INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Knowledge Distillation for End-to-End Monaural Multi-talker ASR System

Wangyou Zhang[*], *Xuankai Chang*[*], *Yanmin Qian*[+]

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

wyz-97@sjtu.edu.cn, xuank@sjtu.edu.cn, yanminqian@sjtu.edu.cn

Abstract

End-to-end models for monaural multi-speaker automatic speech recognition (ASR) have become an important and interesting approach when dealing with the multi-talker mixed speech under cocktail party scenario. However, there is still a large performance gap between the multi-speaker and single-speaker speech recognition systems. In this paper, we propose a novel framework that integrates teacher-student training with the attention-based end-to-end ASR model, which can do the knowledge distillation from the single-talker ASR system to multi-talker one effectively. First the objective function is revised to combine the knowledge from both single-talker and multi-talker labels. Then we extend the original single attention to speaker parallel attention modules in the teacher-student training based end-to-end framework to boost the performance more. Moreover, a curriculum learning strategy on the training data with an ordered signal-to-noise ratios (SNRs) is designed to obtain a further improvement. The proposed methods are evaluated on two-speaker mixed speech generated from the WSJ0 corpus, which is commonly used for this task recently. The experimental results show that the newly proposed knowledge transfer architecture with an end-to-end model can significantly improve the system performance for monaural multi-talker speech recognition, and more than 15% relative WER reduction is achieved against the traditional end-to-end model.

Index Terms: multi-talker speech recognition, cocktail party problem, attention-based end-to-end, knowledge distillation, teacher-student learning

1. Introduction

Recent advances in deep learning have proved that the single-speaker automatic speech recognition (ASR) systems can be trained in the end-to-end manner [1, 2, 3, 4], directly predicting the character sequences. In the end-to-end speech recognition model, a single deep neural network folds the acoustic model (AM), pronunciation and language model (LM) so that they can be optimized simultaneously, which are three individual parts in the traditional deep neural network (DNN) and hidden Markov model (HMM) based hybrid ASR systems [5, 6, 7]. In the past few years, various end-to-end (E2E) models have been developed and they can be categorized into connectionist temporal classification (CTC) based models [8, 9], sequence to sequence (S2S) based models [10, 11] and the models combining CTC and S2S [1]. Even though much progress has been achieved in speech recognition when the speech is involved with stationary noise, it is still challenging in the scenario with nonstationary noise, such as reverberation and speech from other speakers, which is known as the cocktail party problem.

In this work, we aim to address the monaural multi-speaker speech separation and recognition problem. A large amount of research has been done on this problem in recent years. In [12, 13], a method called deep clustering (DPCL) was proposed to separate the mixed speech by mapping each time-frequency (T-F) unit of the signal into a high-dimensional embedding space using a neural network. Then the clustering is performed in the embedding space so that the units from the same dominant speaker are close and farther away otherwise. Another method, called permutation invariant training (PIT), was proposed for both the speech separation [14, 15] and the recognition [16, 17, 18, 19, 20] tasks to train a deep neural network by optimizing the objective of the best output-target assignment at the utterance level. In [21, 22, 23], an end-to-end multi-speaker speech recognition model was brought up using the joint CTC/attention-based encoder-decoder framework [1, 2], in which the encoder first separates the mixed speech and then the attention-based decoder generates the output sequences.

In this paper, we proposed to exploit the knowledge distillation [24] in the end-to-end (E2E) multi-speaker speech recognition system. Most conventional knowledge distillation methods were explored as a compression method by teaching a small model with labels from a more complicated model [25, 26]. In this work, however, the knowledge distillation technique is used to take advantage of the soft label vectors generated from single-speaker model in order to improve the performance of the multi-speaker speech recognition system [20]. Unlike the "one-hot" label vectors, the soft label vector is a less confidential representation thus brings the regularization property to the model. In addition to the information from texts, the soft label vectors introduce supplementary information from the clean signal, so the multi-speaker model can be trained better. Moreover, we adopted the curriculum learning [27, 28] technique to further improve the performance.

The remainder of the paper is organized as follows: In Section 2, the end-to-end monaural multi-speaker ASR model is described. In Section 3, we present the knowledge distillation and the curriculum learning proposed in end-to-end multi-speaker ASR. In Section 4, we evaluate the proposed approaches on the 2-speaker mixed WSJ0 dataset, and the experimental results and analysis are given. Finally the paper is concluded in Section 5.

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

The end-to-end speech recognition model used in our work is the joint CTC/attention-based encoder-decoder proposed in [1, 2, 29]. The advantage of this model is that it uses CTC as a secondary task to enhance the alignment ability of the attention-based encoder-decoder. Later, this model was modified to be fitted in the multi-speaker scenario [22, 23] by introducing a sep-

[*] Equal contribution
[+] Corresponding author

http://dx.doi.org/10.21437/Interspeech.2019-3192

aration stage in the encoder. The input speech mixture is first explicitly separated into multiple sequences of vectors in the encoder, each representing a speaker source. These sequences are fed into the decoder to compute the conditional probabilities.

$\mathbf{O}$ denotes the input speech mixture of S speakers. The encoder consists of three stages: $\mathrm{Encoder}_{\mathrm{Mix}}$, $\mathrm{Encoder}_{\mathrm{SD}}$ and $\mathrm{Encoder}_{\mathrm{Rec}}$. $\mathrm{Encoder}_{\mathrm{Mix}}$, the mixture encoder, encodes $\mathbf{O}$ as an intermediate representation $\mathbf{H}$. Secondly, the representation $\mathbf{H}$ is processed by S independent speaker-different (SD) encoders, $\mathrm{Encoder}_{\mathrm{SD}}$, with S outputs $\mathbf{H}^s$ $(s = 1, \cdots, S)$, each corresponding to the representation of one speaker. In the last stage, for each stream s $(s = 1, \cdots, S)$, $\mathrm{Encoder}_{\mathrm{Rec}}$ transforms the feature sequences $\mathbf{H}^s$ to high-level representations $\mathbf{G}^s$. The encoder can be written as the following steps:

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

$$\mathbf{H}^s = \mathrm{Encoder}_{\mathrm{SD}}{}^s(\mathbf{H}), s = 1, \cdots, S \tag{2}$$

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

A CTC objective function is concatenated after the encoder, whose benefits come in with two folds. The first is to train the encoder of the sequence-to-sequence model as an auxiliary task [1, 2, 29]. The second is that in the multi-speaker case, the CTC objective function is used to perform the permutation-free training shown as in Eq.4, which is also referred to as permutation invariant training (PIT) in [30, 14, 16, 17, 18, 19, 20].

$$\hat{\pi} = \arg\min_{\pi \in \mathcal{P}} \sum_s \mathrm{Loss}_{\mathrm{ctc}}(\mathbf{Y}^s, \mathbf{R}^{\pi(s)}), \tag{4}$$

where $\mathbf{Y}^s$ is the output sequence variable computed from the representation $\mathbf{G}^s$, $\pi(s)$ is the s-th element in a permutation π of $\{1, \cdots, S\}$, and $\mathbf{R}$ is the reference labels for S speakers. Later, the permutation $\hat{\pi}$ with the minimum CTC loss is used for the reference labels in the attention-based decoder in order to reduce the computational cost.

An attention-based decoder network is used to decode each stream $\mathbf{G}^s$ and generates the corresponding output label sequences $\mathbf{Y}^s$. For each pair of representation and reference label index $(s, \hat{\pi}(s))$, the decoding process is described as the following equations:

$$p_{\mathrm{att}}(Y^{s,\hat{\pi}(s)}|\mathbf{O}) = \prod_n p_{\mathrm{att}}(y_n^{s,\hat{\pi}(s)}|\mathbf{O}, y_{1:n-1}^{s,\hat{\pi}(s)}) \tag{5}$$

$$c_n^{s,\hat{\pi}(s)} = \mathrm{Attention}^s(a_{n-1}^{s,\hat{\pi}(s)}, e_{n-1}^{s,\hat{\pi}(s)}, \mathbf{G}^s) \tag{6}$$

$$e_n^{s,\hat{\pi}(s)} = \mathrm{Update}(e_{n-1}^{s,\hat{\pi}(s)}, c_{n-1}^{s,\hat{\pi}(s)}, y_{n-1}^{\hat{\pi}(s)}) \tag{7}$$

$$y_n^{s,\hat{\pi}(s)} \sim \mathrm{Decoder}(c_n^{s,\hat{\pi}(s)}, y_{n-1}^{\hat{\pi}(s)}) \tag{8}$$

where $c_n^{s,\hat{\pi}(s)}$ denotes the context vector, $e_n^{s,\hat{\pi}(s)}$ is the hidden state of the decoder, and $r_n^{\hat{\pi}(s)}$ is the n-th element in the reference label sequence. During training, the reference label $r_{n-1}^{\hat{\pi}(s)}$ in $\mathbf{R}$ is used as a history in the manner of teacher-forcing, instead of $y_{n-1}^{\hat{\pi}(s)}$ in Eq.7 and Eq.8. Eq.5 defines the probability of the target label sequence $\mathbf{Y} = \{y_1, \cdots, y_N\}$ that the attention-based encoder-decoder predicts, in which the probability of y_n at n-th time step is dependent on the previous sequence $y_{1:n-1}$.

The final loss function is defined as

$$\mathcal{L}_{\mathrm{mtl}} = \lambda \mathcal{L}_{\mathrm{ctc}} + (1 - \lambda)\mathcal{L}_{\mathrm{att}}, \tag{9}$$

$$\mathcal{L}_{\mathrm{ctc}} = \sum_s Loss_{\mathrm{ctc}}(\mathbf{Y}^s, \mathbf{R}^{\hat{\pi}(s)}), \tag{10}$$

$$\mathcal{L}_{\mathrm{att}} = \sum_s Loss_{\mathrm{att}}(\mathbf{Y}^{s,\hat{\pi}(s)}, \mathbf{R}^{\hat{\pi}(s)}), \tag{11}$$

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

3. Knowledge Distillation for End-to-End Multi-speaker ASR

In this section, we introduce several techniques to improve the end-to-end multi-speaker ASR system. First, we describe a method called speaker parallel attention that is beneficial for the separation proposed in [23]. Next, we propose to use the teacher-student learning for the knowledge distillation. Third, the curriculum learning is adopted to dig the information underlying the data to improve the training.

3.1. Speaker Parallel Attention

An modification of the attention-based decoder was proposed in [23], called the speaker parallel attention. The motivation is to compensate for the separation ability of the encoder, enhancing the separation performance of the model. The idea was to use individual attention modules for different streams, by virtue of the selective property to filter the noisy information. And the change is simply in the Equation 6:

$$c_n^{s,\hat{\pi}(s)}, a_n^{s,\hat{\pi}(s)} = \mathrm{Attention}^s(a_{n-1}^{s,\hat{\pi}(s)}, c_{n-1}^{s,\hat{\pi}(s)}, \mathbf{G}^s) \tag{12}$$

3.2. Knowledge Distillation

Compared with the hard targets used in the cross entropy criterion, it is claimed that soft targets can provide additional helpful information, leading to better performance [24]. In the multi-speaker speech recognition tasks, we can also use this method to improve the accuracy of attention-based decoder network. To obtain the soft label vectors, the parallel individual speaker's speech goes through the model trained with the speech that contains only one speaker. The soft label vectors contain supplementary information hidden by the overlapping speech as well as the insight from the single-speaker model which has better modelling ability.

The model architecture is shown in Fig.1. The mixed speech and the corresponding individual speech are denoted as $\mathbf{O}$ and $\mathbf{O}^s$ $(s = 1, \cdots, S)$ respectively. Thus, the end-to-end teacher model takes the source speech $\mathbf{O}^s$ as the input to compute teacher logits for each step in the target sequence. And the corresponding outputs, denoted as $\mathbf{Y}_T^s$ $(s = 1, \cdots, S)$, are treated as the target distribution for the student model. Thus the loss function for the teacher-student learning can be expressed as the following:

$$\mathcal{L}_{\mathrm{att-CE}} = \sum_s Loss_{\mathrm{CE}}(\mathbf{Y}^{s,\hat{\pi}(s)}, \mathbf{Y}_T^{\hat{\pi}(s)}) \tag{13}$$

where the knowledge distillation loss $Loss_{\mathrm{CE}}(\mathbf{Y}^s, \mathbf{Y}_T^{\hat{\pi}(s)})$ after the attention-based decoder is computed as the cross entropy between the predictions of the student model and the teacher model, $\hat{\pi}$ is the best permutation determined by the CTC loss. The cross entropy loss can be written as

$$Loss_{\mathrm{CE}}(\mathbf{Y}^s, \mathbf{Y}_T^s) = -\sum_{n=1}^{N} \sum_{c=1}^{|C|} Q(y_{Tn}^s = c|\mathbf{y}_{T0:n-1}^s, \mathbf{O}^s; \theta_T) \times \log P(y_n^s = c|y_{0:n-1}^s, \mathbf{O}; \theta) \tag{14}$$

where, θ_T corresponds to the parameters in the teacher model; θ corresponds to the learning parameters in the student model; $Q(\cdot)$ and $P(\cdot)$ represent the distributions for every speaker from the teacher and student model respectively.

In this paper, we modified our loss function of the attention-based decoder $\mathcal{L}_{att}$. The new form is the weighted sum of the

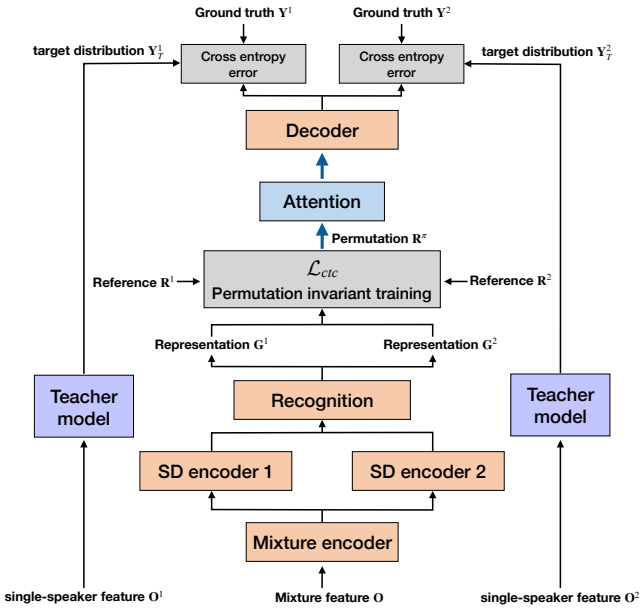

Figure 1: *The proposed knowledge distillation architecture for end-to-end multi-speaker speech recognition in the 2-speaker case.*

original loss based on cross entropy (CE) and the term based on knowledge distillation loss, namely

$$\mathcal{L}^*_{att} = \eta\mathcal{L}_{att} + (1-\eta)\mathcal{L}_{att-CE} \tag{15}$$

where η is the weight coefficient.

3.3. Curriculum learning

In previous works, the end-to-end multi-speaker ASR systems were trained disregarding the similarities and differences underlying the data. In some researches [27, 28], however, it is claimed that the order of the data has an influence on the training process, called curriculum learning strategy. Thus, here we would like to find a pattern from the data to make the training procedure more stable and to boost the performance. According to [19], one observation is that the signal-to-noise ratio (SNR) between the overlapped speech has a great influence on the separation performance. In utterances with small SNRs, the speeches from different speakers are distorted with similar energy. On the contrary, a large SNR means the speeches are distorted in an unbalanced condition with one dominant speech.

In our work, we focus on the SNR level of the overlapping speech, which is defined as the energy ratio between the source speech from two speakers. Other factors may also be used, but the methodology is the same. When generating the mixed speech, the energy ratio is randomly selected in order to simulate the real conditions. When the SNR is larger, the high energy speech is clearer, but the speeches with lower energy are ill behaved. On the contrary, when the SNR is smaller, each utterance in the mixed speech can be recognized with similar performance, thus the model can learn the knowledge from each speakers. We rearranged the training data in the way described in Algorithm 1. Specifically, in the beginning of training, we iterate through minibatches in the training set in the ascending order of the SNR of speech from speaker 1. Afterwards, the training reverts back to the random order over minibatches.

Algorithm 1: Curriculum learning

```
Load the training dataset X;
Sort the training data in X in ascending order of the
  SNR of utterances;
while model is not converged do
    for each i in all minibatches of training data do
        Feed minibatch i into the model and perform
          gradient descent;
    end
end
Shuffle the training data randomly and divide them into
  minibatches;
Feed each minibatch into the model iteratively and
  updata the model;
Repeat step 8 and step 9 until converge.
```

4. Experiments

4.1. Experimental Setup

To evaluate our proposed methods, we artificially generated the single-channel two-speaker mixed signals based on the Wall Street Journal (WSJ0) speech corpus [31], using the tool released by MERL[1]. The training, development and evaluation data were generated from the WSJ0 SI-84, Dev93 and Eval92 respectively, and the durations of each dataset are as follows: 88.2 hr for training, 1.1 hr for development, and 0.9 hr for evaluation.

The input features are the 80-dimensional log-Mel filterbank coefficients with pitch features on each frame, concatenated with their delta and delta delta coefficients. All features were extracted using the Kaldi toolkit [32] and normalized to zero mean and unit variance.

In this work, the neural network models in different approaches have the same depth and similar size so that their performance is comparable. The encoder consists of two VGG-motivated CNN blocks and three bidirectional long-short term memory recurrent neural networks with projection (BLSTMP), while the decoder network has only one unidirectional long-short term memory (LSTM) layer with 300 cells. All networks were built based on the ESPnet [33] framework. The AdaDelta optimizer [34] with $\rho = 0.95$ and $\epsilon =$ 1e-8 was used for training. The interpolation factor λ in Eq.9 was set to 0.2 during training.

For teacher-student training, an end-to-end teacher model was first trained on the original clean speech training dataset from WSJ0. In our experiments, the WER of the teacher model on WSJ0 Dev93 and Eval92 are 8.0% and 2.1% respectively. Then we fed the mixed speech data and the corresponding individual speech data into the teacher-student module simultaneously. The best performance was achieved when the weight coefficient η in Eq.15 was set to 0.5 in our experiments.

In the decoding phase, we combined both the joint CTC/attention score and the score of the pretrained word-level RNN language model (RNNLM), which has 1-layer LSTM with 1000 cells and was trained on the transcriptions from WSJ0 SI-84, in a shallow fusion manner. The beam width was set to 30. The interpolation factor λ used during decoding was 0.3, and the weight for RNNLM was 1.0.

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

4.2. Experiments on teacher-student training and curriculum learning

Table 1: *Performance (Avg. CER & WER) (%) on 2-speaker mixed **WSJ0** corpus. Comparison between End-to-End multi-speaker joint CTC/attention-based encoder-decoder systems*

Model	dev CER	eval CER
multi-speaker (baseline)	13.72	15.31
+ parallel Att	12.48	14.51
+ TS	11.27	14.69
+ TS + parallel Att	11.46	13.54
++ CL	**10.84**	**11.97**

Model	dev WER	eval WER
multi-speaker (baseline)	21.24	23.41
+ parallel Att	20.28	23.04
+ TS	18.29	22.82
+ TS + parallel Att	18.84	21.64
++ CL	**17.78**	**19.80**

We first evaluated the performance of the baseline end-to-end methods and our proposed methods on the mixed speech test dataset in WSJ0. The results are presented in Table 1. The first method is the joint CTC/attention-based encoder-decoder network for multi-speaker speech, where the attention-decoder module is shared among representations of each speaker. The second method extends the single attention to speaker parallel attention modules. We treated these two methods as the baseline systems.

Then the teacher-student learning and curriculum learning were applied step by step. With the teacher-student training, it can be observed that the performances of the both baseline systems are improved on both the dev and eval dataset. A larger performance boost is even achieved on the speaker parallel attention method, 7% and 6% relative reduction of the average WER on the dev and eval dataset respectively. This proves that speaker parallel attention method has stronger capability of eliminating irrelevant information for current individual speaker, and that it can learn better with the knowledge from the attention output distribution of the teacher model. Next we applied the curriculum learning strategy on the teacher-student framework to further improve the performance. As we can see in Table 1, our proposed end-to-end method combining teacher-student training, speaker parallel attention and curriculum learning significantly improves the performance of two-speaker mixed speech recognition, with more than 15% relative improvement in both WER and CER.

4.3. Experiments on different curriculum learning strategies

To investigate the impact of the curriculum learning strategy on the performance of the models, we explored different strategies. We tested on the end-to-end model with teacher-student training and speaker parallel attention, with two different strategies: sorting the training data in the ascending order of SNR and in the descending order of SNR. The experimental results are shown in Table 2.

When the training data is sorted in the descending order of SNR (absolute value), the model performed worse than the one trained with the opposite order, even worse than the model trained with randomly sorted data, which proves our conjecture in Section 3.3. When the SNR is small, the energy difference between two speakers is subtle and the model learns the separation ability. Later, the accuracy performance is enhanced with the data having larger SNRs.

Table 2: *Performance (Avg. CER & WER) (%) of different curriculum learning strategies on the test dataset of the 2-speaker mixed **WSJ0** corpus.*

Model	CER	WER
TS + parallel Att	13.54	21.64
++ CL (ascending SNRs)	**11.97**	**19.80**
++ CL (descending SNRs)	14.49	22.18

5. Conclusion

In this work, we have applied the sequence-level knowledge distillation and the curriculum learning techniques to the multi-speaker end-to-end speech recognition system based on the joint CTC/attention-based encoder-decoder framework. A single-speaker end-to-end speech recognition teacher model was used to compute the soft label vectors as the target distribution to compute the final loss function. To make the best use of the training data, we further rearrange the data in the ascending order of the SNR. Finally, our proposed model achieved over 15% relative improvement on CER & WER.

In our future work, we would like to investigate other curriculum learning strategies including other factors. And knowledge distillation is only applied on the attention-based decoder, which can also be extended to the CTC part.

6. Acknowledgements

This work was supported by the China NSFC projects (No. 61603252 and No. U1736202). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

7. References

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

[2] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *J. Sel. Topics Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[3] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[4] Z. Chen, Q. Liu, H. Li, and K. Yu, "On modular training of neural acoustics-to-word model for lvcsr," in *(ICASSP)*, 2018, pp. 4819–4823.

[5] 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.

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

[7] 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 (ICASSP)*, 2017, pp. 5255–5259.

[8] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, 2014, pp. 1764–1772.

[9] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *IEEE Workshop on (ASRU)*, 2015, pp. 167–174.

[10] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: first results," *arXiv preprint arXiv:1412.1602*, 2014.

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

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

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

[14] 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 (ICASSP)*, 2017, pp. 241–245.

[15] 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 (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[16] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *(INTERSPEECH)*, 2017, pp. 2456–2460.

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

[18] X. Chang, Y. Qian, and D. Yu, "Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks," in *(INTERSPEECH)*, 2018, pp. 1586–1590.

[19] 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.

[20] T. Tan, Y. Qian, and D. Yu, "Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition," in *IEEE (ICASSP)*, 2018.

[21] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4819–4823.

[22] 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 *(ACL) (Volume 1: Long Papers)*, 2018, pp. 2620–2630.

[23] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker asr system without pretraining," in *IEEE (ICASSP)*, 2019.

[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[25] R. Pang, T. Sainath, R. Prabhavalkar, S. Gupta, Y. Wu, S. Zhang, and C.-C. Chiu, "Compression of end-to-end models," in *Proc. Interspeech 2018*, 2018, pp. 27–31.

[26] S. Wang, Y. Yang, T. Wang, Y. Qian, and K. Yu, "Knowledge distillation for small foot-print deep speaker embedding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6021–6025.

[27] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[28] D. Amodei and et al, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. JMLR.org, 2016, pp. 173–182.

[29] T. Hori, S. Watanabe, and J. Hershey, "Joint ctc/attention decoding for end-to-end speech recognition," in *(ACL) (Volume 1: Long Papers)*, vol. 1, 2017, pp. 518–529.

[30] 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.

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

[32] 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," IEEE Signal Processing Society, Tech. Rep., 2011.

[33] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[34] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012. [Online]. Available: http://arxiv.org/abs/1212.5701