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

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

ADAPTIVE PERMUTATION INVARIANT TRAINING WITH AUXILIARY INFORMATION FOR MONAURAL MULTI-TALKER SPEECH RECOGNITION

Xuankai Chang[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
{xuank@sjtu.edu.cn, yanminqian@tencent.com, dyu@tencent.com}

Lu Huang

2018-08-01

Single-channel Multi-talker

- Linearly mixed single-microphone signal

$$y_t = \sum_n x_t^n$$

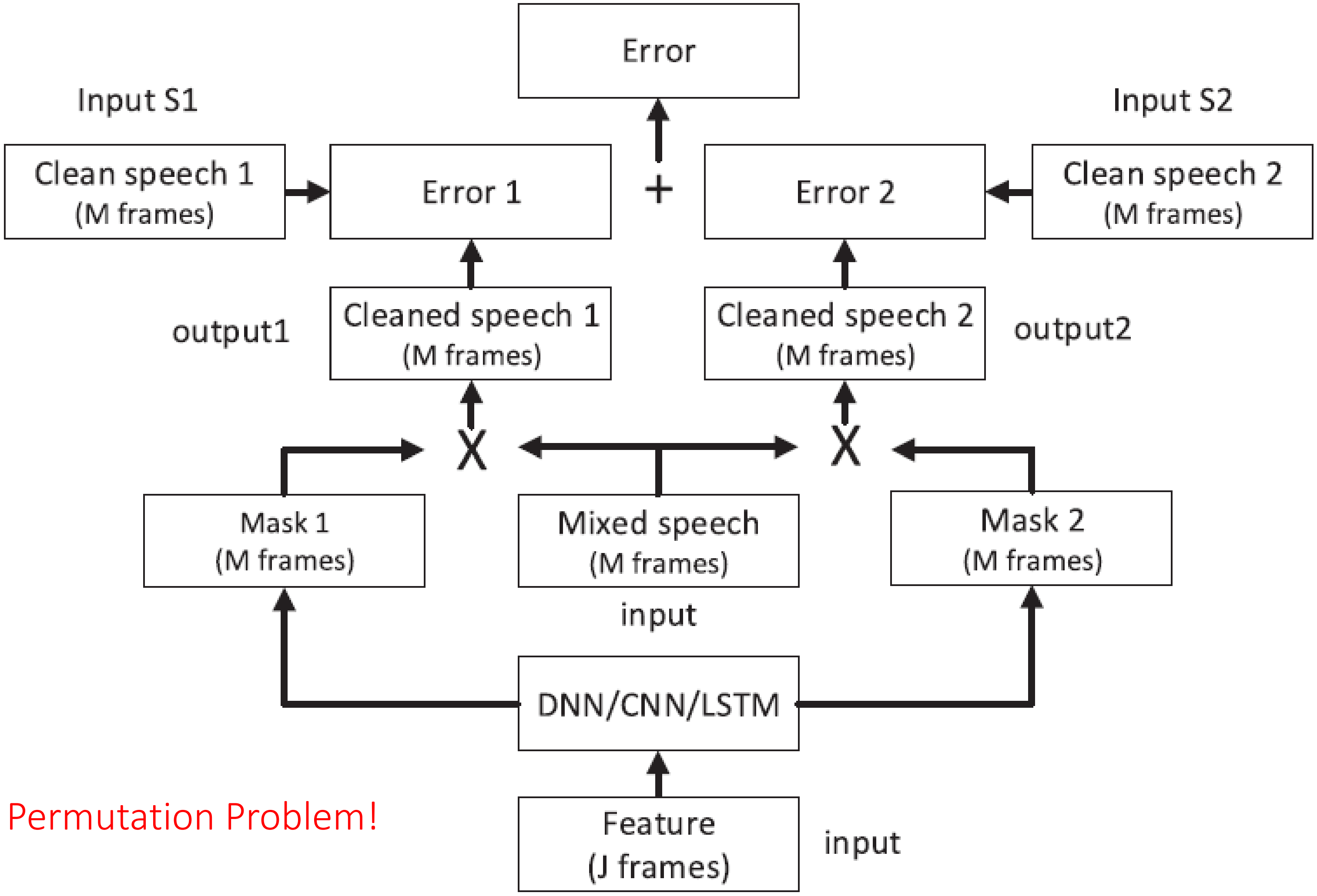
- Mask based spectra estimation
  - Amplitude Mask (AM)

$$\mathcal{J}_{\text{mask}} = \sum_n \text{MSE}(\hat{\mathbf{M}}_n \odot \mathbf{Y}, \mathbf{X}_n)$$

  - Phase Sensitive Mask(PSM)

$$\mathcal{J}_{\text{mask}} = \sum_n \text{MSE}(\hat{\mathbf{M}}_n \odot \mathbf{Y}, \mathbf{X}_n \odot \cos(\theta_y - \theta_n))$$

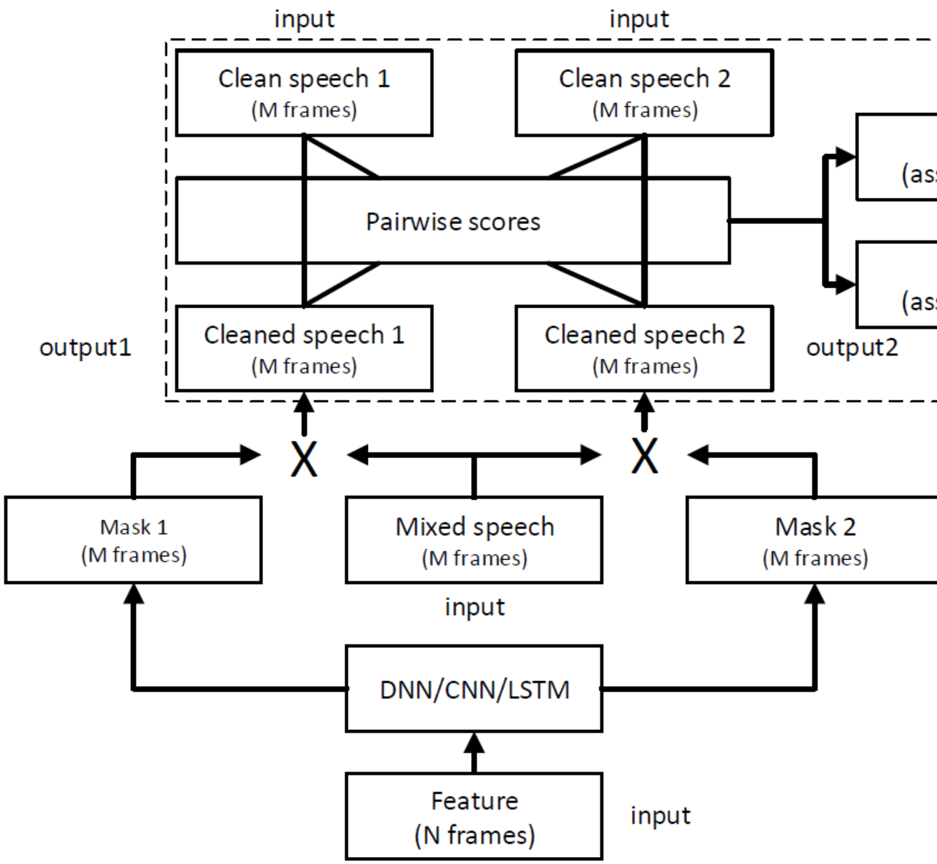
Conventional Speech Separation



Label Permutation Problem!

Kolbæk, Morten, et al. "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks." TASLP 25.10 (2017): 1901-1913.

Permutation Invariant Training



Speaker Tracing Problem!

$$\mathbf{M}_s = \frac{|\mathbf{X}_s|}{|\mathbf{Y}|}$$

$$J_m = \frac{1}{T \times F \times S} \sum_{s=1}^{S} \|\tilde{\mathbf{M}}_s - \mathbf{M}_s\|^2$$

$$J_x = \frac{1}{T \times F \times S} \sum_{s=1}^{S} \||\tilde{\mathbf{X}}_s| - |\mathbf{X}_s|\|^2$$

$$\mathcal{J}_{\text{f-PIT}} = \sum_u \sum_t \min_{s'} \sum_n \text{MSE}(\hat{\mathbf{M}}_{utn} \odot \mathbf{Y}_{ut}, \mathbf{X}^{s'}_{utn})$$

$$\mathcal{J}_{\text{c-PIT}} = \sum_u \sum_c \min_{s'} \sum_t \sum_n \text{MSE}(\hat{\mathbf{M}}_{uctn} \odot \mathbf{Y}_{uct}, \mathbf{X}^{s'}_{uctn})$$

Yu, Dong, et al. “Permutation invariant training of deep models for speaker-independent multi-talker speech separation.” *ICASSP 2017.*

Utterance-level PIT

- Extended to utterance level

$$\mathcal{J}_{\text{u-PIT}} = \sum_{u} \min_{s'} \sum_{t} \sum_{n} \text{MSE}(\hat{\mathbf{M}}_{utn} \odot \mathbf{Y}_{ut}, \mathbf{X}^{s'}_{utn})$$

- Speaker tracing by
  - Utterance level training criterion: forcing frames belonging to the same speaker to the same output streams
  - Deep (B)LSTM: Capturing long-term dependency
- How to recognize?
  - Separation -> Recognition
  - Directly recognition (PIT-CE)

Kolbæk, Morten, et al. "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks." TASLP 25.10 (2017): 1901-1913.

uPIT CE

- Loss

$$\mathcal{J}_{\text{PIT-CE}} = \sum_{u} \min_{s'} \sum_{t} \sum_{n} \mathrm{CE}(\mathbf{O}_{unt}, l^{s'}_{unt})$$

- CNTK NN + Kaldi Decode

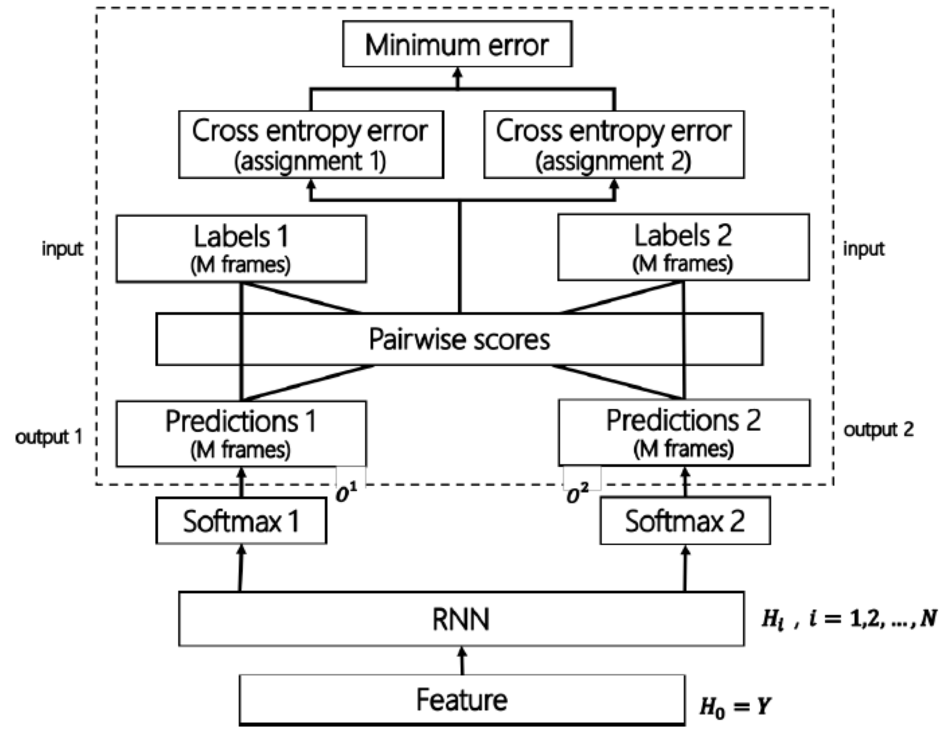


Table 2: *WER (%) of the baseline BLSTM-RNN system on two-talker mixed AMI IHM speech*

SNR Condition	High E Spk	Low E Spk
0db	85.0	100.5
5db	68.8	110.2
10db	51.9	114.9
15db	39.3	117.6
20db	32.1	118.7

Table 3: *WER (%) of the propsoed PIT-ASR model on two-talker mixed AMI IHM speech*

SNR Condition	High E WER	Low E WER
0db	49.74	56.88
5db	40.31	60.31
10db	34.38	65.52
15db	31.24	73.04
20db	29.68	80.83

Yu, Dong, Xuankai Chang, and Yanmin Qian. "Recognizing Multi-talker Speech with Permutation Invariant Training." Interspeech 2017.

ADAPTIVE PERMUTATION INVARIANT TRAINING WITH AUXILIARY INFORMATION FOR MONAURAL MULTI-TALKER SPEECH RECOGNITION

Xuankai Chang[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
{xuank@sjtu.edu.cn, yanminqian@tencent.com, dyu@tencent.com}

Motivation

- PIT-CE's WER is still much higher than that in single-talker case.
- Speaker adaptation reduces the mismatch between the training and testing speakers in single-talker ASR.
- PIT-MSE and CE are much harder on same-gender task, so the gender information may be useful.

Speaker Adaptation

- Three kinds of methods:
  - Feature space transform, eg. CMLLR, fMLLR;
  - Adapting all or part of parameters of NN;
  - Auxiliary features, like i-vector, pitch, T60...
- In this paper
  - Pitch + i-vector
  - Speaker-dependent feature can make the speaker tracing of PIT-CE more easier.

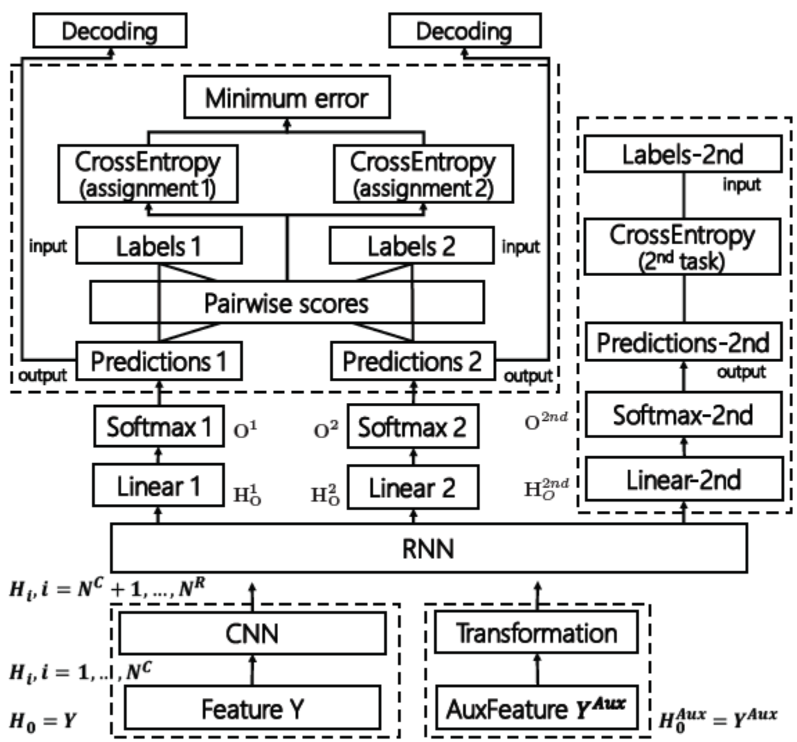
Gender-pair Prediction with MTL

- 3-dimensional one-hot vector:
  - Male + Male, Female + Female, Opposite-Gender.
- Multi-task learning
  - Add a branch to predict the gender-pair information
  - Will be deprecated when decoding

$$J^{MTL} = J + \lambda \sum_{t} CE(\ell_t^{2nd}, \mathbf{O}_t^{2nd})$$

CNN-BLSTM

- Add CNN layers before BRNN:
  - CNN layers only use the normal features (LFBK)
  - Modeling the correlation along frequency axis



Experiment Setup

- Data
  - Mixed two-talker AMI IHM corpus.
  - 80hr train; 8hr evaluation.
- Tools: CNTK NN + Kaldi Decode
- Features: 40-dim LFBk, 3-dim pitch, 10-dim i-vector
- Multi-task Learning: $\lambda = 0.3$
- Model
  - 6-layer BLSTM(768)
  - 2-layer CNN + 4-layer BLSTM (CNN-BLSTM)
    - 11×40; 32*(9×9-1×1); 64*(3×3-2×2)

Experiment Results

- BLSTM vs. CNN-BLSTM

Model	Gender Combination	WER 1	WER 2
BLSTM	All	**55.21**	**64.23**
	opposite	52.41	61.61
	same	58.48	67.27
CNN-BLSTM	All	**51.93**	**60.13**
	opposite	49.40	57.90
	same	54.89	62.72

Experiment Results

- With/without auxiliary feature

Model	Adapt on	WER 1	WER 2
BLSTM	—	55.21	64.23
	pitch	51.88	60.54
	i-vector	51.61	59.99
	pitch + i-vector	51.29	59.78
CNN-BLSTM	pitch + i-vector	50.64	58.78

Experiment Results

- gender-pair prediction with multi-task Learning

Model	2nd Task	Adapt on	WER 1	WER 2
BLSTM	—	—	55.21	64.23
	gender	—	52.47	60.31
		pitch+i-vector	51.11	59.35
CNN-BLSTM	—	—	51.93	60.13
	gender	—	51.10	58.76
		pitch+i-vector	**50.21**	**58.17**

Conclusions

- CNN-BLSTM outperforms BSLTM.
- Auxiliary feature and gender-pair prediction with MTL benefits the PIT-CE.
  - PIT can be combined with advanced techniques, like adaptation and MTL.
  - PIT is a nice modeling technique for multi-taker speech separation or recognition.

ICASSP 2018

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

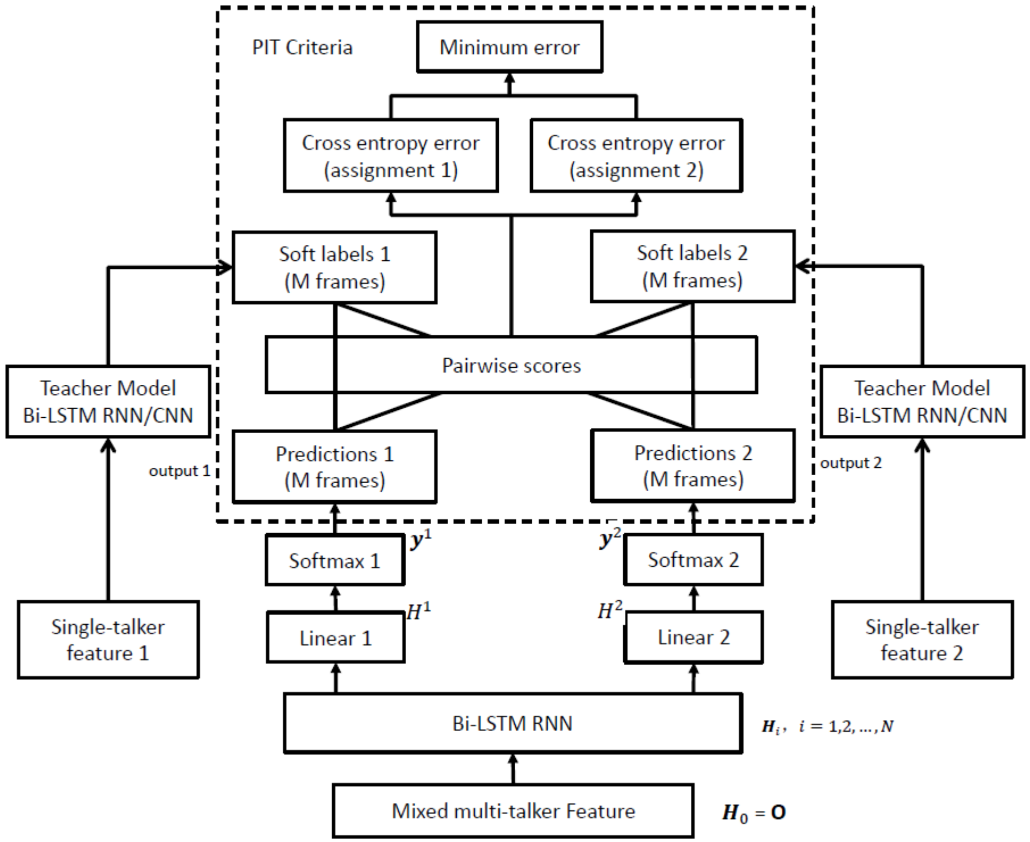
Tian Tan[1], *Yanmin Qian*[1†], *Dong Yu*[2]

[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}

Motivation

- The performance gap between the multi-talker and single-talker speech recognition is still large.
- Distills knowledge from the single-talker model to improve the multi-talker model in the PIT framework.
- How to further improve the PIT-CE system using multiple teachers?
- How to use the untranscribed data for data augmentation or domain adaptation?

Teacher-student Training



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

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

Multiple Teachers

- Average soft label

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

- Progressive ensemble learning scheme

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
```

Unsupervised Knowledge

- Use only soft labels

$$J = \frac{1}{S} \min_{\mathbf{s}' \in permu(S)} \sum_{s} \sum_{t} \sum_{y} p_{teacher}(y|\mathbf{o}_t^{\mathbf{s}'_s}) \log p_\theta(y|\mathbf{o}_t)$$

- Data augmentation
  - Use the unlabeled data to improve the model
- Domain adaptation
  - Use the in-domain data to adapt the general model

Experiment Setup

- Data
  - Mixed two-talker AMI IHM corpus.
  - 400hr train (80hr for fast training); 8hr evaluation
  - WSJ two-talker mixed speech (40hr train, 5hr test)
- Tools: CNTK NN + Kaldi Decode
- Features: 40-dim LFBK with CMVN
- Teacher Model: CNN; BLSTM(3*768)
- Student Model: BLSTM(6*768)

Experiment Results

- Teacher-student training

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

Model	**WER**
CNN	26.6
BLSTM	27.0

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**

Experiment Results

- Multiple teachers
  - CNN outperforms BLSTM
    - Because the BLSTM-RNN is used in PIT-ASR model while CNN is not and thus provides more complementary information.
    - Posteriors provided by CNNs are more informative.

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	WER 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**

Experiment Results

- Data augmentation using untranscribed data

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	WER 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**

Experiment Results

- Domain adaptation from AMI to WSJ
  - Meeting -> reading

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**

Conclusions

- Knowledge transferred from signal-talker model significantly improve the accuracy of multi-taker model.
- Progressive teacher ensemble technique can improve the results of multiple teachers.
- The teacher-student architecture can use unlabeled data for augmentation or in-domain data for adaptation of the general model.

Thx & QA