





Improving End-to-End Single-Channel Multi-Talker Speech Recognition

Wangyou Zhang , *Student Member, IEEE*, Xuankai Chang , *Student Member, IEEE*, Yanmin Qian , *Senior Member, IEEE*, and Shinji Watanabe , *Senior Member, IEEE*

Abstract—Although significant progress has been made in single-talker automatic speech recognition (ASR), there is still a large performance gap between multi-talker and single-talker speech recognition systems. In this article, we propose an enhanced end-to-end monaural multi-talker ASR architecture and training strategy to recognize the overlapped speech. The single-talker end-to-end model is extended to a multi-talker architecture with permutation invariant training (PIT). Several methods are designed to enhance the system performance, including speaker parallel attention, scheduled sampling, curriculum learning and knowledge distillation. More specifically, the speaker parallel attention extends the basic single shared attention module into multiple attention modules for each speaker, which can enhance the tracing and separation ability. Then the scheduled sampling and curriculum learning are proposed to make the model better optimized. Finally the knowledge distillation transfers the knowledge from an original single-speaker model to the current multi-speaker model in the proposed end-to-end multi-talker ASR structure. Our proposed architectures are evaluated and compared on the artificially mixed speech datasets generated from the WSJ0 reading corpus. The experiments demonstrate that our proposed architectures can significantly improve the multi-talker mixed speech recognition. The final system obtains more than 15% relative performance gains in both character error rate (CER) and word error rate (WER) compared to the basic end-to-end multi-talker ASR system.

Index Terms—Multi-talker mixed speech recognition, permutation invariant training, end-to-end model, knowledge distillation, curriculum learning.

I. INTRODUCTION

THANKS to the advances in deep learning, automatic speech recognition (ASR) has achieved a huge progress. Deep neural networks (DNN) and hidden Markov model (HMM) based hybrid systems have achieved a very good performance, which are comparable with, or even surpassing, human performance [1]–[3]. Recently, there have been growing interests in

developing end-to-end systems for speech recognition, in which multiple modules in the hybrid systems, such as the acoustic model (AM), lexicon model, and language model (LM), are folded into a single neural network model, so that they can be optimized simultaneously. Over the past few years, a variety of end-to-end (E2E) models have been proposed and they can be mainly categorized into connectionist temporal classification (CTC) based models [4], [5], and sequence to sequence (S2S) based models [6], [7]. The combined mode with both CTC and S2S [8] is also designed to further improve the end-to-end ASR system. The end-to-end systems have shown promising results according to existing works [8]–[10]. On the other hand, although a huge progress has been achieved on ASR, the current systems mainly focus on single-talker speech, and there is still a large performance gap between single-talker and multi-talker speech recognition. Processing the multi-talker mixed speech is a key problem when multi-talker mixed speech commonly exists in the complex real-world conditions, especially under the cocktail party scenarios [11]–[13].

In this work, we aim to address the monaural multi-speaker speech separation and recognition problem. A large amount of research has been done to tackle this problem in recent years. In [14], [15], a speech separation 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, where T-F units dominated by the same speaker are close and those dominated by different speakers are farther away. Later, a simple yet effective technique, called permutation invariant training (PIT), was proposed for both multi-talker speech separation [16], [17] and speech recognition [18]–[22], which trains a deep neural network by optimizing the objective of the best output-target pair assignment at the utterance level. In [23], [24], an end-to-end model for multi-speaker speech recognition was brought up using the joint CTC/attention-based encoder-decoder framework [8], [25], in which the encoder first separates the mixed speech and then the attention-based decoder generates the output sequences based on the separated streams.

Given the success of an end-to-end model for multi-speaker speech recognition [23], [24], this paper further extends the prior study by focusing a novel neural network architecture and training methods. The novelties are summarized as follows:

- i) We revise the permutation invariant training (PIT) based model in [24], using speaker parallel attention modules for each speaker to enhance the speaker tracing and

Manuscript received December 2, 2019; revised March 7, 2020; accepted April 13, 2020. Date of publication April 20, 2020; date of current version May 14, 2020. This work was supported by the China NSFC Project No. U1736202. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jinyu Li. (*Corresponding author: Yanmin Qian.*)

Wangyou Zhang and Yanmin Qian are with the SpeechLab, Department of Computer Science and Engineering & MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wyz-97@sjtu.edu.cn; yanminqian@sjtu.edu.cn).

Xuankai Chang and Shinji Watanabe are with the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: xchang14@jhu.edu; shinjiw@jhu.edu).

Digital Object Identifier 10.1109/TASLP.2020.2988423

separation ability as well as to alleviate the burden of the encoder.

- ii) We adopt the scheduled sampling technique [26] to mitigate the training-inference discrepancy caused by teacher-forcing during training.
- iii) We design a curriculum learning method [9], [27] to exploit the data scheduling scheme which can make the model better optimized, and three modes are explored and compared.
- iv) Finally, an architecture is designed to distill knowledge from the single-speaker model to the multi-speaker model in the end-to-end framework, which is motivated by the knowledge distillation [28]–[30].

This paper is an extension of our previous study [31], [32], which proposes several new methods to improve the multi-talker speech recognition with an end-to-end architecture. In this paper, we summarize these proposed methods with a consistent formulation and further extend the curriculum learning with various modes. Additional experiments are also conducted to clearly specify the effectiveness of the proposed methods under the same experimental condition as in [32].

We evaluate the proposed architectures on the artificially generated WSJ data with two-talker mixed speech. The experimental results show that our proposed architectures can significantly improve the performance of end-to-end single-channel multi-talker speech recognition.

The rest of the paper is organized as follows. In Section II we define the problem of the single-channel multi-talker mixed speech recognition. In Section III we propose the new methods based on the end-to-end architecture to recognize the multi-talker mixed speech. The experimental results and analysis are presented in Section IV and finally the conclusion is given in Section V.

II. SINGLE-CHANNEL MULTI-TALKER SPEECH RECOGNITION

In this paper, we assume that only a single microphone is available, which is common in many real-world conditions, therefore a single-channel speech signal $y[n]$ is observed. And $y[n]$ is a mixture signal and can be assumed to be a linear combination of multiple speech sources, i.e.

$$y[n] = \sum_{s=1}^S x_s[n], \quad (1)$$

where $x_s[n]$ ($s = 1, \dots, S$) represents streams from S different speakers. Then our goal is to separate these streams and recognize each of them simultaneously. However, given the single-channel speech signal $y[n]$, the problem of separating S different streams is underdetermined, because the number of possible combinations of $x_s[n]$ is infinite. Therefore, it is difficult to derive well-separated streams in the signal level before speech recognition. On the other hand, the sparsity assumption of the speech signals in the time-frequency domain has been adopted in many works on speech enhancement [33]–[35] and speech separation [36], [37], which assumes that the speech from a certain speaker only occupies some part of the feature representation of the mixed signal in the time-frequency domain, and

thus is separable in the feature space. Under such assumption, we can separate different streams in the feature space and perform speech recognition on these separated features, which can be done by designing a separation module and a recognition module in the model architecture [24].

Note that speech recognition of multiple speakers is much harder than that of a single speaker. In the single speaker case, i.e. $S = 1$, the problem is significantly simplified because the input stream to be recognized only consists of one single speaker, thus it can be cast as a simple supervised optimization problem. When multiple speakers are involved, however, the problem becomes much more complicated. We not only need to separate different streams of multiple speakers, but also have to handle the label ambiguity or permutation problem. In the case of two speakers, while the input is the feature of mixed speech from speaker s_1 and s_2 , the output of the model is two labels corresponding to speaker s_1 and s_2 respectively. However, the permutation of output labels (\hat{Y}^1, \hat{Y}^2) is not guaranteed to be invariant, i.e. we do not know whether \hat{Y}^1 corresponds to speaker s_1 or speaker s_2 . Previous works such as [16] and [38] have demonstrate the influence of the label ambiguity problem on training with conventional supervised approach for speech separation. In order to address the label ambiguity problem, several techniques such as permutation invariant training [16], [18]–[22], deep clustering [14], [15], [39] and deep attractor network (DANet) [40], [41] can be used, as described in Section I.

III. END-TO-END MULTI-TALKER SPEECH RECOGNITION

While there are several existing solutions to the label ambiguity problem, as introduced in the last section, DPCL and DANet are not as straightforward as those in PIT [16], [18], [22], and cannot be easily applied to direct recognition of multiple streams of speech without first separation in training or evaluation. Therefore, in this paper, we adopt the permutation invariant training method in end-to-end single-channel multi-talker mixed speech recognition. For simplicity and without loss of generality, we always assume there are two talkers in the mixed speech when describing our architectures in this section.

In this section, we first describe the basic end-to-end multi-talker ASR system that has been used in [24]. Then we introduce the proposed four techniques to improve the end-to-end multi-talker ASR system, including speaker parallel attention, scheduled sampling, curriculum learning and knowledge distillation.

A. End-to-End Multi-Talker ASR With Joint CTC/Attention-Based Encoder-Decoder

The basic framework of end-to-end multi-speaker ASR system in this paper is the joint CTC/attention-based encoder-decoder framework proposed in [8], [25], [42], which is illustrated in Fig. 1. It takes advantage of CTC as a secondary task to enhance the alignment ability of the attention-based encoder-decoder. Later, this model was extended to be applied in the multi-speaker scenario [24], [31] by introducing a separation stage in the encoder and allowing the permutation-free training in the objective function.

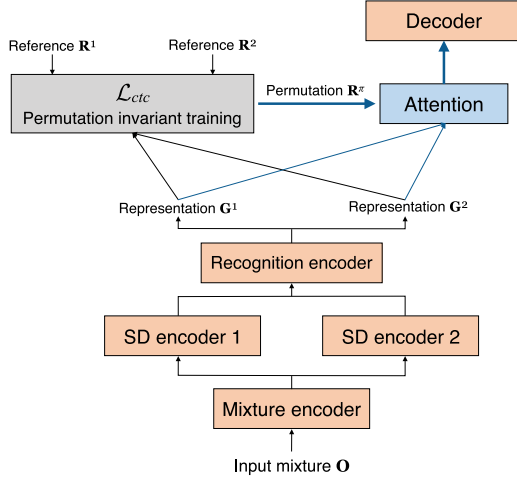


Fig. 1. End-to-End multi-speaker speech recognition model in the 2-Speaker case.

As illustrated in Fig. 1, the input speech mixture \mathbf{O} is first fed into the encoder, which is composed of three stages: $\text{Encoder}_{\text{Mix}}$, $\text{Encoder}_{\text{SD}}$ and $\text{Encoder}_{\text{Rec}}$. $\text{Encoder}_{\text{Mix}}$ is the mixture encoder that encodes \mathbf{O} as an intermediate representation \mathbf{H} , which is then processed by S independent speaker-differentiating (SD) encoders $\text{Encoder}_{\text{SD}}$. Each SD encoder outputs a representation \mathbf{H}^s ($s = 1, \dots, S$) of one speaker and prepare it for recognition. The recognition encoder corresponds to an acoustic model that transforms the single-speaker feature to high-level representations \mathbf{G}^s for the final decoding. The encoder can be formulated as follows:

$$\mathbf{H} = \text{Encoder}_{\text{Mix}}(\mathbf{O}), \quad (2)$$

$$\mathbf{H}^s = \text{Encoder}_{\text{SD}}^s(\mathbf{H}), s = 1, \dots, S, \quad (3)$$

$$\mathbf{G}^s = \text{Encoder}_{\text{Rec}}(\mathbf{H}^s), s = 1, \dots, S. \quad (4)$$

A CTC objective function with permutation invariant training is concatenated after the encoder, whose benefits come in with two folds. The first is to jointly train the encoder of the sequence-to-sequence model as an auxiliary task [8], [25], [42]. The second is to solve the label ambiguity problem by performing permutation invariant training as shown in Eq. (5):

$$\hat{\pi} = \arg \min_{\pi \in \mathcal{P}} \sum_{s=1}^S \text{Loss}_{\text{ctc}}(\mathbf{Y}^s, \mathbf{R}^{\pi(s)}), \quad (5)$$

where \mathcal{P} is the set of all permutations on $\{1, \dots, S\}$, \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 , and \mathbf{R} is the set of reference labels for S speakers. Later, the permutation $\hat{\pi}$ with the minimum CTC loss is chosen for the order of reference labels in the attention-based decoder to reduce the computational cost.

The autoregressive attention-based decoder network decodes each stream \mathbf{G}^s and generates the corresponding output label sequence \mathbf{Y}^s . For each pair of representation and reference label

index $(s, \hat{\pi}(s))$, the decoding process is described as follows:

$$p_{\text{att}}(\mathbf{Y}^{s, \hat{\pi}(s)} | \mathbf{O}) = \prod_n p_{\text{att}}(y_n^{s, \hat{\pi}(s)} | \mathbf{O}, y_{1:n-1}^{s, \hat{\pi}(s)}), \quad (6)$$

$$c_n^{s, \hat{\pi}(s)} = \text{Attention}(e_{n-1}^{s, \hat{\pi}(s)}, \mathbf{G}^s), \quad (7)$$

$$e_n^{s, \hat{\pi}(s)} = \text{Update}(e_{n-1}^{s, \hat{\pi}(s)}, c_{n-1}^{s, \hat{\pi}(s)}, r_{n-1}^{s, \hat{\pi}(s)}), \quad (8)$$

$$y_n^{s, \hat{\pi}(s)} \sim \text{Decoder}(c_n^{s, \hat{\pi}(s)}, r_{n-1}^{s, \hat{\pi}(s)}), \quad (9)$$

where $c_n^{s, \hat{\pi}(s)}$ denotes the context vector obtained with an attention mechanism,

$e_n^{s, \hat{\pi}(s)}$ is the hidden state of the decoder, and $r_n^{s, \hat{\pi}(s)}$ is the n -th element in the reference label sequence. During training, the reference label $r_{n-1}^{s, \hat{\pi}(s)}$ in \mathbf{R} is used as a history in the manner of teacher-forcing, instead of the output label $y_{n-1}^{s, \hat{\pi}(s)}$ in Eq. (8) and (9). The probability of the target label sequence $\mathbf{Y} = \{y_1, \dots, y_N\}$ predicted by the attention-based encoder-decoder is defined in Eq. (6), where the probability of y_n at the n -th time step is dependent on the previous sequence $y_{1:n-1}$.

The final loss function \mathcal{L}_{mtl} of the system is defined as the combination of two objective functions:

$$\mathcal{L}_{\text{mtl}} = \lambda \mathcal{L}_{\text{ctc}} + (1 - \lambda) \mathcal{L}_{\text{att}}, \quad (10)$$

$$\mathcal{L}_{\text{ctc}} = \sum_s \text{Loss}_{\text{ctc}}(\mathbf{Y}^s, \mathbf{R}^{\hat{\pi}(s)}), \quad (11)$$

$$\mathcal{L}_{\text{att}} = \sum_s \text{Loss}_{\text{att}}(\mathbf{Y}^{s, \hat{\pi}(s)}, \mathbf{R}^{\hat{\pi}(s)}), \quad (12)$$

where λ is the interpolation factor, and $0 \leq \lambda \leq 1$. Loss_{ctc} and Loss_{att} are the cross entropy (CE) loss functions.

B. Proposed Method (i): Speaker Parallel Attention Modules

Since the acoustic characteristics and energy of different speakers are usually very different, the encoder has to compensate for the differences while separating the mixed speech. Our motivation is to alleviate the burden for the encoder and to make the attention-decoder module learn to filter the separated speech as well while keeping the model compact. In light of [20], we propose to use independent attention modules for different streams, called speaker parallel attention [31]. Fig. 2 illustrates the architecture of the model, in which *Attention 1* and *Attention 2* are two independent modules for each speaker respectively. The computation process in Eq. (7) is then changed in a stream-specific way, in particular for the s -th stream, as:

$$c_n^{s, \hat{\pi}(s)} = \text{Attention}^s(c_{n-1}^{s, \hat{\pi}(s)}, \mathbf{G}^s). \quad (13)$$

We hope the speaker parallel attention modules can enhance the model ability on the speaker tracing and separation, which should be useful for the final recognition.

C. Proposed Method (ii): Scheduled Sampling

Different from the teacher-forcing method used in the basic end-to-end multi-speaker ASR model in Section III-A, we adopt the scheduled sampling [26] to alleviate the exposure bias [43],

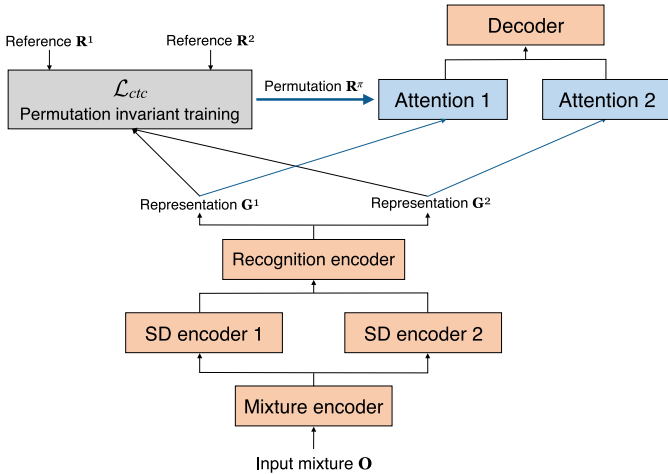


Fig. 2. End-to-End multi-speaker speech recognition model with speaker parallel attention modules in the 2-speaker case.

[44], since the model is never exposed to its own prediction errors during training with the teacher-forcing mode. It can also reduce the mismatch between training and inference, because we only have access to the predicted token y_n from the model itself in the inference phase, which is important especially in the multi-speaker speech recognition task due to the label permutation problem. The scheduled sampling technique changes the training process from the teacher-forcing scheme to a more flexible scheme, which utilizes the predicted token with a large probability and the reference token with a small probability. During training, whether the history information is chosen from the ground truth label or the prediction is determined randomly, with a probability of p from the prediction and $(1 - p)$ from the ground truth.

After the modifications, Eq. (8) and Eq. (9) should be changed as:

$$e_n^{s, \hat{\pi}(s)} = \text{Update}(e_{n-1}^{s, \hat{\pi}(s)}, c_{n-1}^{s, \hat{\pi}(s)}, h), \quad (14)$$

$$y_n^{s, \hat{\pi}(s)} \sim \text{Decoder}(c_n^{s, \hat{\pi}(s)}, h), \quad (15)$$

where

$$b \sim \text{Bernoulli}(p), \quad (16)$$

$$h = \begin{cases} r_{n-1}^{\hat{\pi}(s)}, & \text{if } b = 0, \\ y_{n-1}^{s, \hat{\pi}(s)}, & \text{if } b = 1. \end{cases} \quad (17)$$

D. Proposed Method (iii): Curriculum Learning for Multi-Talker ASR

According to some research [9], [27], the order of the data is proven to have an influence on the model optimization, which is called the curriculum learning strategy. It will start the training with some simple samples and then progressively increase the difficulty of training data, which is similar as the curriculum learning process for humans. Therefore, here we would like to utilize the curriculum learning to better optimize the model and

further improve the system performance on multi-talker speech recognition.

According to [21], one observation is that the signal-to-noise ratio (SNR, the energy ratio between the source speech from two speakers) between the overlapped speech can significantly influence the separation performance. When $|\text{SNR}|^1$ is small, i.e. the energy of the target speech is similar to that of the interfering speech, each utterance in the mixed speech can be recognized with similar performance, thus the model can learn the knowledge from each speaker. On the contrary, a large $|\text{SNR}|$ means the energy in the mixed speech is unbalanced. It is dominated by the high-energy speech from one speaker, which is clearer and easier to recognize, but the recognition of the interfering speech with lower energy is much more difficult. Therefore, we can sort the mixed speech data in the ascending order of $|\text{SNR}|$ between the two speakers, and start with mixtures where both speakers have similar energy levels in the training phase. We think that the more balanced energy in the mixed speech is easier for the model training, especially at the beginning of the optimization. And this is proven effective in our previous study [32].

Another useful information is the genders of the speakers in the mixed speech. As observed in [45], different-gender (M-F) mixtures have a better separability than the same-gender (M-M or F-F) mixtures, and the M-M mixtures seems slightly easier than the F-F mixtures.² Therefore, we can also sort the mixed speech data with the order of different gender combinations, i.e. following the order of M-F \rightarrow M-M \rightarrow F-F.

In addition, the length of input speech also indicates the difficulty of separation and recognition. On one hand, it is similar to the case of single-speaker speech recognition in which longer sequences tend to be harder to model. On the other hand, as the length of the mixed speech becomes longer, the risk of speaker permutation occurring grows in the middle. Thus we can sort the mixed speech data in the ascending order of length as well.

In our work, we evaluate the proposed curriculum learning with all three data scheduling schemes on the end-to-end multi-talker ASR, and the common strategy is described in Algorithm 1. More specifically, we iterate through minibatches on the training set with one of the specific orders in the first several training epochs. After that, the model is further finetuned with the random order over minibatches.

E. Proposed Method (iv): Knowledge Distillation for End-to-End Multi-Talker ASR

When training models described in previous sections, we only use the hard labels in the cross entropy criterion. In [30], however, it is reported that the soft targets from another model can provide additional valuable information such as the similarity structure over the data, leading to a better performance. This method is called knowledge distillation [30] (also known as the teacher-student learning in some application scenarios [28], [29]). Different from the former applications using knowledge

¹ $|\cdot|$ denotes the absolute value.

² M-F denotes the mixture of a male and a female speaker, F-F denotes the mixture of two female speakers, and M-M denotes the mixture of two male speakers.

Algorithm 1: Curriculum learning for end-to-end multi-talker ASR

```

1 Load the training dataset  $\mathbf{X}$ ;
2 if  $type = SNR$  then
3   Sort the training data in  $\mathbf{X}$  in ascending order of
    $|\text{SNR}|$ ;
4 else if  $type = length$  then
5   Sort the training data in  $\mathbf{X}$  in ascending order of the
   length of utterances;
6 else if  $type = gender$  then
7   Sort the training data in  $\mathbf{X}$  in the order of the type of
   gender combinations: M-F  $\rightarrow$  M-M  $\rightarrow$  F-F;
8 end
9 while model is not converged do
10  for each  $i$  in all minibatches of training data do
11    Feed minibatch  $i$  into the model and perform
    gradient descent;
12  end
13 end
14 Shuffle the training data randomly and divide them into
   minibatches;
15 Feed each minibatch into the model iteratively and
   update the model;
16 Repeat step 14 and step 15 until converge.
    
```

distillation for model compression in most previous works [46]–[48], we propose a knowledge distillation framework [32] to transfer knowledge from a single-speaker model, i.e. the teacher, to a multi-speaker one, i.e. the student. It is beneficial to reducing the gap between the single-talker and multi-talker ASR systems, and to improving the recognition accuracy of multi-talker ASR model.

To obtain the soft label vectors, the parallel original single-speaker speech from each speaker is fed into the teacher model, which is built with the single-speaker speech in advance.

The whole architecture is shown in Fig. 3. The mixed speech and the corresponding individual speech are denoted as \mathbf{O} and \mathbf{O}^s ($s = 1, \dots, S$) respectively. As we can see, the end-to-end single-speaker teacher model takes the source speech \mathbf{O}^s as the input to compute teacher logits (the inputs to the final softmax) for each step in the target sequence. And the corresponding outputs, denoted as \mathbf{Y}_T^s ($s = 1, \dots, S$), are treated as the target distribution for the multi-speaker student model. Thus the loss function for the teacher-student learning can be formulated as follows:

$$\mathcal{L}_{\text{att-CE}} = \sum_s \text{Loss}_{\text{CE}}(\mathbf{Y}^s, \hat{\mathbf{Y}}_T^s), \quad (18)$$

where the knowledge distillation loss $\text{Loss}_{\text{CE}}(\mathbf{Y}^s, \hat{\mathbf{Y}}_T^s)$ 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 still the best permutation determined by the CTC loss. The cross entropy loss can be written as

$$\begin{aligned} \text{Loss}_{\text{CE}}(\mathbf{Y}^s, \hat{\mathbf{Y}}_T^s) &= - \sum_{n=1}^N \sum_{c=1}^{|C|} Q(y_{Tn}^s = c | \mathbf{y}_{T0:n-1}^s, \mathbf{O}^s; \theta_T) \\ &\times \log P(y_n^s = c | \mathbf{y}_{0:n-1}^s, \mathbf{O}; \theta), \end{aligned} \quad (19)$$

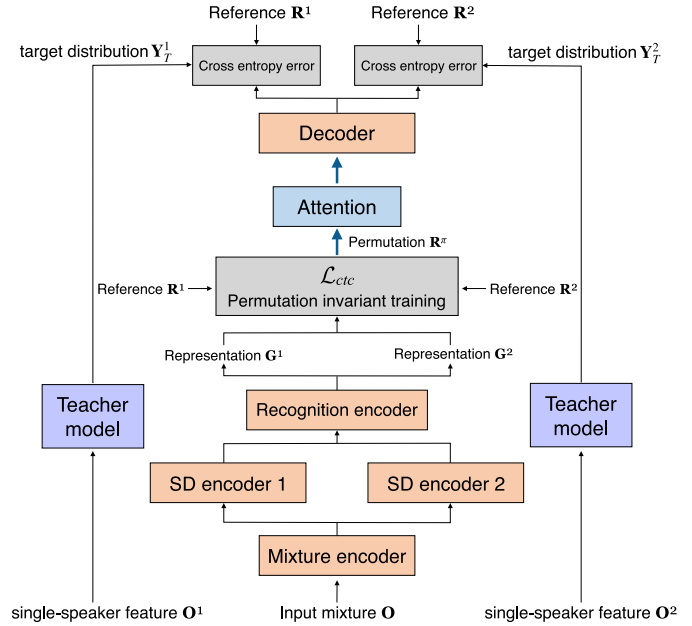


Fig. 3. The proposed knowledge distillation architecture for end-to-end multi-speaker speech recognition in the 2-speaker case.

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 combine both hard targets from the true labels and soft targets from the teacher model to optimize the performance. The loss function of the attention-based decoder \mathcal{L}_{att} is thus modified to the weighted sum of the original loss based on cross entropy and the term based on knowledge distillation loss, namely

$$\mathcal{L}_{\text{att}}^* = \eta \mathcal{L}_{\text{att}} + (1 - \eta) \mathcal{L}_{\text{att-CE}}, \quad (20)$$

where η is the interpolation factor and $0 \leq \eta \leq 1$.

IV. EXPERIMENT

To evaluate the performance of our proposed architectures, all experiments were conducted on the same two-talker mixed speech dataset, which is artificially generated from the Wall Street Journal (WSJ0) reading speech corpus [49].

In this section, we will first introduce the experimental setup in this work, and then the experimental results on the WSJ0-2mix dataset are presented and discussed.

A. Experimental Setup

We artificially generated the single-channel two-speaker mixed speech based on the WSJ0 corpus, using the tool released by MERL.³ Note that the length of the generated mixture is determined by the longer sample when mixing speech from two

³[Online]. Available: <http://www.merl.com/demos/deep-clustering/create-speaker-mixtures.zip>

speakers. The sampling rate of the dataset is 16 kHz. The training, development and evaluation data were generated from the WSJ0 SI-84, Dev93 and Eval92 respectively, and the duration is 88.2 hours for training, 1.1 hours for development, and 0.9 hours for evaluation individually. The SNR of one speaker against the other is randomly sampled from a uniform distribution on the interval $[-5, 5]$ dB.

The input features for the ASR systems 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 [50] and normalized to zero mean and unit variance.

In our experiments, all the neural network models in different approaches have the same depth and a similar size so that their performance is comparable. The encoder is composed of two VGG-style convolutional neural network (CNN) blocks [51] and three bidirectional long-short term memory recurrent neural networks with projection (BLSTMP) [52], while the decoder network has only one unidirectional long-short term memory (LSTM) layer with 300 cells. Note that all networks were built based on the ESPnet framework [53] with the Pytorch back-end. The AdaDelta optimizer [54] with the running average parameter $\rho = 0.95$ and the constant $\epsilon = 10^{-8}$ was used for training. The interpolation factor λ in Eq. (10) was set to 0.2 during training. All models are trained at most 15 epochs, but the training may be finished early if no performance improvement is observed for 3 consecutive epochs. During training, the model is evaluated on the development set after each epoch. After training, the model with the best performance on the development set is used for final evaluation on the evaluation set.

In the decoding phase, we combined both the joint CTC/attention score and the score of the word-level RNN language model (RNNLM), which has a 1-layer LSTM with 1000 cells and was trained on the transcriptions of WSJ0 SI-84, in a shallow fusion manner. The beam width for the beam search process was 30. The interpolation factor λ in Eq. (10) was set to 0.3 during decoding, and the weight for RNNLM was 1.0.

B. Evaluation of PIT-E2E Model on WSJ0-2mix

Firstly, we compare the performance of the usual end-to-end single-speaker model and the basic multi-speaker model (called PIT-E2E from now on) described in Section III-A for recognizing the multi-talker mixed speech.

The single-speaker model is the joint CTC/attention-based encoder-decoder network trained on the original WSJ0 dataset, and word error rate (WER) on WSJ0 Dev93 and Eval92 is 8.0% and 2.1% respectively. Its architecture is similar to that in Fig. 1, and the difference is two-fold. First, the encoder module is composed of a three-layer BLSTMP following the CNN block, rather than divided into three stages. Second, there is only one representation after the encoder and thus no permutation invariant training is demanded. For both models, each BLSTMP layer has 1024 memory cells in each direction. The experimental results on the generated WSJ0-2mix dataset are presented in Table I. Note that the CER and WER of the single-speaker model

TABLE I
PERFORMANCE (AVG. CER & WER) (%) COMPARISON OF THE END-TO-END SINGLE-SPEAKER MODEL AND PIT-E2E MULTI-SPEAKER MODEL ON THE WSJ0-2MIX DATASET

Model	dev CER	eval CER
E2E single-speaker	75.46	77.80
PIT-E2E multi-speaker	13.72	15.31

Model	dev WER	eval WER
E2E single-speaker	113.26	115.94
PIT-E2E multi-speaker	21.28	23.41

TABLE II
PERFORMANCE (AVG. CER & WER) (%) EVALUATION OF THE SPEAKER PARALLEL ATTENTION ARCHITECTURE ON THE WSJ0-2MIX DATASET

Model	dev CER	eval CER
PIT-E2E	13.72	15.31
+ speaker parallel attention	12.48	14.51

Model	dev WER	eval WER
PIT-E2E	21.28	23.41
+ speaker parallel attention	20.28	23.04

is measured by comparing the output against the reference labels of both speakers.

As we can see in the table, the capability of the single-speaker model is very limited on the two-talker mixed speech, and its performance on the overlapped speech degrades severely compared to the single-talker utterance, even more than 100% WER. In contrast, the speech recognition system designed for multiple speakers can significantly improve the performance on the multi-talker mixed speech, with more than 80% relative reduction on both average CER and WER. This demonstrates the effectiveness of the end-to-end multi-speaker speech recognition architecture to recognize the overlapped speech.

In addition, another baseline can be a two-stage method performing speech separation at the first stage and E2E single-speaker ASR for each separated stream at the second stage. Interested readers can refer to Section 4.2.4 in [24], which shows that the purely E2E multi-speaker ASR system has a comparable performance to the two-stage method proposed in [23]. Therefore, in this work, we will focus on improving the PIT-E2E multi-speaker ASR system.

C. Evaluation of Speaker Parallel Attention on WSJ0-2mix

In this subsection, the proposed speaker parallel attention architecture is evaluated and compared on the WSJ0-2mix dataset. On the basic PIT-E2E model, the single shared attention module is extended to two independent attention modules for each speaker source. The rest of the network is kept the same as the PIT-E2E model, containing a 2-layer CNN Encoder_{Mix}, an 1-layer BLSTMP Encoder_{SD}, a 2-layer BLSTMP Encoder_{Rec}, and a shared 1-layer LSTM as the decoder network. The results are illustrated in Table II.

We can observe that the speaker parallel attention module achieves an obvious reduction on both the average CER and WER. This result demonstrates the better separation capability of the proposed speaker parallel attention, and the independent

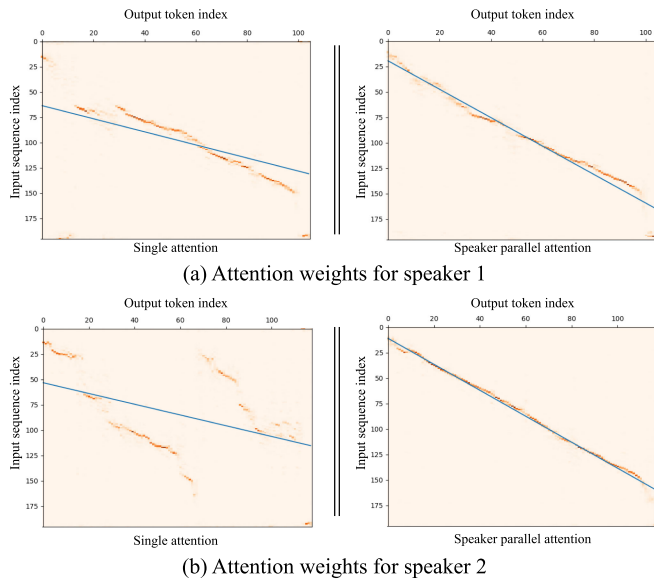


Fig. 4. Visualization of the sequences of attention weights for two overlapped speakers. The left part is from the basic PIT-E2E model with single attention and the right part is from the PIT-E2E model with speaker-parallel-attention. Attention weights are depicted in orange, while the best fitting curve of linear regression is depicted in blue.

attention modules for each speaker can improve the multi-speaker ASR based on the PIT-E2E architecture.

For further investigation on the efficacy of the proposed speaker parallel attention, we visualize the attention weight sequences for two overlapped speakers in one utterance, which are generated by the basic PIT-E2E model with single attention and the enhanced PIT-E2E model with speaker-parallel-attention respectively. In Fig. 4, the horizontal axis represents the sequence of output tokens and the vertical axis represents the input sequence to the attention module. And the attention weights are depicted in orange. The left parts of Fig. 4(a) and (b) show the attention weights for speaker 1 and speaker 2 generated by the PIT-E2E model with single attention, while the right parts show the attention weights generated by the enhanced speaker-parallel-attention model. We can observe that the right parts are more smooth and clear, and the attention weights are more concentrated. This observation conforms with the characteristics of alignments between the output sequence and the input sequence for speech recognition, which further shows the superiority of the proposed speaker parallel attention model. In addition, we also compute the variance of linear regression on the evaluation set. The linear regression on each attention sample is weighted by its attention weights for both curve fitting and calculating the variance. The best-fitting curve is depicted in blue in Fig. 4. For the basic PIT-E2E model, the average variance is 2055.89; for our proposed model with speaker parallel attention, the average variance is 1718.65. This observation also illustrates the effectiveness of the proposed method.

D. Evaluation of Scheduled Sampling for PIT-E2E ASR Model

In this subsection, we evaluate the scheduled sampling technique on PIT-E2E ASR model for overlapped speech

TABLE III
PERFORMANCE (AVG. CER & WER) (%) EVALUATION OF THE SCHEDULED SAMPLING ON THE PIT-E2E FOR THE WSJ0-2MIX DATASET

Model	dev CER	eval CER
PIT-E2E	13.72	15.31
+ scheduled sampling	11.73	14.61
++ speaker parallel attention	11.65	14.29

Model	dev WER	eval WER
PIT-E2E	21.28	23.41
+ scheduled sampling	18.96	22.83
++ speaker parallel attention	18.75	22.19

TABLE IV
PERFORMANCE (AVG. CER & WER) (%) COMPARISON OF DIFFERENT CURRICULUM LEARNING STRATEGIES ON THE WSJ0-2MIX DATASET

Model	dev CER	eval CER
PIT-E2E	13.72	15.31
+ CL (ascending SNRs)	11.09	13.52
+ CL (gender:FM→MM→FF)	10.99	14.36
+ CL (ascending lengths)	11.36	13.93

Model	dev WER	eval WER
PIT-E2E	21.28	23.41
+ CL (ascending SNRs)	18.11	20.79
+ CL (gender:FM→MM→FF)	17.19	21.75
+ CL (ascending lengths)	18.06	21.56

recognition, and the results are shown in Table III. The sampling probability p in Eq. (16) is 0.3 in our experiments. It is noted that although scheduled sampling slows the convergence of our model by about 2 epochs during training, the total time of training is not influenced.

It is observed that the performance of the basic PIT-E2E model can be further improved by applying the scheduled sampling approach during training. The scheduled sampling approach is capable of reducing the mismatch between the training and inference, thus leading to performance improvement. We further applied the scheduled sampling to the enhanced PIT-E2E with speaker parallel attention, and the results are shown as the last row in Table III. It shows that scheduled sampling still works well on the enhanced architecture with speaker parallel attention, and combining both strategies can get a further improvement.

E. Evaluation of the Curriculum Learning for PIT-E2E ASR

Then we implemented the curriculum learning strategies described in Section III-D on the PIT-E2E ASR model. In our experiments, the curriculum learning strategy is applied for the first 3 training epochs, and then random shuffling is used for the remaining 12 epochs. Several strategies of curriculum learning are compared, including reordering the data by SNR, gender mixture and utterance length, and the results are shown in Table IV. It can be observed that our proposed curriculum learning strategies all can bring decent performance improvement compared to the basic PIT-E2E system. All three strategies can achieve $\sim 10\%$ relative WER improvement, indicating that an appropriate data scheduling scheme can significantly affect the multi-speaker ASR system. For the three modes, doing the curriculum learning with ascending SNRs seems the best on the eval set, which will also be utilized in the following experiments. Overall, the results

TABLE V
PERFORMANCE (AVG. WER) (%) COMPARISON OF DIFFERENT MODELS
(PIT-E2E/PIT-E2E + CL) ON THE WSJ0-2MIX EVALUATION SET

Mixture type	SNR	High E WER	Low E WER
F-F	0 – 2.5 dB	55.34 / 49.68	55.67 / 49.82
	2.5 – 5 dB	52.25 / 56.69	46.48 / 46.16
M-M	0 – 2.5 dB	24.27 / 21.91	32.03 / 26.32
	2.5 – 5 dB	25.94 / 19.62	34.94 / 30.76
M-F	0 – 2.5 dB	8.69 / 8.68	15.08 / 13.41
	2.5 – 5 dB	9.70 / 9.43	18.11 / 14.86

Sample Length	SNR	High E WER	Low E WER
≤ 10s	0 – 2.5 dB	16.15 / 15.01	24.98 / 20.24
	2.5 – 5 dB	19.66 / 15.25	30.36 / 24.00
> 10s	0 – 2.5 dB	24.43 / 22.44	27.92 / 25.31
	2.5 – 5 dB	21.48 / 20.91	25.90 / 24.97

show the flexibility of the proposed method, enabling us to use different strategies when the respective information (e.g. SNR) is available in different applications.

To better illustrate how different SNR levels, gender combinations and lengths of utterances influence the recognition performance, we evaluate the basic PIT-E2E model and the model trained with ascending SNRs (2nd row in Table IV) on different subsets of the evaluation set. The results are presented in Table V, which show the recognition performance on samples with different gender combinations and different lengths. For each case, we evaluate the WER on the high energy (High E) source and low energy (Low E) source in the mixed speech, under either low SNR conditions or high SNR conditions. Since the SNR of one source against the other ranges from -5 to 5 dB, the value of $|\text{SNR}|$ lies between 0 and 5 dB, and we simply take 2.5 dB as the threshold for high SNRs. From the first subtable, we can clearly see that the recognition performance of both systems degrades as the mixture type changes from M-F to M-M to F-F, and our method outperforms the baseline in almost all cases.⁴ When $|\text{SNR}|$ is relatively low, the WERs of both speakers are close; when $|\text{SNR}|$ is relatively high, the WER of the low energy speaker is usually increased. These observations further confirm the assumptions in Section III-D and demonstrate the effectiveness of our proposed curriculum learning method. Similar conclusions can also be drawn from the second subtable, which shows that our method brings a consistent improvement on samples of different lengths, and the shorter utterances are easier to be recognized than the longer ones.

F. Evaluation of Knowledge Distillation for PIT-E2E ASR

The proposed knowledge distillation is evaluated for the end-to-end multi-talker ASR model in this subsection. We first apply the teacher-student learning to the basic PIT-E2E model. The teacher model used in our experiments is the single-speaker end-to-end model with joint CTC/attention, which was trained

⁴For F-F mixtures with SNRs ranging from 2.5 to 5 dB, the WER of the high energy source is higher than that of the low energy source, which may be due to the small amount of such samples (only 4.5% of the entire evaluation set).

TABLE VI
PERFORMANCE (AVG. CER & WER) (%) EVALUATION OF THE KNOWLEDGE
DISTILLATION FOR PIT-E2E ASR MODEL ON THE WSJ0-2MIX SPEECH

Model	dev CER	eval CER
PIT-E2E	13.72	15.31
+ teacher-student learning	11.27	14.69
++ speaker parallel attention	11.46	13.54
+++ CL (ascending SNRs)	10.84	11.97

Model	dev WER	eval WER
PIT-E2E	21.28	23.41
+ teacher-student learning	18.29	22.82
++ speaker parallel attention	18.84	21.64
+++ CL (ascending SNRs)	17.78	19.80

on the clean single-speaker speech training dataset from the original WSJ0 corpus. To perform teacher-student learning, we feed the multi-speaker mixed speech feature and the corresponding single-speaker feature into the teacher-student architecture simultaneously. The best performance was achieved when the weight coefficient η in Eq. (20) was set to 0.5 in our experiments. Then the other techniques described in Sections III-B, III-C and III-D are further integrated to boost the performance. All methods are evaluated on the WSJ0-2mix dataset and the results are shown in Table VI. It is noted that the scheduled sampling technique is applied for the last two rows in Table VI, and is only used for the student model, while the teacher model still uses history information from the ground truth label for decoding.

It is observed that the knowledge distillation from the single-speaker model to the multi-speaker model is useful for the proposed PIT-E2E architecture when recognizing overlapped speech, and PIT-E2E with teacher-student learning can obtain a substantial improvement on both dev and eval datasets. Then the previously proposed speaker parallel attention and curriculum learning with $|\text{SNR}|$ reordering are also applied within this knowledge distillation framework. We can see that all the approaches can be combined to achieve a better performance, and the best system is more than 15% relative lower on both WER and CER upon the basic PIT-E2E model for multi-speaker ASR.

G. Performance Comparison With Previous Works on the Benchmark WSJ0-2mix Speech

Finally, we compared the final system in our work with other related works on multi-talker speech recognition. We trained and tested our model on the benchmark WSJ0-2mix dataset, which was released by MERL [14]. Compared to the above used WSJ0-2mix dataset introduced in Section IV-A, the benchmark dataset is relatively small, with approximately 30 hours of training data and 10 hours of validation data. The top part of Table VII shows the WER comparison of the previous work on this dataset, including DNN-HMM hybrid systems using PIT-Hybrid-ASR proposed in [21] and using DPCL-based speech separation proposed in [15], and the basic end-to-end ASR systems constructed in [24]. Both [15] and [21] use a trigram language model obtained by a standard Kaldi recipe, while [24] uses the character and word level RNNLMs pretrained on the WSJ text corpus, and our system uses the same RNNLM as our previous experiments which is described in Section IV-A

TABLE VII

WER (%) COMPARISON OF OUR NEWLY PROPOSED MODEL AND OTHER RELATED WORKS ON THE BENCHMARK WSJ0-2MIX DATASET RELEASED BY MERL [14]

Model	avg WER
DPCL+ASR [15]	30.8
PIT-Hybrid-ASR [21]	28.2
End-to-end ASR (Char/Word-LM) [24]	28.2
PIT-E2E with our proposed methods	23.4

of this paper. The result using all our proposed approaches is illustrated as the last row of Table VII. Note that the model in [24] was trained on a different, larger training dataset than that used in other experiments. From Table VII, we can see that our new system constructed by the proposed methods is significantly better than all the previous architectures on this benchmark WSJ0-2mix speech.

V. CONCLUSION

In this paper, we proposed an enhanced end-to-end architecture to recognize single-channel multi-talker mixed speech, which is based on permutation invariant training for solving the label ambiguity problem.

The basic joint CTC/attention-based encoder-decoder framework was enhanced with several approaches, including speaker parallel attention, scheduled sampling, curriculum learning and knowledge distillation. The speaker parallel attention can enhance the tracing and separation ability on multiple streams, and the scheduled sampling and curriculum learning can make the training easier and better optimized. The end-to-end based knowledge distillation transfers the knowledge from single-speaker ASR to multi-speaker ASR. All the proposed new approaches were evaluated and compared on the artificially generate the WSJ0-2mix dataset with two-talker mixed speech. The experiments on the WSJ0-2mix dataset demonstrate that all the proposed methods are very useful to improve the end-to-end multi-speaker ASR system, and the best system can obtain more than 15% relative improvement on both CER and WER.

Although significant improvement is achieved by the proposed approach, there is still an accuracy gap when compared to the usual single-speaker speech recognition. It is even more difficult and challenging when facing the more spontaneous speech under real noisy scenarios such as the AMI meeting corpus [55] and the LibriCSS conversation corpus [56]. In our future work, we would like to develop more advanced multi-talker ASR architectures, which can show better noise-robustness in such real noisy environments. Moreover, we will also extend our work from single-channel to multi-channel conditions, where we can exploit spatial information to achieve a better performance.

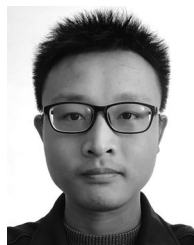
ACKNOWLEDGMENT

Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

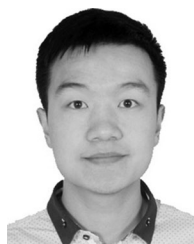
REFERENCES

- [1] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
- [2] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 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 Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5934–5938.
- [4] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1764–1772.
- [5] Y. Miao, M. Gowayyed, and F. Metze, "EESSEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 167–174.
- [6] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *NIPS 2014 Deep Learn. Repres. Learn. Workshop*, 2014.
- [7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4960–4964.
- [8] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 4835–4839.
- [9] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.
- [10] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4945–4949.
- [11] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.
- [12] 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?" *J. Comparative Psychol.*, vol. 122, no. 3, pp. 235–251, 2008.
- [13] Y. Qian, C. Weng, X. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 40–63, Jan. 2018.
- [14] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.
- [15] 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.
- [16] 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 Int. Conf. Acoust., Speech Signal Process.*, 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.
- [18] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *Proc. ISCA Interspeech*, 2017, pp. 2456–2460.
- [19] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 184–196, Jan. 2018.
- [20] 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.
- [21] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Commun.*, vol. 104, pp. 1–11, 2018.
- [22] T. Tan, Y. Qian, and D. Yu, "Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5714–5718.
- [23] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4819–4823.

- [24] 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. Assoc. Comput. Linguist.*, Jul. 2018, pp. 2620–2630.
- [25] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.
- [26] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1171–1179.
- [27] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
- [28] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2654–2662.
- [29] 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.
- [30] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS 2014 Deep Learn. Representation Learn. Workshop*, 2014, pp. 1–9.
- [31] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6256–6260.
- [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] D. Wu, W.-P. Zhu, and M. Swamy, "On sparsity issues in compressive sensing based speech enhancement," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2012, pp. 285–288.
- [34] V. Abrol, P. Sharma, and A. K. Sao, "Speech enhancement using compressed sensing," in *Proc. ISCA Interspeech*, 2013, pp. 3274–3278.
- [35] S. Kammi and M. R. K. Mollaei, "Noisy speech enhancement with sparsity regularization," *Speech Commun.*, vol. 87, pp. 58–69, 2017.
- [36] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 620–633, Mar. 2014.
- [37] S. Gannot *et al.*, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
- [38] M. Kolbaek, 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, Oct. 2017.
- [39] 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.
- [40] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 246–250.
- [41] Z. Chen *et al.*, "Cracking the cocktail party problem by multi-beam deep attractor network," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2018, pp. 437–444.
- [42] T. Hori, S. Watanabe, and J. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguist. (Volume 1: Long Papers)*, 2017, pp. 518–529.
- [43] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proc. ICLR*, 2016.
- [44] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," in *Proc. EMNLP*, 2016, pp. 1296–1306.
- [45] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1535–1546, Jul. 2017.
- [46] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. EMNLP*, 2016, pp. 1317–1327.
- [47] R. Pang *et al.*, "Compression of end-to-end models," in *Proc. ISCA Interspeech*, 2018, pp. 27–31.
- [48] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 126–136, Jan. 2018.
- [49] LDC, *LDC Catalog: CSR-1 (WSJ0) Complete*, University of Pennsylvania, 1993. [Online]. Available: www.ldc.upenn.edu/Catalog/LDC93S6A.html
- [50] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Autom. Speech Recognit. Understanding*, 2011.
- [51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.
- [52] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. ISCA Interspeech*, 2014, pp. 338–342.
- [53] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. ISCA Interspeech*, 2018, pp. 2207–2211.
- [54] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.
- [55] I. McCowan *et al.*, "The AMI meeting corpus," in *Proc. 5th Int. Conf. Methods Techn. Behavioral Res.*, vol. 88, 2005, pp. 137–140.
- [56] Z. Chen *et al.*, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE ICASSP*, 2020, pp. 7284–7288.



Wangyou Zhang (Student Member, IEEE) received the B.Eng. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2018. He is currently working toward the Ph.D. degree in Shanghai Jiao Tong University, Shanghai, China, under the supervision of Yanmin Qian. His current research interests include robust speech recognition, speech signal processing and deep learning.



Xuankai Chang (Student Member, IEEE) received the B.Eng. and M.Eng. degrees in 2016 and 2019, respectively, from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is currently working toward the Ph.D. degree in the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA, advised by Shinji Watanabe. His current research interests include end-to-end speech recognition and the cocktail party problem. He was also the recipient of the Best Paper Award of ASRU in 2019.



Yanmin Qian (Senior Member, IEEE) received the B.S. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2012. Since 2013, he has been with the Department of Computer Science and Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, China, where he is currently an Associate Professor. From 2015 to 2016, he also worked as an Associate Research in the Speech Group, Cambridge University Engineering Department, Cambridge, U.K. His current research interests include the acoustic and language modeling in speech recognition, speaker and language recognition, key word spotting, and multimedia signal processing.



Shinji Watanabe (Senior Member, IEEE) received the B.S., M.S., and Ph.D. (Dr. Eng.) degrees in 1999, 2001, and 2006, respectively, from Waseda University, Tokyo, Japan. He is an Associate Research Professor with Johns Hopkins University, Baltimore, MD, USA. From 2001 to 2011, he was a Research Scientist at NTT Communication Science Laboratories, Kyoto, Japan. From January to March, in 2009, he was a Visiting Scholar with the Georgia Institute of Technology, Atlanta, GA, USA. From January 2012 to June 2017, he was a Senior Principal Research Scientist at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. He has authored or co-authored more than 200 papers in top journals and conferences. His research interests include Bayesian machine learning and speech and spoken language processing. He was the recipient of several awards including the Best Paper Award from the IEICE in 2003. He served as an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and he is a Member of several technical committees including the IEEE Signal Processing Society Speech and Language Technical Committee.