INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Learning Contextual Language Embeddings for Monaural Multi-talker Speech Recognition

Wangyou Zhang, Yanmin Qian†

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai
{wyz-97, yanminqian}@sjtu.edu.cn

Abstract

End-to-end multi-speaker speech recognition has been a popular topic in recent years, as more and more researches focus on speech processing in more realistic scenarios. Inspired by the hearing mechanism of human beings, which enables us to concentrate on the interested speaker from the multi-speaker mixed speech by utilizing both audio and context knowledge, this paper explores the contextual information to improve the multi-talker speech recognition. In the proposed architecture, the novel embedding learning model is designed to accurately extract the contextual embedding from the multi-talker mixed speech directly. Then two advanced training strategies are further proposed to improve the new model. Experimental results show that our proposed method achieves a very large improvement on multi-speaker speech recognition, with $\sim$25% relative WER reduction against the baseline end-to-end multi-talker ASR model.

Index Terms: multi-talker speech recognition, cocktail party problem, attention-based end-to-end, contextual embedding

1. Introduction

Over the past few years, much progress has been achieved in single-speaker automatic speech recognition (ASR). Both end-to-end systems and hybrid systems based on deep neural networks (DNN) and hidden Markov model (HMM) have shown surprisingly good performance [1, 2, 3, 4, 5, 6, 7]. However, it is still a challenging task when multiple speakers are involved, which is known as the cocktail party problem [8, 9, 10].

A lot of research has been conducted to tackle the single-channel multi-speaker speech separation and recognition problem. One of the core problems is known as the label ambiguity or permutation problem. In [11, 12, 13], a speech separation method called deep clustering (DPCL) was proposed to separate the speech mixture in a high-dimensional embedding space, where embeddings from the same speaker are close to each other and farther away otherwise. DPCL was then integrated into the speech recognition framework as the separation frontend [14, 15]. Following DPCL, [16] proposed a similar technique called deep attractor network (DANet), which forces time-frequency embeddings to cluster around different centroids representing different speakers in the high-dimensional space. In [17, 18], a simple yet effective speech separation method named permutation invariant training (PIT) was proposed to solve the label ambiguity problem during training, by optimizing the objective of the best output-target pair assignment. It was later applied to multi-speaker speech recognition [19, 20, 21, 22, 23, 24] and showed promising performance.

† corresponding author

In this paper, we aim to further improve the robustness and performance of the end-to-end single-channel multi-speaker ASR systems. When humans recognize the interested speaker from the mixed speech, in addition to the audio signal itself, people will also utilize the context knowledge to better attend, separate and recognize the target speaker's speech. Inspired by this hearing mechanism of human beings, we propose a novel multi-talker speech recognition framework that can learn to extract contextual embeddings from the input mixture, and then use it to boost the speech recognition. This method no longer requires extra knowledge in advance for usage once the model is trained, and it is very flexible and practical in real-world applications. Moreover, we propose two advanced training strategies to further optimize the proposed architecture.

The remainder of the paper is organized as follows. In Section 2, we introduce the baseline end-to-end monaural multi-speaker ASR system. In Section 3, we describe the proposed framework with the contextual language embedding learning and the enhanced training strategies are also given. Experimental results are presented and discussed in Section 4 and conclusions are given in Section 5.

2. End-to-End Multi-speaker Joint CTC/Attention-based Encoder-Decoder

In this section, we revisit the basic end-to-end monaural multi-speaker ASR system proposed in [25], which is the baseline model in our experiments. It extends the joint CTC/attention-based encoder-decoder system proposed in [4, 26, 27] to multi-speaker cases by introducing a separation stage in the encoder and applying permutation invariant training in the objective function. The model architecture can be illustrated in the left part of Figure 1, but without the additional information from the contextual knowledge.

The input speech mixture $\mathbf{O}$ of J speakers is first fed into the multi-speaker encoder, where it is explicitly separated into J sequences of vectors, each representing a speaker source. The multi-speaker encoder module is composed of three stages, i.e. $\mathrm{Encoder}_{\mathrm{Mix}}$, $\mathrm{Encoder}_{\mathrm{SD}}$ and $\mathrm{Encoder}_{\mathrm{Rec}}$, which can be illustrated as follows:

$$\mathbf{H} = \mathrm{Encoder}_{\mathrm{Mix}}(\mathbf{O}), \tag{1}$$

$$\mathbf{H}^j = \mathrm{Encoder}^j_{\mathrm{SD}}(\mathbf{H}), \quad j = 1, \cdots, J, \tag{2}$$

$$\mathbf{G}^j = \mathrm{Encoder}_{\mathrm{Rec}}(\mathbf{H}^j), \ j = 1, \cdots, J. \tag{3}$$

The encoded representations $\mathbf{G}^j$ are then fed into the joint CTC-attention module, which is trained in a multitask manner. The CTC objective function with permutation invariant training

http://dx.doi.org/10.21437/Interspeech.2020-2015

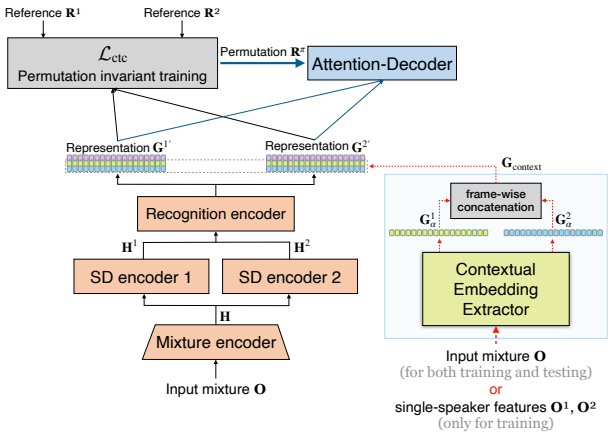

Figure 1: *The proposed architecture for multi-speaker ASR with contextual language embeddings.*

is not only used as an auxiliary task to jointly train the encoder, but also a solution to the label ambiguity problem as shown in Eq. (4):

$$\hat{\pi} = \arg\min_{\pi \in \mathcal{P}} \sum_{j=1}^{J} \text{Loss}_{\text{ctc}}\left(\mathbf{Y}^j, \mathbf{R}^{\pi(j)}\right), \tag{4}$$

where $\mathcal{P}$ denotes the set of all possible permutations on $\{1, \cdots, J\}$, $\pi(j)$ is the j-th element in a permutation $\pi \in \mathcal{P}$, $\mathbf{Y}^j$ denotes the output sequence computed by CTC from the representation $\mathbf{G}^j$, and $\mathbf{R}$ is the set of reference labels for J speakers in the input mixture.

The best permutation $\hat{\pi}$ with the minimum CTC loss is then used in the attention-based decoder to determine the reference label for each decoder output. For each pair of representation and reference label index $(j, \hat{\pi}(j))$, the decoding process can be formulated as follows:

$$y_n^j \sim \text{Attention-Decoder}(\mathbf{G}^j, h_{n-1}), \tag{5}$$

where subscript n denotes the n-th time step of decoding, and h_{n-1} is the $(n-1)$-th element in either the reference label sequence $\mathbf{R}^{\pi(j)}$ or the predicted label sequence $\mathbf{Y}^j$. The technique of choosing h_{n-1} during training is also known as scheduled sampling [28, 24], which can be described as the following equations:

$$b \sim \text{Bernoulli}(p), \tag{6}$$

$$h_{n-1} = \begin{cases} r_{n-1}^{\hat{\pi}(j)}, & \text{if } b = 0, \\ y_{n-1}^j, & \text{if } b = 1, \end{cases} \tag{7}$$

where the history information h_{n-1} is chosen with a probability of p from the the prediction and $(1-p)$ from the ground truth.

The final loss function of the system is defined as the combination of two objectives:

$$\begin{aligned} \mathcal{L} = \sum_j \Big(& \lambda \, \text{Loss}_{\text{ctc}}(\mathbf{Y}^j, \mathbf{R}^{\hat{\pi}(j)}) \\ & + (1-\lambda) \, \text{Loss}_{\text{att}}(\mathbf{Y}^{j,\hat{\pi}(j)}, \mathbf{R}^{\hat{\pi}(j)}) \Big), \end{aligned} \tag{8}$$

where λ is the interpolation factor, and $0 \leq \lambda \leq 1$.

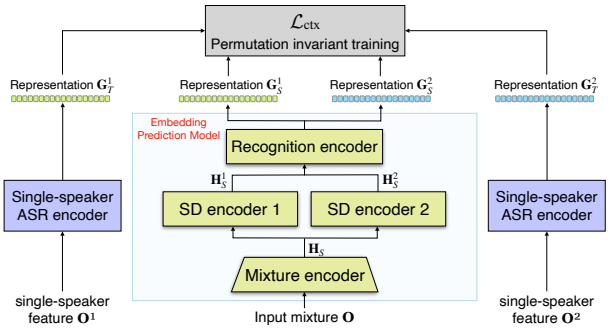

Figure 2: *The proposed architecture for learning contextual language embeddings from the multi-talker mixed speech.*

3. End-to-End Multi-speaker ASR with Contextual Language Embeddings

In this section, we describe the proposed approach for improving the end-to-end single-channel multi-speaker ASR system. First, we propose a novel method that exploits contextual language embeddings learning. Next, we introduce a multi-stage training and an embedding sampling strategy that can further improve the proposed architecture.

3.1. Contextual language embeddings learning

Monaural multi-speaker speech recognition is much more challenging than the single-speaker case, because separating multiple streams from the input mixture is an underdetermined problem, with an infinite number of possible combinations of speech streams. When humans recognize the target speaker from the mixed speech, in addition to the audio signal itself, people will also utilize the context-related information to better attend-separate-recognize the target speaker's speech. Inspired by this human mechanism, we also want to explore the contextual information for recognizing the multi-speaker mixture.

One straightforward way is to utilize the intermediate representations from the single-speaker end-to-end ASR system, whose input is the parallel single-speaker speech in the mixture. The original single-speaker speech can be fed into a pretrained ASR model, and the outputs of the encoder can be regarded as the contextual language embedding for that utterance.

Although it is practical to get the contextual embeddings in this way for training, it is unreasonable to get the contextual embeddings in the same way for testing, as the original clean speech is usually unavailable. In order to address this problem, we need to estimate the contextual embedding for each speaker from the mixed speech directly. In this paper, we design a novel knowledge distillation method to learning the contextual embeddings for the mixed speech. Different from the traditional knowledge distillation work, which usually forces the student model to mimic the output distribution of the teacher [29, 30, 31, 23, 32], we do the knowledge distillation between single-speaker contextual embeddings and the multi-speaker contextual embeddings, and perform the teacher-student learning on the encoder representations of the single-speaker ASR.

Figure 2 illustrates the knowledge distillation framework for learning the contextual embeddings for the mixed speech. The teacher is the encoder module of a pretrained end-to-end single-speaker ASR system, which takes the individual speech of each speaker as input and outputs the corresponding rep-

resentation $\mathbf{G}_T^j$ $(j = 1, \cdots, J)$. The student is the embedding prediction model, with a similar architecture to the encoder of the baseline ASR model introduced in Section 2. It consists of three stages: The mixture encoder, $\mathrm{Encoder}_{\mathrm{Mix}}$, first encodes the input mixture $\mathbf{O}$ as an intermediate representation $\mathbf{H}_S$, which is further processed by J independent speaker-differentiating (SD) encoders $\mathrm{Encoder}_{\mathrm{SD}}$. The outputs $\mathbf{H}_S^j$ $(j = 1, \cdots, J)$ of different SD encoders correspond to different speakers in the mixture. Finally, the recognition encoder, $\mathrm{Encoder}_{\mathrm{Rec,S}}$, transforms the features $\mathbf{H}_S^j$ to high-level representations $\mathbf{G}_S^j$. Our goal is to learn the individual contextual representations of both speakers directly from the mixture, thus the loss function for the knowledge distillation can be formulated as follows:

$$\mathcal{L}_{\mathrm{ctx}} = \sum_j \mathcal{L}_{1;\mathrm{smooth}} \left(\mathbf{G}_S^j, \mathbf{G}_T^{\hat{\pi}_{\mathrm{ctx}}(j)} \right) , \tag{9}$$

$$\mathcal{L}_{1;\mathrm{smooth}}(x, y) = \begin{cases} 0.5(x-y)^2, & \text{if } |x-y| < 1, \\ |x-y| - 0.5, & \text{otherwise,} \end{cases} \tag{10}$$

where $\hat{\pi}_{\mathrm{ctx}}$ is the best permutation that minimizes $\mathcal{L}_{\mathrm{ctx}}$ through permutation invariant training, and the smooth $l1$-loss is used for calculating the final loss. Note that the models in Figure 2 are trained separately, and are then used to provide contextual embeddings to the multi-talker ASR model in Figure 1.

Figure 1 shows the newly proposed multi-speaker ASR architecture that integrates the contextual embeddings into the ASR model. The major difference lies in the encoder output, which can be formulated as follows:

$$\mathbf{G}^{j\prime} = \mathrm{Concat} \left(\mathbf{G}^j, \mathbf{G}_{\mathrm{context}} \right), \quad j = 1, \cdots, J, \tag{11}$$

$$\mathbf{G}_{\mathrm{context}} = \mathrm{Concat} \left(\mathbf{G}_\alpha^1, \cdots, \mathbf{G}_\alpha^J \right), \alpha \in \{T, S\}, \tag{12}$$

where $\mathbf{G}^j$ is the representation generated in Eq. (3), $\mathrm{Concat}(\cdot)$ denotes frame-wise concatenation. Note that the contextual embedding $\mathbf{G}_{\mathrm{context}}$ can be from either the single-speaker teacher encoder ($\alpha = T$) or the embedding prediction model ($\alpha = S$) in multi-speaker ASR training, however only the predicted contextual embedding can be utilized in testing.

3.2. Advanced training strategies for the proposed model

In this subsection, we introduce two enhanced training strategies to further improve the model performance.

The first training strategy divides the training process into two stages. At the first stage, the multi-speaker ASR model without contextual embedding is trained normally for several epochs. Then at the second stage, we exploit the contextual embeddings as described in Figure 2, and continue training the model for the rest epochs. Our motivation is that the contextual embeddings already contain enough acoustic information for recognition, thus involving these features too early may result in an underfitted multi-speaker encoder, which can be suboptimal for training. Therefore, we propose the two-stage training strategy to allow the multi-speaker ASR model to be moderately trained before the contextual embeddings are introduced, which can prevent the model from abusing or overemphasizing the context information.

The second training strategy utilizes both the oracle contextual embeddings from the single-speaker ASR encoder and the predicted contextual embeddings from the prediction model in training, while only the predicted contextual embedding is used in testing. During training, we randomly sample from a Bernoulli distribution in Eq. (6) to determine the source of the contextual embeddings. More specifically, the contextual embeddings come from the oracle contextual embeddings with a probability of p and from the predicted contextual embeddings with a probability of $(1-p)$. We refer to this strategy as embedding sampling, and it shares some similarities with the scheduled sampling technique described in Section 2. It is also capable of mitigating the mismatch between training and testing, and enhances the generalization of proposed multi-talker ASR model with contextual embeddings.

4. Experiments

To evaluate the performance of our proposed methods, we use the single-channel two-speaker mixed speech dataset artificially generated in [32, 33], which is based on the Wall Street Journal (WSJ0) speech corpus [34].

In this section, we first describe the experimental setup in this work. Then the experimental results on the generated 2-speaker mixed WSJ dataset are presented and discussed.

4.1. Experimental setup

As described in Section 4.1 in [32], the 2-speaker mixed WSJ dataset is artificially simulated using the tool released by MERL[1]. The sampling rate of the generated samples is 16 kHz. In each sample, the SNR of one speaker against the other is uniformly sampled from $[-5, 5]$ dB. The duration of the training, development and evaluation sets is 88.2 hr, 1.1 hr and 0.9 hr respectively. Note that this is a larger dataset than the benchmark WSJ0-2mix [11] released by MERL, so that our models can be fully trained.

The input features for all models are the 80-dimensional log-Mel filterbank coefficients with pitch features on each frame, together with their first- and second-order differences. The features were extracted using the Kaldi toolkit [35], and normalized to zero mean and unit variance for training.

The multi-speaker encoder used in Figure 1, as well as the student model for contextual embedding prediction in Figure 2, is composed of two VGG-motivated CNN blocks ($\mathrm{Encoder}_{\mathrm{Mix}}$), one bidirectional long-short term memory layer with projection (BLSTMP) for each speaker ($\mathrm{Encoder}_{\mathrm{SD}}$), and two shared BLSTMP layers ($\mathrm{Encoder}_{\mathrm{Rec}}$). The encoder of the single-speaker ASR teacher model in Figure 2 has a similar structure, with two VGG-motivated CNN blocks followed by three BLSTMP layers. The decoders of both multi-speaker and single-speaker ASR models consist of a single unidirectional long-short term memory (LSTM) layer with 300 cells. All networks were built based on the ESPnet [36] framework with the PyTorch backend.

In the training phase, the AdaDelta optimizer [37] with $\rho = 0.95$ and $\epsilon = 10^{-8}$ was used, and the interpolation factor in Eq. (8) was set to $\lambda = 0.2$. In the decoding phase, a word-level RNN language model (RNNLM) [38] was introduced for rescoring, which was pretrained on the transcriptions from WSJ0 SI-84 and has a single LSTM layer with 1,000 cells. And the interpolation factor λ was set to 0.3, while the weight for RNNLM was set to 1.0. As for knowledge-distillation learning, we used the same single-speaker ASR teacher model as in [32, 33], which was trained on the original WSJ0 corpus. But only the encoder module was used for later knowledge distillation, as described in Section 3.1. The probabilities for scheduled

[1] http://www.merl.com/demos/deep-clustering/create-speaker-mixtures.zip

Table 1: *Performance (Avg. WER) (%) of the proposed contextual embedding method on 2-speaker mixed WSJ corpus.*

Model	dev WER	eval WER
PIT-E2E (baseline) [32]	21.28	23.41
+ scheduled sampling	18.96	22.83
++ context (oracle)	**16.02**	**16.66**
++ context (predicted)	16.55	18.83

sampling and embedding sampling are $p = 0.4$ and $p = 0.7$ respectively in our experiments.

All models were trained for at most 15 epochs, and the model with the best performance on the development set was chosen for the final evaluation.

4.2. Evaluation on the proposed architecture

We first evaluate the performance of the baseline end-to-end model and our proposed new model with contextual embedding on the generated mixed speech evaluation set. The results are presented in Table 1. The first baseline is the end-to-end multi-speaker ASR system in [32], denoted as PIT-E2E, which has the same architecture as in Section 2 but without using the scheduled sampling technique in Eq. (6) $\sim$ (7). It was trained in a teacher-forcing manner, where the history information h_{n-1} in Eq. (7) is always from the ground truth label $r_{n-1}^{\hat{\pi}(j)}$. We also applied scheduled sampling to PIT-E2E, as shown in the second row in Table 1, which serves as the second baseline. It can be observed from Table 1 that the performance of the PIT-E2E model can be slightly improved after applying scheduled sampling during training, as it mitigates the training-inference discrepancy caused by teacher-forcing during training. Therefore, all our proposed methods apply scheduled sampling during training by default.

Then we evaluate the upper bound of our proposed contextual embedding method, where the contextual embeddings in both training and testing come from the single-speaker teacher encoder, denoted as $\mathrm{context\ (oracle)}$. As we can see in Table 1, the performances on both development set and evaluation set are significantly improved after exploiting the contextual information, with more than 15% and 27% relative improvement on the development set and the evaluation set respectively. However, such contextual embeddings are not always available for decoding, as the parallel clean speech from each speaker is required. Therefore, we further evaluate the performance of using the contextual embeddings from the prediction model, denoted as $\mathrm{context\ (predicted)}$, which does not rely on the parallel data for testing. Although we can observe a performance degradation when comparing the predicted embedding with the oracle embedding, it still significantly outperforms the baseline methods, with over 12% and 17% relative improvement on the development set and the evaluation set respectively.

4.3. Evaluation on embedding integration positions

The contextual embedding method used in the last subsection performs embedding integration after the last encoder layer ($\mathrm{Encoder_{Rec}}$). In this subsection, we further investigate how different positions influence the performance of our proposed method. We trained and evaluated the multi-speaker ASR model with embedding integration after the mixture encoder ($\mathrm{Encoder_{Mix}}$), which is a relatively shallow level, and the results are illustrated in Table 2. We can observe that the performance is dramatically degraded when integrating contextual embeddings after $\mathrm{Encoder_{Mix}}$ compared to that after $\mathrm{Encoder_{Rec}}$. The appropriate embedding integrating position is very important for the proposed method, and we will use integration after $\mathrm{Encoder_{Rec}}$ in the following experiments.

Table 2: *Performance (Avg. WER) (%) of different embedding integration positions for the proposed contextual embedding method on 2-speaker mixed WSJ corpus.*

Model	Integration Position	dev WER	eval WER
context (oracle)	after $\mathrm{Encoder_{Rec}}$	**16.02**	**16.66**
context (oracle)	after $\mathrm{Encoder_{Mix}}$	17.79	21.94

Table 3: *Performance (Avg. WER) (%) of different training strategies for the proposed contextual embedding method on 2-speaker mixed WSJ corpus.*

Model	dev WER	eval WER
context (predicted)	**16.55**	18.83
+ embedding sampling	16.98	18.11
+ two-stage training	16.87	17.90
++ embedding sampling	16.90	**17.70**

4.4. Evaluation on the different training strategies

In this subsection, we further explore different training strategies proposed in Section 3.2 for optimizing the usage of the contextual embeddings. Table 3 shows the performance of models trained with different strategies. As we can see, both proposed training strategies can still bring a moderate but consistent improvement compared to the basic training procedure with contextual embedding. We further investigate the combination of the two proposed strategies in training, as shown in the last row of Table 3. As we can see, the performance is further boosted, and $\sim$25% relative WER improvement is achieved finally compared to the baseline model. This demonstrates the effectiveness of our newly proposed method.

5. Conclusions

In this paper, we have proposed a novel context-aware multi-speaker speech recognition framework, which is capable of learning contextual embeddings directly from the input mixture to improve the multi-talker ASR system. Different embedding integration positions are investigated, and two training strategies are designed to further improve the performance. The new architecture can enable the system to act as human beings, utilizing both audio and context information to attend to and separate each target speaker in the mixed speech. The experimental results on the artificially generated two-speaker mixed speech corpus show that the newly proposed method can significantly improve the multi-talker ASR performance.

6. Acknowledgements

This work was supported by the China NSFC project No. U1736202. Experiments have been carried out on the PI supercomputers at Shanghai Jiao Tong University.

7. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE ICASSP*, 2013, pp. 8614–8618.

[3] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *Proc. IEEE ICASSP*, 2018, pp. 5934–5938.

[4] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE ICASSP*, 2017, pp. 4835–4839.

[5] D. Amodei and et al, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. ICML*, 2016, pp. 173–182.

[6] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 4945–4949.

[7] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Trans. ASLP.*, vol. 24, no. 12, pp. 2263–2276, 2016.

[8] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[9] M. A. Bee and C. Micheyl, "The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it?" *Journal of comparative psychology*, vol. 122, no. 3, p. 235, 2008.

[10] 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.

[11] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[12] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, 2016, pp. 545–549.

[13] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE ICASSP*, 2018.

[14] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. IEEE ICASSP*, 2018, pp. 4819–4823.

[15] T. Menne, I. Sklyar, R. Schlüter, and H. Ney, "Analysis of Deep Clustering as Preprocessing for Automatic Speech Recognition of Sparsely Overlapping Speech," in *Proc. ISCA Interspeech*, 2019, pp. 2638–2642. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1728

[16] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE ICASSP*, 2017, pp. 246–250.

[17] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–245.

[18] 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 Trans. ASLP.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[19] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *Proc. ISCA Interspeech*, 2017, pp. 2456–2460.

[20] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Trans. ASLP.*, vol. 26, no. 1, pp. 184–196, Jan 2018.

[21] X. Chang, Y. Qian, and D. Yu, "Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks," in *Proc. ISCA Interspeech*, 2018, pp. 1586–1590.

[22] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, 2018.

[23] T. Tan, Y. Qian, and D. Yu, "Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5714–5718.

[24] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," in *Proc. IEEE ICASSP*, 2019, pp. 6256–6260.

[25] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proc. ACL*, Jul. 2018, pp. 2620–2630.

[26] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, 2017, pp. 1240–1253.

[27] T. Hori, S. Watanabe, and J. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 518–529.

[28] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[29] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Proc. NIPS*, 2014, pp. 2654–2662.

[30] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. ISCA Interspeech*, 2014, pp. 1910–1914.

[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," pp. 1–9, 2014.

[32] W. Zhang, X. Chang, and Y. Qian, "Knowledge distillation for end-to-end monaural multi-talker ASR system," in *Proc. ISCA Interspeech*, 2019, pp. 2633–2637.

[33] W. Zhang, X. Chang, Y. Qian, and S. Watanabe, "Improving end-to-end single-channel multi-talker speech recognition," *IEEE/ACM Trans. ASLP.*, vol. 28, pp. 1385–1394, 2020.

[34] LDC, *LDC Catalog: CSR-I (WSJ0) Complete*, University of Pennsylvania, 1993, `www.ldc.upenn.edu/Catalog/LDC93S6A.html`.

[35] 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," in *Proc. IEEE ASRU*, 2011.

[36] 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 *Proc. ISCA Interspeech*, 2018, pp. 2207–2211.

[37] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[38] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based RNN language models," in *Proc. IEEE SLT*, 2018, pp. 389–396.