KNOWLEDGE TRANSFER IN PERMUTATION INVARIANT TRAINING FOR SINGLE-CHANNEL MULTI-TALKER SPEECH RECOGNITION

Tian Tan[1], *Yanmin Qian*[1†], *Dong Yu*[2]

[1]SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2]Tencent AI Lab, Tencent, Bellevue, WA, USA
{ tantian@sjtu.edu.cn, yanminqian@tencent.com, dyu@tencent.com}

ABSTRACT

This paper proposes a framework that combines teacher-student training and permutation invariant training (PIT) for single-channel multi-talker speech recognition. In contrast to most of conventional teacher-student training methods that aim at compressing the model, the proposed method distills knowledge from the single-talker model to improve the multi-talker model in the PIT framework. The inputs to the teacher and student networks are the single-talker clean speech and the multi-talker mixed speech, respectively. The knowledge is transferred to the student through the soft labels generated by the teacher. Furthermore, the ensemble of multiple teachers is exploited with a progressive training scheme to further improve the system. In this framework it is easy to take advantage of data augmentation and perform domain adaptation for multi-talker speech recognition using only untranscribed data. The proposed techniques were evaluated on artificially mixed two-talker AMI speech data. The experimental results show that the teacher-student training can cut the word error rate (WER) by relative 20% against the baseline PIT model. We also evaluated our unsupervised domain adaptation method on an artificially mixed WSJ0 corpus and achieved relative 30% WER reduction against the AMI PIT model.

Index Terms— permutation invariant training, knowledge distillation, multi-talker speech recognition, unsupervised training

1. INTRODUCTION

Tremendous progresses have been made in near-field single-talker automatic speech recognition (ASR) in past several years [1, 2, 3, 4, 5, 6, 7, 8]. However, under the far-field multi-talker scenario the ASR system still performs poorly. This is because the signal to noise ratio (SNR) between the target speaker and the interfering speaker is much lower than that when close-talk microphones are used.

In this paper, we aim to attack the single-channel multi-talker speech recognition problem. Many attempts have been made to address this problem. In [9], a deep learning model with two-branches was developed in which the senone labels for each branch was assigned according to the instantaneous energy. In [10, 11], the deep clustering (DPCL) technique was exploited to separate the multi-talker mixed speech into multiple streams. An ASR engine was then applied to these streams to recognize speech. In [12] the deep attractor network (DANet) was proposed. In contrast to DPCL, several cluster centers, referred to as attractor points, were created in the embedding space to pull together the time-frequency bins corresponding to the same source. In [13, 14], permutation invariant training (PIT) was proposed to attack the multi-talker speech separation problem using a simple but effective training criterion. More recently, PIT was extended to conduct multi-talker speech recognition. Promising results were reported in [15, 16, 17, 18]. Despite all these progresses, the performance gap between the multi-talker and single-talker speech recognition is still large [16].

In this work, we propose to exploit the teacher-student training in the PIT framework to improve the multi-talker speech recognition. In the conventional teacher-student training, knowledge is usually transfered from a large and complicated teacher network to a small student network[19, 20, 21, 22, 23] to help reduce model footprint. The student tries to mimic the teacher by using the soft labels estimated by the teacher. In [22], the soft label is referred to as *dark knowledge* and is considered more important than hard labels for deep learning. In [24, 25], the teacher-student training was proposed to transfer knowledge from the clean-speech recognition model to the noisy-speech recognition model.

Different from these prior arts, we aim at transferring knowledge from the single-talker ASR model to the multi-talker ASR model in the PIT framework. In our work, the inputs to the teacher and student networks are the single-talker clean speech and the multi-talker mixed speech, respectively. The knowledge is transferred to the student through the soft labels estimated by the teacher. Furthermore, the ensemble of multiple teachers is exploited with a progressive training scheme to further improve the system. In this framework it is easy to take advantage of data augmentation and perform domain adaptation for multi-talker speech recognition using only untranscribed data[1].

The paper is organized as follows: In Section 2 we briefly introduce PIT for single-channel multi-talker speech recognition. We describe knowledge distillation and transfer in the PIT framework in Section 3. In Section 4 we report experimental results. We conclude the paper in Section 5

2. PERMUTATION INVARIANT TRAINING FOR SINGLE-CHANNEL MULTI-TALKER ASR

Permutation invariant training (PIT) [13, 15] is an efficient and effective technique for solving the label ambiguity problem in deep learning based multi-talker speech recognition. The basic architecture of PIT for multi-talker speech recognition (PIT-ASR), proposed

[†] Yanmin Qian and Dong Yu are the corresponding authors and now Yanmin Qian is with Tencent AI Lab, Tencent, Bellevue, WA, USA.

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

[1] We noticed that a similar idea was proposed in [26] and posted to arXiv few weeks before this submission. Our work was conducted independently of theirs. In contrast to their work, we proposed and studied different architectures and conducted more comprehensive investigation including the exploitation of the ensemble of teachers and the unsupervised adaptation.

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

in our previous work [15], is depicted in the middle part of Figure 1. In this model, the mixed speech $\mathbf{O}$ is sent to the deep learning model to estimate the state-level posterior for each talker. For better ability of modeling long-range dependency, which can improve speaker tracing, recurrent neural networks (RNNs) are usually used. In this work, we apply bidirectional LSTM-RNNs in all models.

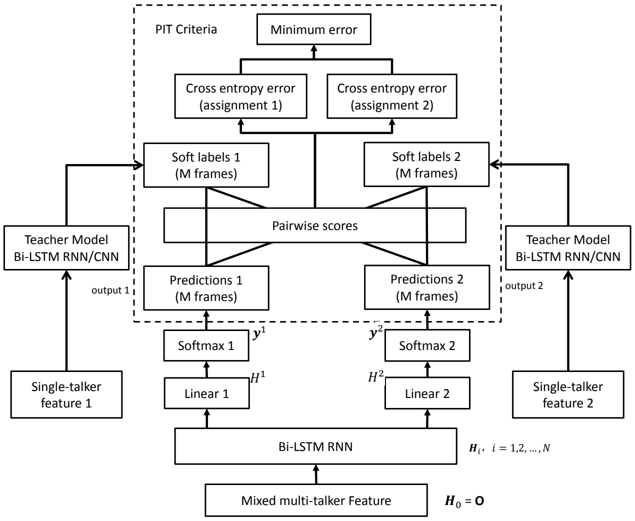

Fig. 1. The basic permutation invariant training architecture and the proposed knowledge distillation architecture for multi-talker speech recognition

The key ingredient in PIT is its training criterion. Let $\mathbf{l}^s$ and $\mathbf{y}^s$ be the ground truth alignment and the estimated state-posterior of stream s, respectively, the objective function in PIT is defined as

$$J = \frac{1}{S} \min_{\mathbf{s}' \in permu(S)} \sum_s \sum_t \mathrm{CE}(l_t^{\mathbf{s}'_s}, \mathbf{y}_t^s) \tag{1}$$

where $\mathbf{s}'$ is a permutation of $[1, 2, \cdots, S]$ and $l_t^{\mathbf{s}'_s}$ is the ground truth label of stream $\mathbf{s}'_s$ at frame t. PIT aims to minimize the minimal average cross entropy (CE) of the whole utterance among all possible assignments of the reference to the estimated posterior.

With this criterion, the network can automatically estimate the assignment. The CE is computed over the whole utterance so that all frames that belong to the same speaker are forced to be aligned with the same output segment (or branch). Moreover, compared to DPLC [10] or DANet [12], this structure is much simpler since it allows direct multi-talker mixed speech recognition without explicit separation. After the PIT model training, the individual output posterior stream can be used for decoding as normal to obtain the final recognition result.

3. KNOWLEDGE DISTILLATION FROM SINGLE-TALKER ACOUSTIC MODEL WITHIN PIT

3.1. Knowledge transfer within PIT

3.1.1. Conventional teacher-student training

Instead of using only hard labels in traditional machine learning tasks, the teacher-student training additionally uses the posterior probability generated by the teacher as the supervision. The objective function in the teacher-student training is defined as

$$\mathrm{CE}(\theta; \mathbf{O}, \mathbf{L}) \;=\; -\sum_t \sum_y p'(y|\mathbf{o}_t) \log p_\theta(y|\mathbf{o}_t) \tag{2}$$

$$p'(y|\mathbf{o}_t) \;=\; \lambda p_{teacher}(y|\mathbf{o}_t) + (1-\lambda) p_t^{\mathrm{ref}}(y) \tag{3}$$

where $p_\theta(y|\mathbf{o}_t)$ is the posterior generated by the student model, $p_t^{\mathrm{ref}}(y)$ is the reference distribution and is represented by a Kronecker delta function $p_t^{\mathrm{ref}}(y) = \delta(y, l_t^{ref})$, and l_t^{ref} is the ground truth label, which is referred to as *hard* label usually. $p_{teacher}(y|\mathbf{o}_t)$ is the posterior probability estimated by the teacher model, which is also referred to as *soft* label because it is not a one-hot vector. λ is the interpolation weight, which is a hyper parameter. [22] suggested that this new label $p'(y|\mathbf{o}_t)$ can encode correlations among different classes and is better than the hard label.

3.1.2. Teacher-student Training in PIT

In most previous works the teacher-student training was used to transfer knowledge from a larger teacher model to a smaller student model to reduce model footprint. For this reason, the inputs to both the teacher and student models are the same. Our proposed method, however, transfers knowledge from a single-talker speech recognition model (the teacher) to a multi-talker one (the student) to improve the recognition accuracy of the student model. For this reason, parallel data of original individual single-talker speech and the corresponding multi-talker mixed speech, are used. The whole architecture is illustrated in Figure 1. The inputs to the teacher model are original single-talker speech streams $\mathbf{o}_t^{\mathbf{s}'_s}$, and the inputs to the student model are the corresponding multi-talker mixed speech $\mathbf{o}_t$. The training criterion in this architecture is

$$J = \frac{1}{S} \min_{\mathbf{s}' \in permu(S)} \sum_s \sum_t \sum_y p'(y|\mathbf{o}_t^{\mathbf{s}'_s}) \log p_\theta^s(y|\mathbf{o}_t) \tag{4}$$

$$p'(y|\mathbf{o}_t^{\mathbf{s}'_s}) = \lambda p_{teacher}(y|\mathbf{o}_t^{\mathbf{s}'_s}) + (1-\lambda) p_{t,\mathbf{s}'_s}^{\mathrm{ref}}(y) \tag{5}$$

where $p_\theta^s(y|\mathbf{o}_t)$ is the posterior of stream s estimated by the student model, $p_{t,\mathbf{s}'_s}^{\mathrm{ref}}(y) = \delta(y, l_t^{\mathbf{s}'_s})$ is the reference distribution. $p_{teacher}(y|\mathbf{o}_t^{\mathbf{s}'_s})$ is the posterior estimated by the teacher model using original single-talker speech stream $\mathbf{s}'_s$. Different from the basic PIT in Equation 1, we minimize the minimal average CE among all possible assignment between model outputs and soft labels.

3.2. Knowledge Transfer from Ensemble of Teachers

In Section 3.1.2, the soft label is generated by a single teacher. It is reasonable to believe that further performance improvement may be achieved if the soft label is generated using an ensemble of teachers [27] such that

$$p_{teacher}(y|\mathbf{o}_t^{\mathbf{s}'_s}) = \sum_k w_k p_k(y|\mathbf{o}_t^{\mathbf{s}'_s}) \tag{6}$$

where $w_k \in [0,1] and \sum_k w_k = 1$ are the interpolation weights. $p_k(y_t^{\mathbf{s}'_s}|\mathbf{o}_t)$ is the posterior estimated by different teacher models.

Instead of using the soft label averaged over those generated by multiple teachers as the teacher provided supervision, we also developed a progressive ensemble learning scheme which obtained better performance. As shown in algorithm 1, in this alternative approach, the teacher-student learning is applied to the PIT model by using the teachers one by one, in the ascending order of their recognition performance on single-talker tasks.

Algorithm 1 Progressive ensemble teacher-student training

```
1: Sort teacher models in ascending order of the performance on single-talker task
2: for each $i$ in all teachers do
3:     for each $j$ in all minibatches of training data do
4:         Generate soft-targets for minibatch $j$ using teacher model $i$
5:         Update neural network model with minibatch $j$
6:     end for
7:     Repeat 3 until converge
8: end for
```

3.3. Unsupervised Knowledge Transfer for Data Augmentation and Domain Adaptation

The teacher-student training can be further extended by exploiting large amount of unlabeled data. In the conventional teacher-student training, the target distribution is a weighted sum of the *hard* and *soft* labels. When unlabeled data are used, only the posterior estimated by the teacher model is used as the supervision, i.e.,

$$J = \frac{1}{S} \min_{\mathbf{s}' \in permu(S)} \sum_s \sum_t \sum_y p_{teacher}(y|\mathbf{o}_t^{\mathbf{s}'_s}) \log p_\theta(y|\mathbf{o}_t) \quad (7)$$

Parallel data used as inputs to the model are still needed in this setup. However, they are easy to generate since we can just vary the relative energy of the involved talkers without transcribing the source streams.

This approach also allows us to conduct fast domain adaptation with only untranscribed target domain speech. More specifically, we can first train a general PIT-based multi-talker ASR model and then collect in-domain speech waveform without transcription and use them to adapt the general model to generate the domain-dependent model.

4. EXPERIMENT

To evaluate the performance of the proposed methods, experiments were conducted on two artificially generated multi-talker mixed speech datasets. One is based on AMI IHM corpus [28], which was used in our previous work [16] and has been released to public by us. Another is based on Wall Street Journal (WSJ0) corpus [29], which was used in [10] and released by MERL. The multi-talker AMI IHM corpus consists of two setups: (1) full 400hr speech, with the energy ratio of the two-talkers under five different SNR conditions (0dB, 5dB, 10dB, 15dB, 20dB), each of which contains 80hr mixed speech. (2) a subset of 80hr speech, which was used for fast model training and evaluation. For evaluation, 8hr two-talker mixed speech with all SNR levels is generated. More details can be found in [16]. Additional details on the multi-talker WSJ corpus can be found in [10].

4.1. Baseline systems

In this work, all networks were built using the Microsoft Cognitive Toolkit (CNTK2.0) [30] and the decoding was conducted using Kaldi [31]. A single-talker LDA-MLLT-SAT GMM-HMM system was first trained based on the standard Kaldi recipe using AMI IHM data. This model uses 39-dim MFCC feature and has roughly 4K tied-states and 80K Gaussians. This acoustic model was used to generate the senone alignment for neural network training. Then CNN and BLSTM-RNN based acoustic models were constructed, 40 dimensional log filter bank (LFBK) features with CMVN was used to train the baselines. The structure of CNN model is the same as that in [32]. The BLSTM-RNN contains 3 bidirectional LSTM layers. Each BLSTM layer has 768 memory cells. CE was used to train all models. The CNN was optimized by SGD with minibatches of 256 samples, and the BLSTM-RNN was trained using SGD with 4 full-length utterances in each minibatch.

For decoding, a 50K-word dictionary and a trigram LM interpolated from the LMs created using the AMI transcripts and the Fisher English corpus were used. The performance of these two baselines on the original single-speaker AMI IHM corpus are presented in Table 1. We can observe that they achieve comparable performance on this corpus.

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

Model	WER
CNN	26.6
BLSTM	27.0

The basic PIT model proposed in our previous work [16] was built. It contains 6 BLSTM layers, each of which contains 768 memory cells. The gradient was clipped to 0.0003 to guarantee the training stability. This model was trained by PIT-CE on the 80hr AMI IHM-2mix subset, and the result was shown at the first row of Table 2. This original PIT-CE ASR model can recognize two-talker speech. However, there is still a large performance gap (WER is doubled) between the two- and single-talker speech recognition scenario. More advanced technologies are needed to improve the multi-talker ASR system.

4.2. Experiment on teacher-student training

We investigated the different configurations of teacher-student training using the 80hr training subset. In this experiment, the posteriors were obtained from the single-talker BLSTM-RNN model using original single-talker speech features, and the input for the PIT model was still two-talker mixed speech features. The interpolation weights and initialization modes were investigated, and the results are shown in Table 2.

Table 2. WER (%) of the PIT model with teacher-student training using different configurations on the 80hr AMI IHM-2mix dataset. TS means teacher-student training.

Model	Init	λ	WER	
			SPK1	SPK2
PIT	Random	—	55.21	64.23
+TS	PIT	0.5	52.44	60.49
		1	51.84	60.34
	Random	0.5	51.28	59.27
		1	**51.07**	**59.12**

It is observed that PIT-ASR with teacher-student training outperforms the baseline PIT-ASR system for both speakers. Random initialization (training from scratch) achieves better performance than initialization from pre-trained PIT model. Simply using the soft label ($\lambda = 1.0$, i.e., without using the hard label) performs the best. Based on these experiments, random initialization from scratch and $\lambda = 1.0$ were used for all the rest experiments.

4.3. Experiment on different teachers and teacher ensembles

Different teachers and teacher ensembles were evaluated. First, the CNN and BLSTM baseline single-talker models were used as teachers (shown in Table 1). The experimental results are shown in Table 3. It is observed that all single-talker models can effectively improve the multi-talker recognition accuracy with teacher-student training. Using CNN as the teacher obtained 2% absolute WER improvement than that using BLSTM-RNN. We conjecture that this may because the BLSTM-RNN is used in PIT-ASR model while CNN is not and thus provides more complementary information. It is also possible that posteriors provided by CNNs are more informative although the WER obtained by the CNN model is only slightly lower than that achieved by the BLSTM-RNN model.

Table 3. WER (%) of the teacher-student training using ensemble of single-speaker teacher models on 80hr AMI IHM-2mix dataset

Model	Teacher	WER	
		SPK1	SPK2
PIT	—	55.21	64.23
+TS	BLSTM	51.07	59.12
	CNN	48.95	57.52
	BLSTM+CNN: interpolated	49.34	57.78
	BLSTM+CNN: progressive	**48.03**	**56.46**

In addition, different teacher ensembles were evaluated. The interpolation approach and the progressive approach were compared. The results, shown in the bottom part of Table 3, indicate that different ensemble achieves different performance for the teacher-student training within the PIT framework. Surprisingly, the direct interpolation of the two soft labels from individual teachers does not lead to further improvement over the best single-model teacher (CNN in our case). In contrast, the progressive ensemble that uses teachers one by one in the ascending order of performance can achieve further performance improvement compared to the single teacher.

4.4. Experiments on using untranscribed data

4.4.1. *Experiments on data augmentation on AMI full set*

As described in Section 3, we can exploit un-transcribed data to improve the system performance. The full 400hr AMI IHM-2mix data were used here. Except the original 80hr subset, the rest two-talker mixed speech was used without transcription and senone-alignment. Table 4 compares WER achieved with and without untranscribed data augmentation on the AMI IHM-2mix dataset. It is observed that using the knowledge distillation with additional untranscribed data can obtain further improvement.

Shown in the last row of Table 4, using both the teacher ensemble and un-transcribed data augmentation within the PIT-based knowledge distillation framework achieves the best performance for multi-talker ASR on AMI IHM-2mix dataset. Compared to the basic PIT-ASR model, the new approach reduced WERs from 55.21% and 64.23% to 43.56% and 51.29% for two talkers respectively.

4.4.2. *Experiments on fast domain adaptation*

Finally, fast domain adaptation without transcription was explored. The source domain is AMI meeting speech and the target domain is WSJ reading speech. The target multi-talker mixed speech was synthesized by mixing two separated clean utterances from WSJ0 corpus [29]. A PIT-ASR model was trained for meeting speech using AMI IHM-2mix data and adapted to the WSJ reading speech. The standard trigram language model and dictionary for WSJ0 were used here for evaluation, generated by the standard Kaldi recipe. The related results, reported in Table 5, show that the WER is very high when using the AMI-based model to recognize WSJ two-talker mixed speech directly, since there is a big mismatch on the acoustic conditions. Using the proposed domain adaptation technique with un-transcribed data reduced WER from 51.8% to 35.2%.

Table 4. Compare WER (%) with and without using the untranscribed data in the teacher-student training framework on the AMI IHM-2mix dataset

Model	Teacher	Data	Label	WER	
				SPK1	SPK2
PIT	—	80hr	Labeled	55.21	64.23
	—	400hr	Labeled	49.19	57.06
+TS	BLSTM	80hr	Labeled	51.07	59.12
		+320hr	Unlabeled	45.11	53.31
	CNN	80hr	labeled	48.95	57.52
		+320hr	Unlabeled	44.59	52.25
	BLSTM+CNN	80hr	labeled	48.03	56.46
		+320hr	Unlabeled	**43.58**	**51.29**

Table 5. Efficient domain adaptation from AMI Meeting speech to WSJ Reading speech for multi-talker speech recognition with only untranscribed WSJ data. WER (%) on WSJ-2mix

System	Teacher	WER
PIT Baseline AMI 80hr	—	51.81
+ WSJ domain adaptation	AMI BLSTM	38.77
PIT-TS AMI 400hr	AMI BLSTM	43.50
+ WSJ domain adaptation	AMI BLSTM	36.59
PIT-TS AMI 400hr	AMI BLSTM+CNN	38.56
+ WSJ domain adaptation	AMI BLSTM+CNN	**35.21**

5. CONCLUSION

In this work, knowledge distillation and transfer was applied to the PIT-ASR model to improve single-channel multi-talker speech recognition. The knowledge is transferred from the single-talker model to the multi-talker model. We also proposed the progressive teacher ensemble technique to further improve knowledge distillation. We showed that the proposed framework allows exploitation of untranscribed data to either improve the multi-talker speech recognition accuracy or perform fast domain adaptation. The experiments on the multi-talker AMI and WSJ corpora showed that the proposed methods can significantly improves the performance of multi-talker speech recognition.

6. REFERENCES

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

[2] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent Pre-trained Deep Neural Networks for Large-vocabulary Speech Recognition," *IEEE/ACM Transactions on*

Audio, Speech, and Language Processing (TASLP), vol. 20, pp. 30–42, 2012.

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[4] Yanmin Qian, Mengxiao Bi, Tian Tan, and Kai Yu, "Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 12, pp. 2263–2276, 2016.

[5] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts," in *INTERSPEECH*, 2015, pp. 3214–3218.

[6] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," in *ICML*, 2016.

[7] Dong Yu, Wayne Xiong, Jasha Droppo, Andreas Stolcke, Guoli Ye, Jinyu Li, and Geoffrey Zweig, "Deep Convolutional Neural Networks with Layer-Wise Context Expansion and Attention," in *INTERSPEECH*, 2016, pp. 17–21.

[8] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "The Microsoft 2016 Conversational Speech Recognition System," in *ICASSP*, 2017, pp. 5255–5259.

[9] Chao Weng, Dong Yu, Michael L. Seltzer, and Jasha Droppo, "Deep Neural Networks for Single-Channel Multi-talker Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1670–1679, 2015.

[10] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *ICASSP*, 2016, pp. 31–35.

[11] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-Channel Multi-Speaker Separation Using Deep Clustering," in *INTERSPEECH*, 2016, pp. 545–549.

[12] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep Attractor Network for Single-Microphone Speaker Separation," in *ICASSP*, 2017, pp. 246–250.

[13] Dong Yu, Morten Kolbk, Zheng-Hua Tan, and Jesper Jensen, "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation," in *ICASSP*, 2017, pp. 241–245.

[14] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[15] Dong Yu, Xuankai Chang, and Yanmin Qian, "Recognizing Multi-talker Speech with Permutation Invariant Training," in *INTERSPEECH*, 2017, pp. 2456–2460.

[16] Yanmin Qian, Xuankai Chang, and Dong Yu, "Single-Channel Multi-talker Speech Recognition with Permutation Invariant Training," *CoRR*, vol. abs/1707.06527, 2017.

[17] Xuankai Chang, Yanmin Qian, and Dong Yu, "Adaptive Permutation Invariant Training with Auxiliary Information for Monaural Multi-Talker Speech Recognition," in *ICASSP*, 2018.

[18] Zhehuai Chen and Jasha Droppo, "Sequence Modeling in Unsupervised Single-channel Overlapped Speech Recognition," in *ICASSP*, 2018.

[19] Jimmy Ba and Rich Caruana, "Do Deep Nets Really Need to be Deep?," in *NIPS*, 2014, pp. 2654–2662.

[20] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning Small-Size DNN with Output-Distribution-Based Criteria," in *INTERSPEECH*, 2014.

[21] William Chan, Nan Rosemary Ke, and Ian Lane, "Transferring Knowledge from a RNN to a DNN," in *INTERSPEECH*, 2015.

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," *CoRR*, vol. abs/1503.02531, 2015.

[23] Liang Lu, Michelle Guo, and Steve Renals, "Knowledge Distillation for Small-Footprint Highway Networks," in *ICASSP*. IEEE, 2017, pp. 4820–4824.

[24] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey, "Student-Teacher Network Learning With Enhanced Features," in *INTERSPEECH*, 2017, pp. 5275–5279.

[25] Jinyu Li, Michael L. Seltzer, Xi Wang, Rui Zhao, and Yifan Gong, "Large-Scale Domain Adaptation via Teacher-Student Learning," in *INTERSPEECH*, 2017, pp. 2386–2390.

[26] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive Joint Modeling in Unsupervised Single-Channel Overlapped Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 1, pp. 184–196, 2018.

[27] Yevgen Chebotar and Austin Waters, "Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition," in *INTERSPEECH*, 2016, pp. 3439–3443.

[28] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al., "The AMI Meeting Corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005, vol. 88.

[29] Garofolo, John, et al., "CSR-I (WSJ0) Complete LDC93S6A," Philadelphia: Linguistic Data Consortium, 1993.

[30] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Huaming Wang, et al., "An Introduction to Computational Networks and the Computational Network Toolkit," *Microsoft Technical Report MSR-TR-2014–112*, 2014.

[31] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi Speech Recognition Toolkit," in *ASRU*, 2011.

[32] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, "Deep Convolutional Neural Networks for LVCSR," in *ICASSP*. IEEE, 2013, pp. 8614–8618.