j))}\end{aligned} \tag{8}$$

where $score$ is a content-based function

$$score(\mathbf{s}_{t-1}^i, \mathbf{h}_k) = \begin{cases} {\mathbf{s}_{t-1}^i}^{\mathrm{T}} \mathbf{W}_a \mathbf{h}_k & general \\ \mathbf{v}_a^{\mathrm{T}} \tanh\left(\mathbf{W}_a[\mathbf{s}_{t-1}^i; \mathbf{h}_k]\right) & concat \end{cases} \tag{9}$$

where $\mathbf{v}_a$ and $\mathbf{W}_a$ are the learned parameters in the attention model. The context vector $\mathbf{c}_t^i$ is finally computed as a weighted sum over the embedding vector $\mathbf{h}_t$ with a segment size $2N+1$:

$$\mathbf{c}_t^i = \alpha_t^i(1)\mathbf{h}_{t-N} + \cdots + \alpha_t^i(2N+1)\mathbf{h}_{t+N} \tag{10}$$

We believe that the attention module can help continuously attend to a specific speaker on every output by extracting context knowledge using the context vector. As shown in Figure 2, after encoding the input features, two separate attention streams are constructed individually to get the context vector for each speaker respectively. Then the individual context vectors $\mathbf{c}_t^i$ of each speaker are fed into one RNN-based predictor to generate the frame-level senone posterior stream. The entire architecture is optimized with permutation invariant training as the normal PIT-ASR model.

4. Experiments

4.1. Experimental setup and baseline system

To evaluate our proposed model, we conducted all the experiments on the artificially mixed two-talker Wall Street Journal (WSJ0) speech corpus released by MERL [14]. There are about 30hr-speech in the training set and 10hr-speech in the validation set, which are generated by randomly selecting a pair of utterances from different speakers in the WSJ0 training set si_tr_s, and mixing them at various signal-to-noise (SNR) randomly chosen between 0dB and 10dB. Similarly, the evaluation set contains the 5hr-speech generated by mixing utterances of 16 unseen speakers from si_dt_05 and si_et_05. Note that we only test on the evaluation set, since the open-set condition is more interesting and important for the research in this area.

40-dimensional log filter bank features are used as the input for all the models, and the labels for the model training are aligned with a DNN model built following the WSJ recipe in Kaldi [31]. There are totally 3429 tied-states in the output layer.

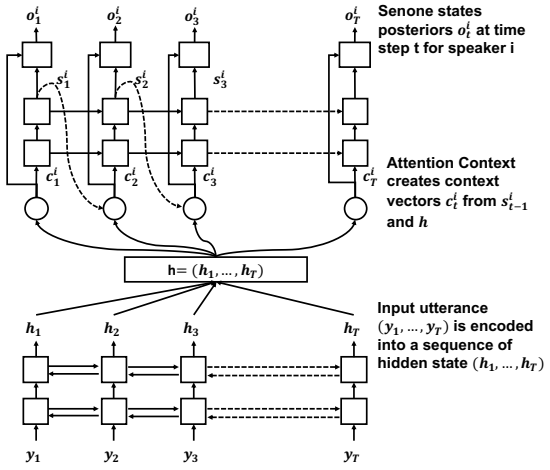

Figure 3: *Attention mechanism.*

All neural network acoustic models proposed in this work are built with Pytorch [32] using SGD on 1 GPU. The learning rate is set to $2e-4$ and the gradients are clipped with L2-norm 500. A standard pipeline in Kaldi WSJ recipe is used for decoding. For scoring, we evaluate the hypotheses pairwisely against the two references, and make the assignment with better WER as the final recognition results for each utterance.

Firstly, a baseline normal PIT-ASR system is constructed as in our previous work [18]. It has a 6-layer BLSTM-RNN model with 384 memory cells in each layer [1]. The output of the last BLSTM layer is sent directly to two output layers with softmax activation, representing two recognition streams. These two outputs are then used in decoding to obtain the hypotheses for two talkers. The averaged WER on two talkers of the baseline is shown as the first line in Table 1. The WER is much higher than that of the single-talker task on this corpus, which demonstrates the big challenge for multi-talker ASR.

4.2. Evaluation on the Gated CNN

We replace some BLSTM layers at the bottom of the baseline PIT-ASR model with the convolutional layers. The usual CNN layer is utilized and the depth of CNN layers is increased from 2 to 6 (from a shallow CNN to a deep CNN). The results are shown in the middle part of Table 1. It is observed that using CNN layers in PIT-ASR can get obvious improvements. Increasing convolutional layer depth brings further gains.

The GCN with 2 Gated convolutional layers plus 4 BLSTM layers is explored. The result, shown in the bottom line of Table 1, indicates that GCN-BLSTM achieved relative 12% and 6% improvements upon the 6L-BLSTM and 2L-CNN-4L-BLSTM respectively. Even the shallow GCN-BLSTM slightly outperforms the deep CNN-BLSTM.

Table 1: *WERs (%) of normal PIT-ASR models with different model structures*

Model	# L CNN/GCN	#L BLSTM	AVG WER
BLSTM	—	6 BLSTM	37.2
CNN-BLSTM	2 CNN	4 BLSTM	34.8
	4 CNN	4 BLSTM	33.5
	6 CNN	4 BLSTM	32.9
GCN-BLSTM	2 GCN	4 BLSTM	**32.7**

[1]We used less cells here compared to our previous works [18, 20] for the fast development, and larger model scale can get a better system performance based on our experiments.

4.3. Evaluation on the Attention Mechanism

The proposed method with attention mechanism is then evaluated. The new model has an encoding transformer that is a 3-layer BLSTM and a predictor module that is a 3-layer LSTM with 384 cells per layer. The context size used in the attention model is set to 10. Two scoring methods in the attention model, i.e. "general" and "concat" shown in Equation 9, are compared, and the results are listed in Table 2. It is observed that the newly proposed architecture with attention mechanism is significantly better than the normal PIT-ASR for multi-talker speech recognition. For the different scoring modes in the attention module, the "concat" obviously outperforms the "general". The "concat" scoring method is utilized in all the following experiments.

Table 2: *WERs (%) of the attention-based PIT-ASR models with different score methods in the attention module*

Model	Score Method	AVG WER
PIT-ASR	—	37.2
+ Attention	general	34.6
	concat	**33.1**

Table 3: *WERs (%) of the attention-based PIT-ASR models with different configurations in the model architecture, including the number of BLSTM layers in the encoding module (***#L Enc***), the number of LSTM layers in the predictor module (***#L Pred***) and the context window size in the attention module (***#N Ctx***)*

Model	#L Enc	#L Pred	#N Ctx	AVG WER
PIT-ASR	—	—	—	37.2
+ Attention	3	3	5	34.6
	3	3	10	33.1
	3	3	15	**32.4**
	4	2	10	32.6
	5	1	10	**32.3**
	4	2	15	**31.0**

Additional experiments were performed for the proposed attention-based PIT-ASR model. Different configurations with regard to the number of layers in the encoding module and the predictor module, and the context window size in the attention module were implemented and compared. All the results are shown in Table 3. Several observations can be made: 1) The larger the context window in the attention model, the better the performance, since attending on longer embedding sequence can generate more accurate context vectors. 2) When keeping the total number of layers unchanged in the proposed model (with the comparable model scale), more layers in the encoding transformer can produce better embedding representations and thus better performing system. 3) The system with both larger context window and deeper encoding module performs best and achieves relative 17% error reduction compared to the baseline.

4.4. Evaluation of the integrated system

Finally we integrate the gated CNN into the attention-based PIT-ASR model. Based on the 4th line of Attention configurations in Table 3, i.e. 4-layer BLSTM and 2-layer LSTM in the encoder and predictor respectively with the context size 10, two BLSTM layers in encoding module are replaced with two Gated CNN layers, and others are kept the same[2]. The system performance comparison is illustrated in Table 4. It is observed that incorporating gated convolutional networks enables the encoding module to generate better embeddings which achieves another gain upon the attention-based PIT-ASR system.

[2]we have not enough time to finish with other configurations before the submission, such as using the longer context window size

Table 4: *WERs (%) comparison of the integrated systems*

Model	Attention	GCN	AVG WER
PIT-ASR	—	—	37.2
+ Attn	√	—	32.6
+ Attn + GCN	√	√	**31.6**

4.5. Analysis on Attention Models

As stated above, the attention mechanism is exploited here in PIT-ASR to better trace speakers and eliminate interferences by attending on the embedding sequence segments. To better understand and validate this, we do the statistics on the attention weights α_t^i, and measure the difference between α_t^i for the individual speakers i at time step t. The angle difference θ based on the cosine-distance is used:

$$\theta\langle\alpha_t^1, \alpha_t^2\rangle = \arccos \frac{\alpha_t^1 \cdot \alpha_t^2}{\|\alpha_t^1\| \, \|\alpha_t^2\|} \tag{11}$$

A larger angle difference, approaching 90°, indicates the better discrimination between the speakers with attentions, and vice versa. We do the statistics on the evaluation set and the distribution is shown in Figure 4. It shows that only a very small part is with close similarity, around 1% with less than 10°, and a large portion of the distance between the attentions is over 50°. The average distance of attentions between speakers on the open-set evaluation is 50°, which indicates the potential ability on speaker tracing and interference elimination from the proposed new architecture for multi-talker speech recognition.

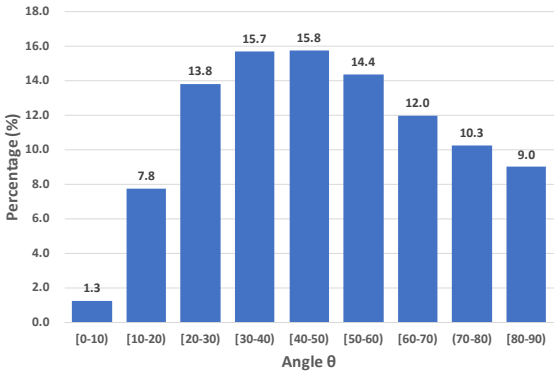

Figure 4: *Distribution of the distance between attentions.*

5. Conclusions

In this paper, we proposed a novel model for monaural multi-talker speech recognition. More specifically, we enhanced the previously developed baseline PIT-ASR model in two ways. First, an attention mechanism that extracts the knowledge from a context segment is designed to provide better speaker tracing and interference elimination ability. Second, gated convolutional networks are incorporated into the encoding transformer to generate better embedding representation. Both techniques lead to performance improvements. The final system reduces WER by more than 15% relatively over our previously proposed baseline PIT-ASR model for monaural multi-talker speech recognition task on the two-talker mixed WSJ0 corpus.

6. References

[1] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, ser. Signals and Communication Technology. Springer London, 2014.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[3] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention." in *Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2016, pp. 17–21.

[4] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8614–8618.

[5] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition." in *Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2013, pp. 3366–3370.

[6] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4580–4584.

[7] M. Bi, Y. Qian, and K. Yu, "Very deep convolutional neural networks for LVCSR," in *Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2015.

[8] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 12, pp. 2263–2276, 2016.

[9] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The Microsoft 2016 conversational speech recognition system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5255–5259.

[10] Y. Qian, T. Tan, H. Hu, and Q. Liu, "Noise robust speech recognition on aurora4 by humans and machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[11] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *accepted by IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2018.

[12] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1670–1679, 2015.

[13] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[14] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2016, pp. 545–549.

[15] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 246–250.

[16] D. Yu, M. Kolbk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[17] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[18] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2456–2460.

[19] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *submitted to Speech Communication,arXiv preprint arXiv:1707.06527*, 2017.

[20] X. Chang, Y. Qian, and D. Yu, "Adaptive permutation invariant training with auxiliary information for monaural multi-talker speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[21] T. Tan, Y. Qian, and D. Yu, "Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[22] Z. Chen and J. Droppo, "Sequence modeling in unsupervised single-channel overlapped speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[23] Y. Qian, C. Weng, X. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, Jan 2018.

[24] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 1, pp. 184–196, Jan 2018.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[26] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[27] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[28] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2015, pp. 577–585.

[29] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[30] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *arXiv preprint arXiv:1710.00343*, 2017.

[31] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit."

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.