Speech Communication 104 (2018) 1–11

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Single-channel multi-talker speech recognition with permutation invariant training

Yanmin Qian[a,*], Xuankai Chang[a], Dong Yu[b]

[a] Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] Tencent AI Lab, Bellevue, USA

ARTICLE INFO

Keywords:
Permutation invariant training
Multi-talker mixed speech recognition
Feature separation
Joint-optimization

ABSTRACT

Although great progress has been made in automatic speech recognition (ASR), significant performance degradation is still observed when recognizing multi-talker mixed speech. In this paper, we propose and evaluate several architectures to address this problem under the assumption that only a single channel of mixed signal is available. Our technique extends permutation invariant training (PIT) by introducing the front-end feature separation module with the minimum mean square error (MSE) criterion and the back-end recognition module with the minimum cross entropy (CE) criterion. More specifically, during training we compute the average MSE or CE over the whole utterance for each possible utterance-level output-target assignment, pick the one with the minimum MSE or CE, and optimize for that assignment. This strategy elegantly solves the label permutation problem observed in the deep learning based multi-talker mixed speech separation and recognition systems. The proposed architectures are evaluated and compared on an artificially mixed AMI dataset with both two- and three-talker mixed speech. The experimental results indicate that against the state-of-the-art single-talker speech recognition system our proposed architectures can cut the word error rate (WER) by relative 45.0% and 25.0% across all speakers when their energies are comparable, for two- and three-talker mixed speech, respectively. To our knowledge, this is the first work on the single-channel multi-talker mixed speech recognition on the challenging speaker-independent spontaneous large vocabulary continuous speech task.

1. Introduction

Thanks to the significant progresses made in recent years (Yu et al., 2010; Seide et al., 2011; Hinton et al., 2012; Dahl et al., 2012; Abdel-Hamid et al., 2012; 2014; Yu and Deng, 2014; Sainath et al., 2015; Bi et al., 2015; Qian et al., 2016; Qian and Woodland, 2016; Mitra and Franco, 2015; Peddinti et al., 2015; Sercu et al., 2016; Amodei et al., 2016; Zhang et al., 2016a; Yu et al., 2016; Xiong et al., 2017), ASR systems have now surpassed the threshold for adoption in many real-world scenarios and have enabled services such as Microsoft Cortana, Apple's Siri and Google Now, where close-talk microphones are commonly used.

However, current ASR systems still perform poorly when far-field microphones are used. This is because many difficulties hidden by close-talk microphones now surface under distant recognition scenarios. For example, the signal to noise ratio (SNR) between the target speaker and the interfering noises is much lower than that when close-talk microphones are used. As a result, the interfering signals, such as background noise, reverberation, and speech from other talkers, become so distinct that they can no longer be ignored.

In this paper, we aim at solving the speech recognition problem when multiple talkers speak at the same time and only a single channel of mixed speech is available. Although multi-channel speech processing is important for many applications now, it is still very necessary (and interesting) to conduct research on single-channel multi-talker speech recognition for four reasons. First, many recording devices, such as those used by reporters, only have one microphone. Second, even with microphone array, single-channel multi-talker ASR is still needed when the two speakers are in the same direction and thus cannot be separated by a beamformer. Third, single-channel results set a lower-bound on what is achievable when multi-channel information is available. Fourth, it sheds lights on new solutions other than beamformer when multi-channel information (esp. with ad hoc mic-array) is available.

Many attempts have been made to address the problem of single-channel multi-talker speech recognition. Before the deep learning era, the most famous and effective model is the factorial GMM-HMM, which outperformed humans in the 2006 monaural speech separation and recognition challenge (Kristjansson et al., 2006; Hershey et al., 2010; Ming et al., 2010; Cooke et al., 2010). The factorial GMM-HMM, however, requires the test speakers to be seen during training so that the

[*] Corresponding author.
E-mail address: yanminqian@sjtu.edu.cn (Y. Qian).

https://doi.org/10.1016/j.specom.2018.09.003
Received 13 November 2017; Received in revised form 13 July 2018; Accepted 3 September 2018
Available online 7 September 2018

interactions between them can be properly modeled. Recently, several deep learning based techniques have been proposed to solve this problem (Weng et al., 2015; Hershey et al., 2016a; Isik et al., 2016; Yu et al., 2017b; Kolbaek et al., 2017; Chen et al., 2017; Chang et al., 2018a; Tan et al., 2018; Chen and Droppo, 2018; Qian et al., 2018; Chen et al., 2018; Chang et al., 2018b). The core issue that these techniques try to address is the label ambiguity or permutation problem (refer to Section 3 for details).

In Weng et al. (2015) a deep learning model was developed to recognize mixed speech directly. To solve the label ambiguity problem, Weng et al. assigned the senone labels of the talker with higher instantaneous energy to output one and the other to output two. Although this addresses the label ambiguity problem, it causes frequent speaker switch across frames. To deal with the speaker switching problem, a two-speaker joint-decoder with a speaker switching penalty was used to trace speakers. This approach has two limitations. First, energy, which is manually selected, may not be the best information to assign labels under all conditions. Second, the frame switching problem introduces burden to the decoder.

In Hershey et al. (2016a); Isik et al. (2016) the multi-talker mixed speech is first separated into multiple streams. An ASR engine is then applied to these streams independently to recognize speech. To separate the speech streams, they proposed a technique called deep clustering (DPCL) (Hershey et al., 2016a). They assume that each time-frequency bin belongs to only one speaker and can be mapped into a shared embedding space. The model is optimized so that in the embedding space the time-frequency bins belonging to the same speaker are closer and those of different speakers are farther away. The work in Isik et al. (2016) further introduced an enhancement network to refine the DPCL output, in which the soft mask can be obtained to achieve a better reconstruction performance. During evaluation, a clustering algorithm is first used upon embeddings to generate a partition of the time-frequency bins, and then the separated audio streams are reconstructed based on the partition. In these works, the speech separation and recognition are usually two separate components.

Chen et al. (2017) proposed a similar technique called deep attractor network (DANet). Following DPCL, their approach also learns a high-dimensional embedding of the acoustic signals. Different from DPCL, however, it creates cluster centers, called attractor points, in the embedding space to pull together the time-frequency bins corresponding to the same source. The main limitation of DANet is the requirement to estimate attractor points during evaluation time and to form frequency-bin clusters based on these points.

In Yu et al. (2017b), Kolbaek et al. (2017), Chang et al. (2018a), Tan et al. (2018), Chen and Droppo (2018), Qian et al. (2018), Chen et al. (2018) and Chang et al. (2018b), a simpler yet equally effective technique named permutation invariant training (PIT) was proposed to address the speaker independent multi-talker speech separation problem. In PIT, the source targets are treated as a set (i.e., order is irrelevant). During training, PIT first determines the output-target assignment with the minimum error at the utterance level based on the forward-pass result. It then minimizes the error given the assignment. This strategy elegantly solved the label permutation problem. However, in these original works PIT was used to separate speech streams from mixed speech. For this reason, a frequency-bin mask was first estimated and then used to reconstruct each stream. The minimum mean square error (MMSE) between the true and reconstructed speech streams was used as the criterion to optimize model parameters. It is noted that a similar permutation free technique was also proposed in Hershey et al. (2016a) but with negative results and conclusions; and this idea was also used in Isik et al. (2016) but within the DPCL framework which is more complex.

Moreover, most previous works on single-channel multi-talker speech still focus on *speech separation* (Hershey et al., 2016a; Isik et al., 2016; Chen et al., 2017; Yu et al., 2017b; Kolbaek et al., 2017). In contrast, single-channel multi-talker *speech recognition* is much harder and there is less related work. There have been some attempts, but the related tasks are relatively simple. For example, the 2006 monaural speech separation and recognition challenge (Cooke et al., 2010; Hershey et al., 2010; Rennie et al., 2010; Weng et al., 2015) was defined on a speaker-dependent, small vocabulary, constrained language model setup, while in Isik et al. (2016) a medium vocabulary reading style corpus was used. We are not aware of any extensive research work on the more real, speaker-independent, spontaneous large vocabulary continuous speech recognition (LVCSR) on single-channel multi-talker mixed speech before our work.

In this paper, we attack the multi-talker mixed speech recognition problem with a focus on the speaker-independent setup given just a single-channel of the mixed speech. Different from Hershey et al. (2016a), Isik et al. (2016), Yu et al. (2017b) and Kolbaek et al. (2017), here we extend and redefine PIT over log filter bank features and/or senone posteriors. In some architectures PIT is defined upon the minimum mean square error (MSE) between the true and estimated individual speaker features to separate speech at the feature level (called PIT-MSE from now on). In some other architectures, PIT is defined upon the cross entropy (CE) between the true and estimated senone posterior probabilities to recognize multiple streams of speech directly (called PIT-CE from now on). Moreover, the PIT-MSE based front-end feature separation can be combined with the PIT-CE based back-end recognition in a joint optimization architecture. We evaluate our architectures on the artificially generated AMI data with both two- and three-talker mixed speech. The experimental results demonstrate that our proposed architectures are very promising and flexible. Note that compared to our previous preliminary attempt in Yu et al. (2017a), this paper gives more comprehensive and detailed exploration, and other architectures are also developed and compared in this work: PIT is performed not only on the front-end feature separation module to obtain better separated feature streams but also on the back-end recognition module to predict the separated senone posterior probabilities directly. Moreover, PIT can be implemented on both the front-end and back-end with a joint-optimization architecture. Then a comprehensive experiment is designed and compared to evaluate the performance of different PIT based architectures for multi-talker speech recognition, and the further improvement and analysis is performed for the proposed framework with the evaluation on both artificially generated multi-talker AMI and WSJ0 corpus.

The rest of the paper is organized as follows. In Section 2 we describe the speaker independent single-channel multi-talker mixed speech recognition problem. In Section 3 we propose several PIT-based architectures to recognize multi-streams of speech. We report experimental results in Section 4 and conclude the paper in Section 5.

2. Single-channel multi-talker speech recognition

In this paper, we assume that a single-microphone signal $y[n]$ is observed. $y[n]$ is a mixture signal and assumed to be a linear combination of multiple speech sources, i.e. $y[n] = \sum_{s=1}^{S} x_s[n]$, where $x_s[n], s = 1, \cdots, S$ are S streams of speech sources from different speakers. Our goal is to separate these streams and recognize every single one of them. In other words, the model needs to generate S output streams, one for each source, at every time step. However, given only the mixed speech $y[n]$, the problem of recognizing all streams is under-determined because there are an infinite number of possible $x_s[n]$ (and thus recognition results) combinations that lead to the same $y[n]$. Fortunately, speech is not a random signal. It has patterns that we may learn from a training set of pairs $\mathbf{y}$ and $\ell^s, s = 1, \cdots, S$, where ℓ^s is the senone label sequence for stream s, i.e. the senone alignment obtain on the original single-talker speech.

In the single speaker case, i.e., $S = 1$, the learning problem is significantly simplified because there is only one speaker's stream that needs to be recognized, thus it can be cast as a simple supervised optimization problem. Given the input to the model, which is some feature representation of $\mathbf{y}$, the output is simply the senone posterior probability conditioned on the input. As in most classification problems, the model can be optimized by minimizing the cross entropy between the senone label and the estimated posterior probability.

When S is greater than 1, however, it is no longer as simple and direct as in the single-talker case and the label ambiguity or permutation becomes a problem in training. In the case of two speakers, because speech sources are symmetric given the mixture (i.e., $\mathbf{x}_1 + \mathbf{x}_2$ equals to $\mathbf{x}_2 + \mathbf{x}_1$ and both $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same characteristics), there is no predetermined way to assign the correct target to the corresponding output layer. Interested readers can find additional information in Yu et al. (2017b) and Kolbaek et al. (2017) on how training fails to progress when the conventional supervised approach is used for the multi-talker speech separation.

3. Permutation invariant training for multi-talker speech recognition

To address the label ambiguity problem, we propose several architectures based on the permutation invariant training (PIT) (Yu et al., 2017b; Kolbaek et al., 2017; Chang et al., 2018a; Tan et al., 2018; Chen and Droppo, 2018; Qian et al., 2018; Chen et al., 2018; Chang et al., 2018b) for single-channel multi-talker mixed speech recognition. For simplicity and without loss the generality, we always assume there are two-talkers in the mixed speech when describing our architectures in this section.

Note that, DPCL (Hershey et al., 2016a; Isik et al., 2016) and DANet (Chen et al., 2017) are alternative solutions to the label ambiguity problem. However these two techniques are not as straightforward as those in Yu et al. (2017b), Kolbaek et al. (2017), Chang et al. (2018a), Tan et al. (2018), Chen and Droppo (2018), Chen et al. (2018) and Chang et al. (2018b), and cannot be easily applied to direct recognition of multiple streams of speech without first separation in training or evaluation.[1] Furthermore the current goals of these works are still only on speech source separation.

3.1. Feature separation with direct supervision

To recognize the multi-talker mixed speech, one straightforward approach is to estimate the features of each speech source given the mixed speech feature and recognize them one by one using a normal single-talker LVCSR system. This idea is depicted in Fig. 1 where we learn a model to recover the filter bank (FBANK) features from the mixed FBANK features and then feed each stream of the recovered FBANK features to a conventional LVCSR system for recognition.

In the simplest architecture, which is denoted as **Arch#1** and illustrated in Fig. 1(a), feature separation can be considered as a multi-class regression problem, similar to many previous works (Huang et al., 2014; Wang et al., 2014; Xu et al., 2014; Weninger et al., 2015; Huang et al., 2015; Du et al., 2016). In this architecture, $\mathbf{Y}$, the mixed speech features, are used as the input to some deep learning models, such as deep neural networks (DNNs), convolutional neural networks (CNNs), and long short-term memory (LSTM) recurrent neural networks (RNNs), to estimate feature representation of each individual talker. If we use the bidirectional LSTM-RNN model, the model will compute

$$\mathbf{H}_0 = \mathbf{Y} \tag{1}$$

$$\mathbf{H}_i^f = RNN_i^f(\mathbf{H}_{i-1}), i = 1, \cdots, N \tag{2}$$

$$\mathbf{H}_i^b = RNN_i^b(\mathbf{H}_{i-1}), i = 1, \cdots, N \tag{3}$$

$$\mathbf{H_i} = Stack(H_i^f, H_i^b), i = 1, \cdots, N \tag{4}$$

$$\hat{\mathbf{X}}^s = Linear^s(\mathbf{H}_N), s = 1, \cdots, S \tag{5}$$

where $\mathbf{H}_0$ is the input, N is the number of hidden layers, $\mathbf{H}_i$ is the ith hidden layer, RNN_i^f and RNN_i^b are the forward and backward RNNs at hidden layer i, respectively, $\hat{\mathbf{X}}^s, s = 1, \cdots, S$ is the estimated separated features from the output layers for each speech stream s.

During training, we need to provide the correct reference (or target) features $\mathbf{X^s}, s = 1, \cdots, S$ for all speakers in the mixed speech to the corresponding output layers for supervision. The model parameters can be optimized to minimize the mean square error (MSE) between the estimated separated feature $\hat{\mathbf{X}}^\mathbf{s}$ and the original reference feature $\mathbf{X^s}$,

$$\mathbf{J} = \frac{1}{S}\sum_{s=1}^{S}\sum_{t} ||\mathbf{X_t^s} - \hat{\mathbf{X}}_\mathbf{t}^\mathbf{s}||^2 \tag{6}$$

where S is the number of mixed speakers. In this architecture, it is assumed that the reference features are organized in a given order and assigned to the output layer segments accordingly. Once trained, this feature separation module can be used as the front-end to process the mixed speech. The separated feature streams are then either fed into a normal single-speaker LVCSR system for decoding directly, or used to first retrain the original single-speaker acoustic model prior to decoding.

3.2. Feature separation with permutation invariant training

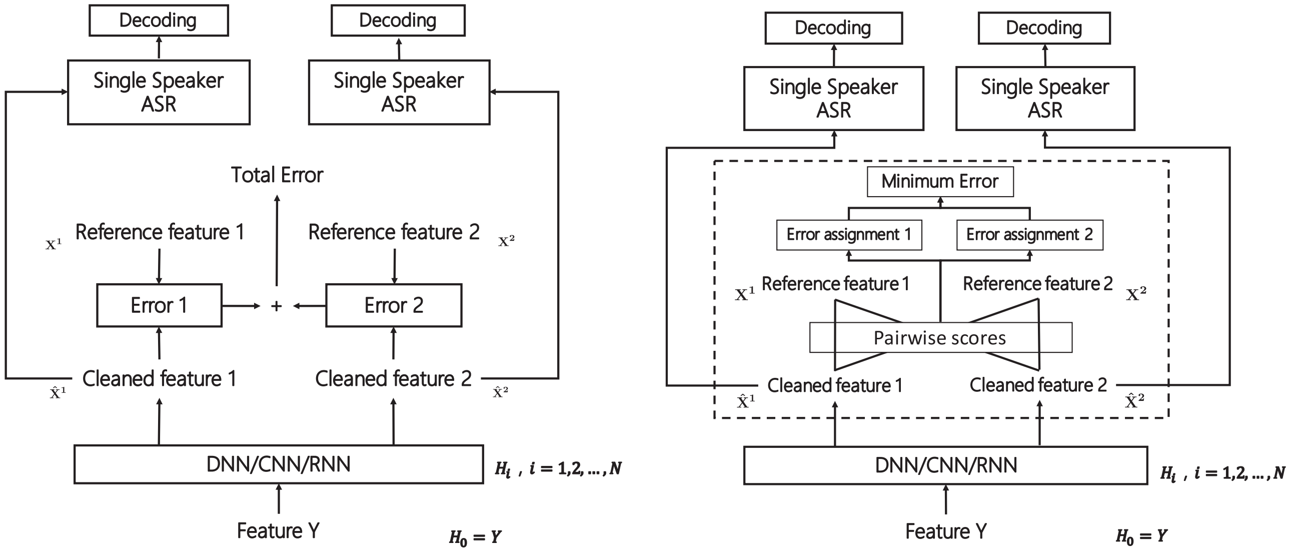
The architecture depicted in Fig. 1(a) is easy to implement but with obvious drawbacks. Since the model has multiple output layer segments (one for each stream), and they depend on the same input mixture, assigning a reference is actually difficult. The fixed reference order used in this architecture is not quite right since the source speech streams are symmetric and there is no clear clue on how to order them in advance. This is referred to as the label ambiguity (or label permutation) problem in Weng et al. (2015), Hershey et al. (2016a) and Yu et al. (2017b). As a result, this architecture may work well on the speaker-dependent setup where the target speaker is known (and thus can be assigned to a specific output segment) during training, but cannot generalize well to the speaker-independent case.

The label ambiguity problem in multi-talker mixed speech recognition was addressed with limited success in Weng et al. (2015) where Weng et al. assigned reference features depending on the energy level of each speech source. In the architecture illustrated in Fig. 1(b), named as **Arch#2**, permutation invariant training (PIT) (Kolbaek et al., 2017; Yu et al., 2017b) is utilized to estimate individual feature streams. In this architecture, the reference feature sources are given as a set instead of an ordered list. The output-reference assignment is determined dynamically based on the current model. More specifically, it first computes the MSE for each possible assignment between the reference $\mathbf{X}^{\mathbf{s}'}$ and the estimated source $\hat{\mathbf{X}}^\mathbf{s}$, and picks the one with minimum MSE. In other words, the training criterion is

$$J = \frac{1}{S}\min_{s'\in permu(S)}\sum_{s}\sum_{t} ||\mathbf{X}_\mathbf{t}^{\mathbf{s}'} - \hat{\mathbf{X}}_\mathbf{t}^\mathbf{s}||^2, s = 1, \cdots, S \tag{7}$$

where *permu*(S) is a permutation of 1, ⋯, S. We note two important ingredients in this objective function. First, it automatically finds the appropriate assignment no matter how the labels are ordered. Second, the MSE is computed over the whole sequence for each assignment. This forces all the frames of the same speaker to be aligned with the same output segment, which can be regarded as performing the feature-level tracing implicitly. With this new objective function, we can simultaneously perform label assignment and error evaluation on the feature level.

[1] Although it is possible to append a recognition module following the structure proposed in Isik et al. (2016) and train it jointly, it is still with the DPCL framework and a clustering stage is firstly demanded to generate the separated stream during the evaluation.



(a) Arch#1: Feature separation with the fixed reference assignment

(b) Arch#2: Feature separation with permutation invariant training

Fig. 1. Feature separation architectures for multi-talker mixed speech recognition.

It is expected that the feature streams separated with PIT (Fig. 1(b)) have higher quality than those separated with a fixed reference order (Fig. 1(a)). As a result, the recognition errors on these feature streams should also be lower. Note that the computational cost associated with permutation is negligible compared to the network forward computation during training, and no permutation (and thus no cost) is needed during evaluation.

Note that in this and other architectures that use PIT, we need to determine the maximum number of simultaneous speakers the system should support and thus the number of output sections in the model during design time. Our statistics has shown that in most cases supporting three simultaneous speakers is sufficient. It is very rare that four or more speakers speak at the same time and all of them have similar energy.

Here we describe several scenarios to better understand PIT. Let A, B and C be three different speakers, and A + B, B + C, and A + C are mixtures of A and B, B and C, and A and C, respectively. If the model only supports two simultaneous speakers (i.e., the number of output sections is two) and we input A + B and B + C to the system as two separate mixtures, B may not be assigned to the same output section (which is done automatically) when it's mixed with A and C. Which output section B will be assigned to depends not only on B but also on which other speaker it is mixed with and the internal state (determined by the history) of the system. This is similar to the "Rock-paper-scissors" game, whether Rock is assigned as the winner (i.e., left output section) depends on the other party. If a speaker (say A) is mixed with the same other speaker (say B), however, in most of the cases, they will be assigned to the same output sections (Say A to 1 and B to 2), and occasionally to different output sections, across utterances. Here using recurrent neural networks (RNNs) is critical because RNNs maintain an internal state that is dependent on the history and contains important information to separate and trace speakers.

Note that when the model only supports two simultaneous speakers we cannot input A + B + C (e.g., first half is A + B and second half is B + C) to the system during training time. When the model supports three simultaneous speakers, however, we can input a mixture, whose first half is A + B and second half is B + C, to the system. When this happens, B will be assigned to the same output section (say 1) over the whole utterance, while A and C are usually assigned to two different output sections (say 2 and 3). The system will be trained to assign B and C to different output sections because the target has three streams.

3.3. Direct multi-talker mixed speech recognition with PIT

In the previous two architectures mixed speech features are first separated explicitly and then recognized independently with a conventional single-talker LVCSR system. Since the feature separation is not perfect, there is mismatch between the separated features and the normal features used to train the conventional LVCSR system. In addition, the objective function minimizing the MSE between the estimated and reference features is less correlated to the recognition performance than the objective functions which minimize the difference between the estimated senone posteriors and the corresponding labels, such as cross-entropy. In this section, we propose an end-to-end architecture that directly recognizes mixed speech of multiple speakers.[2]

In this architecture, denoted as **Arch#3**, we apply PIT to the CE between the reference and estimated senone posterior probability distributions as shown in Fig. 2(a). Given some feature representation **Y** of the mixed speech **y**, this model will compute

$$\mathbf{H}_0 = \mathbf{Y} \tag{8}$$

$$\mathbf{H}_i^f = RNN_i^f(\mathbf{H}_{i-1}), i = 1, \cdots, N \tag{9}$$

$$\mathbf{H}_i^b = RNN_i^b(\mathbf{H}_{i-1}), i = 1, \cdots, N \tag{10}$$

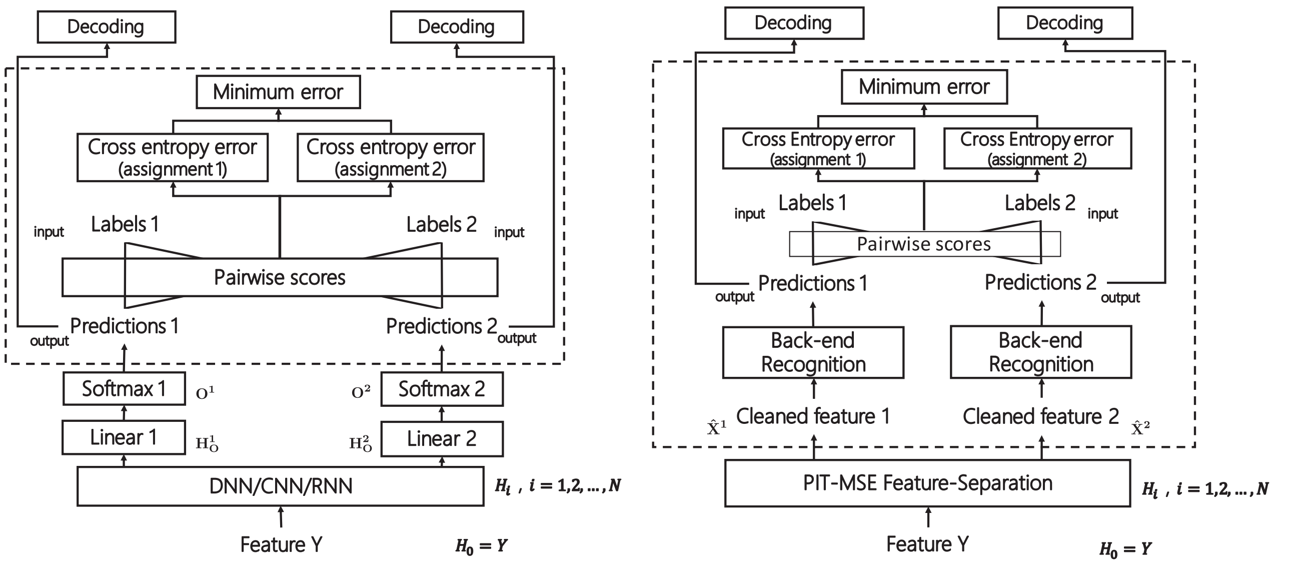
$$\mathbf{H}_i = Stack(\mathbf{H}_i^f, \mathbf{H}_i^b), i = 1, \cdots, N \tag{11}$$

$$\mathbf{H}_o^s = Linear^s(\mathbf{H}_N), s = 1, \cdots, S \tag{12}$$

$$\mathbf{O}^s = Softmax(\mathbf{H}_o^s), s = 1, \cdots, S \tag{13}$$

using a deep bidirectional RNN, where Equations (8) ~ (11) are similar to Eqs. (1) ~ (4). $\mathbf{H}_o^s, s = 1, \cdots, S$ is the excitation at the output layer

[2] Note that the 'end-to-end' term used here does not refer to end-to-end ASR architectures such as attention or CTC. In our work, it indicates the proposed architecture could do the multi-talker speech recognition without explicit speech separation stage, and it can recognize the individual mixed speech stream directly.



(a) Arch#3: Direct multi-talker mixed speech recognition with PIT

(b) Arch#4: Joint optimization of PIT-based feature separation and recognition

Fig. 2. Advanced architectures for multi-talker mixed speech recognition.

for each speech stream s, and $\mathbf{O}^s, s = 1, \cdots, S$ is the output segment for stream s. Note that the parameters of Linear blocks for each stream in Arch#3 are not shared. In contrast to the architectures discussed in previous sections, in this architecture each output segment represents the estimated senone posterior probability for a speech stream. No additional feature separation, clustering or speaker tracing is needed. Although various neural network structures can be used, in this study we focus on bidirectional LSTM-RNNs.

In this direct multi-talker mixed speech recognition architecture, we minimize the objective function

$$J = \frac{1}{S} \min_{s' \in permu(S)} \sum_{s} \sum_{t} CE(\ell_t^{s'}, \mathbf{O}_t^s), s = 1, \cdots, S \tag{14}$$

where $\ell_t^{s'}$ is the senone label for stream s' at the tth timestep.

In other words, we minimize the minimum average CE of every possible output-label assignment. All the frames of the same speaker are forced to be aligned with the same output segment by computing the CE over the whole sequence for each assignment. This strategy allows for the direct multi-talker mixed speech recognition without explicit separation. It is a simpler and more compact architecture for multi-talker speech recognition.

3.4. Joint optimization of PIT-based feature separation and recognition

As mentioned above, the main drawback of the feature separation architectures is the mismatch between the distorted separation result and the features used to train the single-talker LVCSR system. The direct multi-talker mixed speech recognition with PIT, which bypassed the feature separation step, is one solution to this problem. Here we propose another architecture named joint optimization of PIT-based feature separation and recognition, and it is denoted as **Arch#4** and shown in Fig. 2(b).

This architecture contains two PIT-components, the front-end feature separation module with PIT-MSE and the back-end recognition module with PIT-CE. Different from the architecture in Fig. 1(b), in this architecture a new LVCSR system is trained on the output of the feature separation module with PIT-CE. The whole model is trained progressively: the front-end feature separation module is firstly optimized with PIT-MSE; Then the parameters in the back-end recognition module are optimized with PIT-CE while keeping the parameters in the feature separation module fixed. Finally parameters in both modules are jointly refined with PIT-CE using a small learning rate. This progressive training strategy seems to be important in our study. We have tried training jointly without progressive training and it did not converge well. Note that the reference assignment in the recognition (PIT-CE) step is the same as that in the separation (PIT-MSE) step, and the parameters of two back-end recognition blocks in Arch#4 are shared.

$$J_1 = \frac{1}{S} \min_{s' \in permu(S)} \sum_{s} \sum_{t} ||\mathbf{X}_t^{s'} - \hat{\mathbf{X}}_t^{s}||^2, s = 1, \cdots, S \tag{15}$$

$$J_2 = \frac{1}{S} \min_{s' \in permu(S)} \sum_{s} \sum_{t} CE\left(\ell_t^{s'}, \mathbf{O}_t^s\right), s = 1, \cdots, S \tag{16}$$

During decoding, the mixed speech features are fed into this architecture, and the final posterior streams are used for decoding as normal.

4. Experimental results

To evaluate the performance of the proposed architectures, we conducted a series of experiments on artificially generated two- and three-talker mixed speech datasets based on the AMI corpus (Hain et al., 2012).

There are four reasons for us to use AMI: (1) AMI is a speaker-independent spontaneous LVCSR corpora. Compared to small vocabulary, speaker-dependent, read English datasets used in most of the previous studies (Cooke et al., 2010; Hershey et al., 2010; Rennie et al., 2010; Weng et al., 2015), observations made and conclusions drawn from AMI are more likely to generalize to other real-world scenarios; (2) AMI is a really hard task with a variety of noises (mainly in AMI SDM/MDM), truly spontaneous meeting style speech, and strong accents. It reflects the true ability of LVCSR when the training set size is around 100 hours. The state-of-the-art word error rate (WER) on AMI is around 25.0% for the close-talk condition (Povey et al., 2016) and more than 45.0% for the far-field condition with single-microphone (Povey et al., 2016; Zhang et al., 2016b). These WERs are much higher than those on other corpora, such as Switchboard (Godfrey and Holliman, 1997) on which the WER is now below 10.0% (Povey et al., 2016; Sercu and Goel, 2016; Saon et al., 2016; Xiong et al., 2017); (3) Although the close-talk data (AMI IHM) was used to generate mixed speech in this work due to the high WER in the far-field condition, the existence of parallel far-field data (AMI SDM/MDM) allows us to evaluate our architectures based on the far-field data in the future when the single-talker ASR results on it is improved; 4) AMI is a public corpora, therefore, using AMI allows interested readers to reproduce our results more easily.[3]

The AMI IHM (close-talk) dataset contains about 80 h and 8 h speech in training and evaluation sets, respectively (Hain et al., 2012; Swietojanski et al., 2013). Using AMI IHM, we generated a two-talker (IHM-2mix) and a three-talker (IHM-3mix) mixed speech dataset.

To artificially synthesize IHM-2mix, we randomly select two speakers and then randomly select an utterance for each speaker to form a mixed-speech utterance. For easier explanation, the high energy (High E) speaker in the mixed speech is always chosen as the target speaker and the low energy (Low E) speaker is considered as interference speaker. We synthesized mixed speech for five different SNR conditions (i.e. 0 dB, 5 dB, 10 dB, 15 dB, 20 dB) based on the utterance-level energy ratio of the two talkers. More specifically, the energy is calculated for each utterance firstly, and then the selected utterances will be scaled before utterance mixing to satisfy the target SNR. To eliminate easy cases we force the lengths of the selected source utterances to be comparable so that at least half of the mixed speech contains overlapping speech. When the two source utterances have different lengths, the shorter one is padded with random noise with small value at the front and end. The same procedure is used for preparing both the training and testing data. We generated in total 400 h two-talker mixed speech, 80 h per SNR condition, as the training set. A subset of 80 h speech from this 400 h training set was used for fast model training and evaluation. For evaluation, total 40 h two-talker mixed speech, 8 h per SNR condition, is generated and used.

The IHM-3mix dataset was generated similarly. The relative energy of the three speakers in each mixed utterance varies randomly in the training set. Different from the training set, all the speakers in the same mixed utterance have equal energy in the testing set. We generated in total 400 h and 8 h three-talker mixed speech as the training and testing set, respectively.

4.1. Single-speaker recognition baseline

In this work, all the neural networks were built using the latest Microsoft Cognitive Toolkit (CNTK) (Yu et al., 2014) and the decoding systems were built based on Kaldi (Povey et al., 2011). We first followed the officially released Kaldi AMI IHM recipe to build an LDA-MLLT-SAT GMM-HMM model. This model uses 39-dimensional MFCC feature and has roughly 4K tied-states and 80K Gaussians. We then used this acoustic model to generate the senone alignment for neural network training. We trained the DNN and BLSTM-RNN baseline systems with the original AMI IHM data. 40-dimensional log filter bank (LFBK) features with CMVN were used to train the baselines. The DNN has 6 hidden layers each of which contains 2048 Sigmoid neurons. The input feature for DNN contains a window of 11 frames. The BLSTM-RNN has 3 bidirectional LSTM layers which are followed by the softmax layer. Each BLSTM layer has 768 memory cells. The input to the BLSTM-RNN is a single acoustic frame. All the models explored here are optimized with cross-entropy criterion. The DNN is optimized using SGD method with 256 minibatch size, and the BLSTM-RNN is trained using SGD with 4 full-length utterances in each minibatch.

For decoding, we used a 50K-word dictionary and a trigram language model interpolated from the ones created using the AMI transcripts and the Fisher English corpus. The performance of these two baselines on the original single-speaker AMI corpus are presented in Table 1. These results are comparable with those reported by others (Swietojanski et al., 2013) even though the systems did not use adapted fMLLR features.[4] It is noted that adding more BLSTM layers did not show meaningful WER reduction in the baseline.

Table 1
WER (%) of the baseline systems on original AMI IHM single-talker corpus.

Model	WER
DNN	28.0
BLSTM	27.0

Table 2
WER (%) of the baseline BLSTM-RNN single-speaker system on the IHM-2mix dataset.

SNR condition	High E WER	Low E WER
0 dB	85.0	100.5
5 dB	68.8	110.2
10 dB	51.9	114.9
15 dB	39.3	117.6
20 dB	32.1	118.7

To test the normal single-speaker model on the two-talker mixed speech, the above baseline BLSTM-RNN model is utilized to decode the mixed speech directly. During scoring we compare the decoding output (only one output) against the reference of each source utterance, i.e. high energy speaker utterance and low energy speaker utterance, to obtain the individual WER for the mixed two speakers. Table 2 summarizes the recognition results. It is clear, from the table, that the single-speaker model performs very poorly on the multi-talker mixed speech as indicated by the huge WER degradation of the high-energy speaker when SNR decreases. Further more, in all the conditions, the WERs for the low energy speaker are all above 100.0%. These results demonstrate the great challenge in the multi-talker mixed speech recognition.

4.2. Evaluation of two-talker speech recognition architectures

The proposed four architectures for two-talker speech recognition are evaluated here. For the first two approaches (Arch#1 and Arch#2) that contain an explicit feature separation stage (with and without PIT-MSE), a 3-layer BLSTM is used in the feature separation module. The separated feature streams are either (1) fed into a normal 3-layer BLSTM LVCSR system, trained with original single-talker speech, for decoding; or (2) used to first retrain the original single-speaker acoustic model and then fed into this updated acoustic model for decoding. The whole system contains in total six BLSTM layers. For the other two approaches (Arch#3 and Arch#4), in which PIT-CE is used, 6-layer BLSTM models are used so that the number of parameters is comparable to the other two architectures. In all these architectures the input is the 40-dimensional LFBK feature and each layer contains 768 memory cells. To train the latter two architectures that exploit PIT-CE we need to prepare the alignments for the mixed speech. The senone alignments for the two-talkers in each mixed speech utterance are from the single-speaker baseline alignment (using original clean single-speaker AMI corpus). The alignment of the shorter utterance within the mixed speech is padded with the silence state at the front and the end. All the models were trained with a minibatch of 8 utterances. The gradient was clipped to 0.0003 per value to guarantee the training stability. To obtain the results reported in this section we used the 80 h mixed speech training subset.

The recognition results on both speakers are evaluated. For scoring, we evaluated the two hypotheses, obtained from two output sections, against the two references and pick the assignment with better WER to compute the final WER. Note that all results reported in this paper are trained with CE criterion instead of sequence discriminative training since our main focus of this work is to evaluate the techniques for multi-talker speech recognition and CE training is sufficient for us to evaluate these new techniques. We will leave sequence discriminative training,

[3] The generated multi-talker mixed AMI dataset can be accessed and downloaded from: http://speechlab.sjtu.edu.cn/resources/data/ami-2mix

[4] Since our multi-talker models used only FBANK features, we show the baseline results on FBANK in this table for clarity. With fMLLR features, the WERs are 25.9% and 25.0% for DNN and BLSTM systems, respectively, on this original AMI IHM single-talker corpus.

Table 3
WER (%) of the proposed multi-talker mixed speech recognition architectures on the IHM-2mix dataset under 0 dB SNR condition (using 80 h training subset). Arch#1-#4 indicate the proposed architectures described in Sections 3.1 to 3.4, respectively.

Arch	Front-end	Back-end	High E WER	Low E WER
#1	Feat-Sep-baseline	Single-Spk-ASR	72.58	79.61
		+ re-train	65.64	73.47
#2	Feat-Sep-PIT-MSE	Single-Spk-ASR	68.88	75.62
		+ re-train	60.53	67.22
#3	×	PIT-CE	56.04	63.87
#4	Feat-Sep-PIT-MSE	PIT-CE	55.08	63.31

Table 4
WER (%) of the proposed direct PIT-CE-ASR model on the IHM-2mix dataset with full training set.

SNR condition	High E WER	Low E WER
0 dB	47.77	54.89
5 dB	39.25	59.24
10 dB	33.83	64.14
15 dB	30.54	71.75
20 dB	28.75	79.88
original clean AMI IHM single-talker eval	27.0	SILENCE

Table 5
WER (%) of the direct PIT-CE-ASR model using different deep learning models on the IHM-2mix dataset.

Models	**SNR condition**	High E WER	Low E WER
6L-DNN	0 dB	72.95	80.29
	5 dB	65.42	84.44
	10 dB	55.27	86.55
	15 dB	47.12	89.21
	20 dB	40.31	92.45
4L-BLSTM	0 dB	49.74	56.88
	5 dB	40.31	60.31
	10 dB	34.38	65.52
	15 dB	31.24	73.04
	20 dB	29.68	80.83
6L-BLSTM	0 dB	47.77	54.89
	5 dB	39.25	59.24
	10 dB	33.83	64.14
	15 dB	30.54	71.75
	20 dB	28.75	79.88
8L-BLSTM	0 dB	46.91	53.89
	5 dB	39.14	59.00
	10 dB	33.47	63.91
	15 dB	30.09	71.14
	20 dB	28.61	79.34

which can be carried out by using the assignments determined by PIT-CE, as a future work item when we want to achieve the best results on the task.

The results on the 0 dB SNR condition are shown in Table 3. Compared to the 0 dB condition in Table 2, all the proposed multi-talker speech recognition architectures obtain obvious improvement on both speakers. Within the two architectures with the explicit feature separation stage, the architecture with PIT-MSE is obviously better than the baseline feature separation architecture. These results confirmed that the label permutation problem can be alleviated by the PIT-MSE at the feature level. It is further observed that if we use the separated features to retrain the single-speaker model, the system performance can be improved significantly, and the gain from the retraining in Arch#2 is larger than that in Arch#1. Moreover applying PIT-CE on the recognition module (Arch#3 & Arch#4) can further reduce WER by 5.0% ~ 10.0% absolute. This is because these two architectures can significantly reduce the mismatch between the separated feature and the feature used to train the LVCSR model, and it is also because the optimization criterion (PIT-CE) is better correlated to the recognition accuracy. Comparing Arch#3 and Arch#4, we can see that the architecture with joint optimization on PIT-based feature separation and recognition slightly outperforms the direct PIT-CE based model.

Since Arch#3 and Arch#4 achieve comparable results, and the model architecture and training process of Arch#3 is much simpler than that of Arch#4, our further evaluations reported in the following sections are based on Arch#3. For clarity, Arch#3 is named **direct PIT-CE-ASR** from now on.

4.3. Evaluation of the direct PIT-CE-ASR model on large dataset

We evaluated the direct PIT-CE-ASR architecture on the full IHM-2mix corpus. All the 400 h mixed data under different SNR conditions are pooled together for training. The direct PIT-CE-ASR model is still composed of 6 BLSTM layers with 768 memory cells in each layer. All other configurations are also the same as the experiments conducted on the subset.

The results under different SNR conditions are shown in the top part of Table 4. Comparing to the results in Table 3 under the 0 dB SNR condition, achieved with 80 h training subset, we observe that additional absolute ~ 10.0% WER improvement on both speakers can be obtained using the large training set. Compared to the baseline results in Table 2, the direct PIT-CE-ASR model achieved significant improvements on both talkers for all SNR conditions. We also observe that the WER increases slowly when the SNR becomes smaller for the high energy speaker, and the WER improvement is very significant for the low energy speaker across all conditions. In the 0 dB SNR scenario, the WERs on two speakers are very close and are 45.0% less than that achieved with the single-talker ASR system for both high and low energy speakers. At 20 dB SNR, the WER of the high energy speaker is still significantly better than the baseline, and approaches the single-talker recognition result reported in Table 1.

Moreover, we evaluated our proposed direct PIT-CE-ASR system on the original clean single-talker utterances to see whether or not there is any performance degradation. To our surprise, the results, shown in the last row of Table 4, indicate that there is no performance degradation when using the proposed direct PIT-CE-ASR multi-talker system to recognize the original single-talker utterances, although the multi-talker model was trained only on multi-talker speech. More interestingly, one out of the two output streams almost always outputs SILENCE when the proposed direct PIT-CE-ASR system is used to recognize the single-talker speech.

4.4. Permutation invariant training with alternative deep learning models

We investigated the direct PIT-CE-ASR model with alternative deep learning models. The first model we evaluated is a 6-layer feed-forward DNN in which each layer contains 2048 Sigmoid units. The input to the DNN is a window of 11 frames each with a 40-dimensional LFBK feature.

The results of DNN-based PIT-CE-ASR model are reported at the top of Table 5. Although it still gets obvious improvement at the lower SNR conditions over the baseline single-speaker model shown in Table 2, the gain is much smaller with near 20.0% WER degradation in every condition relative to the BLSTM-based PIT-CE-ASR model. The difference between DNN and BLSTM models can be partially attributed to the stronger modeling power of BLSTM models and partially attributed to the better tracing ability of RNNs.

We also compared the BLSTM models with 4, 6, and 8 layers as shown in Table 5. It is observed that deeper BLSTM models perform better. This is different from the single speaker ASR model whose performance peaks at 4 BLSTM layers (Zhang et al., 2016b). This may be because the direct PIT-CE-ASR architecture needs to conduct two tasks - separation and recognition, and thus requires additional modeling power.

```
BLSTM Baseline-speaker1 (#C #S #D #I) 6 5 0 6
REF:  WELL IT WILL BE  YOU    KNOW STILL **** ** ** ** LIMITED VERSION OF YOU          KNOW THE NEXT QUERY SEARCH
HYP:  WELL I  DO   NOT REALLY KNOW STILL WITH R. S. I. ISSUE   AGAIN   A  PRESENTATION ON   THE NEXT ***** OR
EVAL:      S  S    S   S                 I    I  I  I  S       S       S  S            S             D     S
BLSTM Baseline-speaker2 (#C #S #D #I) 6 9 1 3
REF:  **** I AM CONCERNED WHEN YOU    READ THE   THE  R. S. I. ISSUE AGAIN * ************ REPETITIVE STRAIN INJURY
HYP:  WELL I ** DO        NOT  REALLY KNOW STILL WITH R. S. I. ISSUE AGAIN A PRESENTATION ON         THE    NEXT
EVAL: I      D  S         S    S      S    S     S                         I I            S          S      S
```

Fig. 3. Decoding results of baseline single speaker BLSTM-RNN system on a 0 dB two-talker mixed speech sample.

```
PIT model output2-speaker1 (#C #S #D #I) 12 3 1 0
REF:  WELL IT WILL BE YOU KNOW STILL LIMITED VERSION OF YOU KNOW THE NEXT QUERY SEARCH
HYP:  WELL I  WILL BE YOU WILL STILL WANTED  VERSION OF YOU KNOW THE NEXT QUERY ******
EVAL:      S              S          S                                          D
PIT model output1-speaker2 (#C #S #D #I) 10 5 1 0
REF:  I   AM  CONCERNED WHEN YOU READ THE  THE R. S. I. ISSUE AGAIN REPETITIVE STRAIN INJURY
HYP:  YOU CAN STILL     HAVE YOU **** HAVE THE R. S. I. ISSUE AGAIN REPETITIVE STRAIN INJURY
EVAL: S   S   S         S        D    S
```

Fig. 4. Decoding results of the proposed direct PIT-CE-ASR model on a 0 dB two-talker mixed speech sample.

Table 6
WER (%) comparison of the 6-layer-BLSTM direct PIT-CE-ASR model on the mixed speech generated from two male speakers (**M** + **M**), two female speakers (**F** + **F**) and a male and a female speaker (**M** + **F**).

Genders	SNR condition	High E WER	Low E WER
M + M	0 dB	52.18	59.32
	5 dB	42.64	61.77
	10 dB	36.10	63.94
F + F	0 dB	49.90	57.59
	5 dB	40.02	60.92
	10 dB	32.47	65.15
M + F	0 dB	44.89	51.72
	5 dB	37.34	57.43
	10 dB	33.22	63.86

4.5. Analysis on multi-talker speech recognition results

To better understand the results on multi-talker speech recognition, we computed the WER separately for the speech mixed with same and opposite genders. The results are shown in Table 6. It is observed that the same-gender mixed speech is much more difficult to recognize than the opposite-gender mixed speech, which is consistent with similar observations made for speech separation (Wang et al., 2017), and the gap is even larger when the energy ratio of the two speakers is closer to 1. It is also observed that the mixed speech of two male speakers is harder to recognize than that of two female speakers. These results suggest that effective exploitation of gender information may help to further improve the multi-talker speech recognition system. We will explore this in our future work.

We further examined the recognition results with and without using the direct PIT-CE-ASR. An example of these results on a 0 dB two-talker mixed speech utterance is shown in Fig. 3 (using the single-speaker baseline system) and Fig. 4 (with direct PIT-CE-ASR). It is clearly seen that the results are erroneous when the single-speaker baseline system is used to recognize the two-talker mixed speech. In contrast, many more words are recognized correctly with the proposed direct PIT-CE-ASR model.

4.6. Three-talker speech recognition with direct PIT-CE-ASR

In this subsection, we further extend and evaluate the proposed direct PIT-CE-ASR model on the three-talker mixed speech using the IHM-3mix dataset.

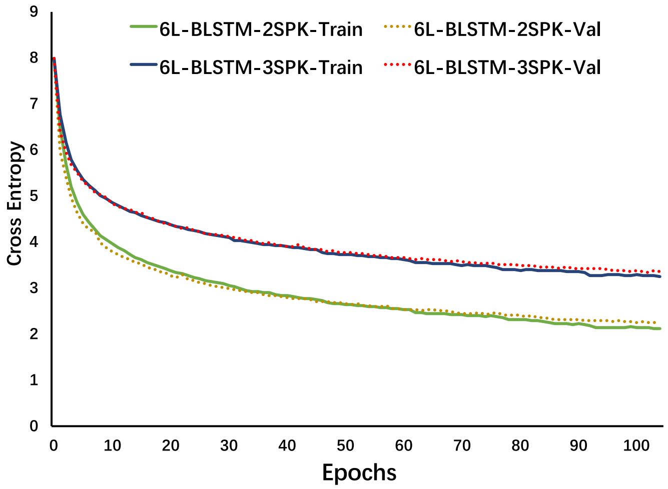


Fig. 5. CE values over epochs on both the IHM-2mix and IHM-3mix training and validation sets with the proposed direct PIT-CE-ASR model.

The three-talker direct PIT-CE-ASR model is also a 6-layer BLSTM model. The training and testing configurations are the same as those for two-talker speech recognition. The direct PIT-CE-ASR training processes as measured by CE on both two- and three-talker mixed speech training and validation sets are illustrated in Fig. 5. It is observed that the direct PIT-CE-ASR model with this specific configuration converges slowly, and the CE improvement progress on the training and validation sets is almost the same. The training progress on three-talker mixed speech is similar to that on two-talker mixed speech, but with an obviously higher CE value. This indicates the huge challenge when recognizing speech mixed with more than two talkers. Note that, in this set of experiments we used the same model configuration as that used in two-talker mixed speech recognition. Since three-talker mixed speech recognition is much harder, using deeper and wider models may help to improve performance. Due to resource limitation, we did not search for the best configuration for the task.

For scoring, the hypotheses from three output sections are evaluated against the three references for all assignments on each mixed-speech utterance, the best WER is used as the final WER. The three-talker mixed speech recognition WERs are reported in Table 7. The WERs on differ-

Table 7
WER (%) comparison of the baseline single-speaker BLSTM-RNN system and the proposed direct PIT-CE-ASR model on the IHM-3mix dataset. **Different** indicates the mixed speech is from different genders, and **Same** indicates the mixed speech is from same gender. Note that Speaker 1, 2 and 3 are arbitrarily numbered in this experiment.

Genders	Model	Speaker1	Speaker2	Speaker3
All	BLSTM-RNN	91.0	90.5	90.8
All	direct PIT-CE-ASR	69.54	67.35	66.01
Different		69.36	65.84	64.80
Same		72.21	70.11	69.78

Table 8
WER (%) when using three-talker direct PIT-CE-ASR model to recognize two-talker mixed IHM-2mix speech and original clean AMI IHM single-talker eval speech.

Model	**SNR condition**	High E WER	Low E WER
Three-Talker PIT-CE-ASR	0 dB	46.63	54.59
	5 dB	39.47	59.78
	10 dB	34.50	64.55
	15 dB	32.03	72.88
	20 dB	30.66	81.63
	single-talker	28.2%	SILENCE
Two-Talker PIT-CE-ASR	0 dB	47.77	54.89
	5 dB	39.25	59.24
	10 dB	33.83	64.14
(copied from Table IV)	15 dB	30.54	71.75
	20 dB	28.75	79.88
Single-talker baseline (Table I)	single-talker	27.0%	—

ent gender combinations are also provided. The WERs achieved with the single-speaker model are listed at the first line in Table 7. Compared to the results on IHM-2mix, the results on IHM-3mix are significantly worse using the conventional single speaker model. Under this extremely hard setup, the proposed direct PIT-CE-ASR architecture still demonstrated its powerful ability to separate, trace and recognize the mixed speech, and achieved 25.0% relative WER reduction across all three speakers. Although the performance gap from two-talker to three-talker is obvious, it is still very promising under this speaker-independent three-talker LVCSR task. Not surprisingly, the mixed speech of different genders is relatively easier to recognize than that of same gender, and the performance gap is about absolute 4.0% ~ 5.0% on WER.

We further used the three-talker direct PIT-CE-ASR model to recognize both the two-talker mixed speech and the single-talker clean speech. For (1) ***3-talker-model#2-talker-speech-test*** condition, the best two of these three output sections with the better assignment against the two speakers' references is used to compute the final WER for each mixed-speech utterance in scoring, and for (2) ***3-talker-model#1-talker-speech-test*** condition, the hypothesis-reference pair with the best WER is made as the final WER in scoring. The results are shown in the top part of Table 8. Surprisingly, the results are almost identical to that obtained using the 6-layer BLSTM based two-talker PIT-CE-ASR model (shown in the middle part of Table 8, which are copied from Table 4 for easy comparison) and slightly worse than the single-talker baseline system (shown as the last line of Table 8, which are copied from Table 1 for easy comparison).

Similar to what is observed in the experiments using the two-talker model to recognize the single-talker speech, the remaining output streams almost always produce SILENCE when the proposed three-talker PIT-CE-ASR model is used to recognize the single- and two-talker speech. Its robustness over variable number of mixed speakers suggests that a single PIT model may be able to recognize mixed speech of different number of speakers without knowing or estimating the number of speakers.

Table 9
WER (%) comparison of the proposed direct PIT-CE-ASR model and the DPCL model on the two-talker WSJ0 corpus.

Model	AVG WER
direct PIT-CE-ASR	28.2
DPCL (Isik et al., 2016)	30.8

4.7. Comparison with DPCL

For the better comparison with other techniques such as DPCL, we evaluated the direct PIT-CE-ASR model on the two-talker mixed WSJ0 corpus, which was used to evaluate DPCL (Hershey et al., 2016a; Isik et al., 2016) and released by MERL (Hershey et al., 2016b) with detailed information in Hershey et al. (2016a). The training procedure of the proposed direct PIT-CE-ASR model is the same as that used in the AMI IHM setup. The results are shown in Table 9, in which the DPCL result is copied from Isik et al. (2016). We observe that the proposed direct PIT-CE-ASR performs well even though it is also much simpler. It is noted that the direct comparison between these two results may be not very fair due to many differences for the implementation details, e.g. the work in Isik et al. (2016) used GMM-HMM for the acoustic model, and dropout, curriculum learning and multi-speaker training for the neural nets training on the separation module, but the results could still provide the useful reference for the interested readers. Moreover, the construction of the proposed direct PIT-CE-ASR is more simplified than using DPCL for multi-talker speech recognition (Hershey et al., 2016a; Isik et al., 2016), and the integration with other advanced techniques in speech recognition will be more straightforward. For example, it can be easily combined with the adaptation and sequence discriminative training techniques as the normal acoustic model to further improve performance.

We noted that the recently work in Settle et al. (2018) also uses DPCL with an end-to-end framework, and achieves promising results for multi-talker speech recognition. It utilizes a different setup, such as the different corpus and evaluation metric, so we can not compare with it directly here. We will leave this comparison in our future work.

5. Conclusion

In this paper, we proposed several architectures for recognizing multi-talker mixed speech given only a single channel of the mixed signal. Our technique is based on permutation invariant training, which was originally developed for separation of multiple speech streams. PIT can be performed on the front-end feature separation module to obtain better separated feature streams or it can be extended on the back-end recognition module to predict the separated senone posterior probabilities directly. Moreover, PIT can be implemented on both the front-end and back-end with a joint-optimization architecture. When using PIT to optimize a model, the criterion is computed over all frames in the whole utterance for each possible output-target assignment, and the one with the minimum loss is picked for parameter optimization. Thus PIT can address the label permutation problem well, and conduct the speaker separation and tracing in one shot. Particularly for the proposed architecture with the direct PIT-CE based recognition model, multi-talker mixed speech recognition can be directly conducted without an explicit separation stage.

The proposed architectures were evaluated and compared on an artificially mixed AMI dataset with both two- and three-talker mixed speech. The experimental results indicate that the proposed architectures are very promising. Our models can obtain relative 45.0% and 25.0% WER reduction against the state-of-the-art single-talker speech recognition system across all speakers when their energies are comparable, for two- and three-talker mixed speech, respectively. Another interesting observation is that there is either no degradation or only slight degradation when using the proposed three-talker model to recognize the single- and two-talker speech directly. This suggests that we can construct one model to recognize speech mixed with a variable number of speakers without knowing or estimating the number of speakers in the mixed speech. To our knowledge, this is the first work on the single-channel multi-talker mixed speech recognition on the challenging speaker-independent spontaneous LVCSR task.

In this work, the multi-talker mixed data is still generated artificially and based on the close-talk utterances, which relatively have the good quality. In contrast, the scenario is much more complex for the real applications, which are usually with a far-field situation and corrupted with more serious noises, e.g. reverberation. It will make the multi-talker speech recognition more challenging than the simulated case in this paper. There are still many work to be explored on this research area, and we will leave the evaluation on the real distant multi-talker speech in our future work.

Acknowledgment

This work was supported by the China NSFC projects (No. 61603252 and No. U1736202), the Shanghai Sailing Program No. 16YF1405300, and the Tencent-Shanghai Jiao Tong University joint project. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., Yu, D., 2014. Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 22, 1533–1545.

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Penn, G., 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4277–4280.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., et al., 2016. Deep speech 2: End-to-end speech recognition in English and Mandarin. In: International Conference on Machine Learning (ICML).

Bi, M., Qian, Y., Yu, K., 2015. Very deep convolutional neural networks for LVCSR. In: Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 3259–3263.

Chang, X., Qian, Y., Yu, D., 2018a. Adaptive permutation invariant training with auxiliary information for monaural multi-talker speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Chang, X., Qian, Y., Yu, D., 2018b. Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks. In: Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 1586–1590.

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

Chen, Z., Droppo, J., Li, J., Xiong, W., 2018. Progressive joint modeling in unsupervised single-channel overlapped speech recognition. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 26 (1), 184–196.

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

Cooke, M., Hershey, J.R., Rennie, S.J., 2010. Monaural speech separation and recognition challenge. Comput. Speech Lang. (CSL) 24, 1–15.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 20, 30–42.

Du, J., Tu, Y., Dai, L.R., Lee, C.H., 2016. A regression approach to single-channel speech separation via high-resolution deep neural networks. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 24, 1424–1437.

Godfrey, J. J., Holliman, E., 1997. Switchboard-1 release 2. Linguistic Data Consortium, Philadelphia.

Hain, T., Burget, L., Dines, J., Garner, P.N., Grézl, F., Hannani, A.E., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., 2012. Transcribing meetings with the AMIDA systems. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 20 (2), 486–498.

Hershey, J.R., Chen, Z., Roux, J.L., Watanabe, S., 2016a. Deep clustering: discriminative embeddings for segmentation and separation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 31–35.

Hershey, J.R., Rennie, S.J., Olsen, P.A., Kristjansson, T.T., 2010. Super-human multi-talker speech recognition: a graphical modeling approach. Comput. Speech Lang. (CSL) 24 (1), 45–66.

Hershey, J. R., Roux, J. L., Watanabe, S., Harsham, B., Isik, Y., Chen, Z., 2016b. Single-channel multi-speaker separation using deep clustering. http://www.merl.com/demos/deep-clustering

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. (SPM) 29, 82–97.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2014. Deep learning for monaural speech separation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1562–1566.

Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2015. Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 23, 2136–2147.

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

Kolbaek, M., Yu, D., Tan, Z.-H., Jensen, J., 2017. Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 25 (10), 1901–1913.

Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., Gopinath, R., 2006. Super-human multi-talker speech recognition: the ibm 2006 speech separation challenge system. In: International Conference on Spoken Language Processing (ICSLP).

Ming, J., Hazen, T.J., Glass, J.R., 2010. Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation. Comput. Speech Lang. 24 (1), 67–76.

Mitra, V., Franco, H., 2015. Time-frequency convolutional networks for robust speech recognition. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 317–323.

Peddinti, V., Povey, D., Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In: Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 3214–3218.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The kaldi speech recognition toolkit. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU).

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S., 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In: Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 2751–2755.

Qian, Y., Bi, M., Tan, T., Yu, K., 2016. Very deep convolutional neural networks for noise robust speech recognition. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 24 (12), 2263–2276.

Qian, Y., Weng, C., Chang, X., Wang, S., Yu, D., 2018. Past review, current progress, and challenges ahead on the cocktail party problem. Front. Inf. Technol. Electron.Eng. 19 (1), 40–63.

Qian, Y., Woodland, P.C., 2016. Very deep convolutional neural networks for robust speech recognition. In: IEEE Spoken Language Technology Workshop (SLT), pp. 481–488.

Rennie, S.J., Hershey, J.R., Olsen, P.A., 2010. Single-channel multitalker speech recognition. IEEE Signal Process. Mag. (SPM) 27, 66–80.

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

Saon, G., Sercu, T., Rennie, S., Kuo, H.-K.J., 2016. The IBM 2016 english conversational telephone speech recognition system. In: Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 7–11.

Seide, F., Li, G., Yu, D., 2011. Conversational speech transcription using context-dependent deep neural networks.. In: Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 437–440.

Sercu, T., Goel, V., 2016. Dense prediction on sequences with time-dilated convolutions for speech recognition. arXiv:1611.09288v1.

Sercu, T., Puhrsch, C., Kingsbury, B., LeCun, Y., 2016. Very deep multilingual convolutional neural networks for LVCSR. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4955–4959.

Settle, S., Roux1, J.L., Hori, T., Watanabe, S., Hershey, J.R., 2018. End-to-end multi-speaker speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4819–4823.

Swietojanski, P., Ghoshal, A., Renals, S., 2013. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 285–290.

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

Wang, Y., Du, J., Dai, L.R., Lee, C.H., 2017. A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 25, 1535–1546.

Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Process. (TASLP) 22, 1849–1858.

Weng, C., Yu, D., Seltzer, M.L., Droppo, J., 2015. Deep neural networks for single-channel multi-talker speech recognition. IEEE/ACM Trans. Audio SpeechLang. Process. (TASLP) 23 (10), 1670–1679.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J., Hershey, J.R., Schuller, B., 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA). Springer-Verlag, New York, Inc., pp. 91–99.

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

Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2014. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. (SPL) 21, 65–68.

Yu, D., Chang, X., Qian, Y., 2017a. Recognizing multi-talker speech with permutation invariant training. In: Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 2456–2460.

Yu, D., Deng, L., 2014. Automatic Speech Recognition: A Deep Learning Approach. Signals and Communication Technology. Springer, London. https://books.google.com/books?id=rUBTBQAAQBAJ

Yu, D., Deng, L., Dahl, G.E., 2010. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition. NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

Yu, D., Eversole, A., Seltzer, M., Yao, K., Huang, Z., Guenter, B., Kuchaiev, O., Zhang, Y., Seide, F., Wang, H., et al., 2014. An introduction to computational networks and the computational network toolkit. Technical Report MSR-TR-2014–112. Microsoft.

Yu, D., Kolbaek, M., Tan, Z.-H., Jensen, J., 2017b. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 241–245.

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

Zhang, S., Jiang, H., Xiong, S., Wei, S., Dai, L., 2016a. Compact feedforward sequential memory networks for large vocabulary continuous speech recognition. In: Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 3389–3393.

Zhang, Y., Chen, G., Yu, D., Yao, K., Khudanpur, S., Glass, J., 2016b. Highway long short--term memory RNNs for distant speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5755–5759.