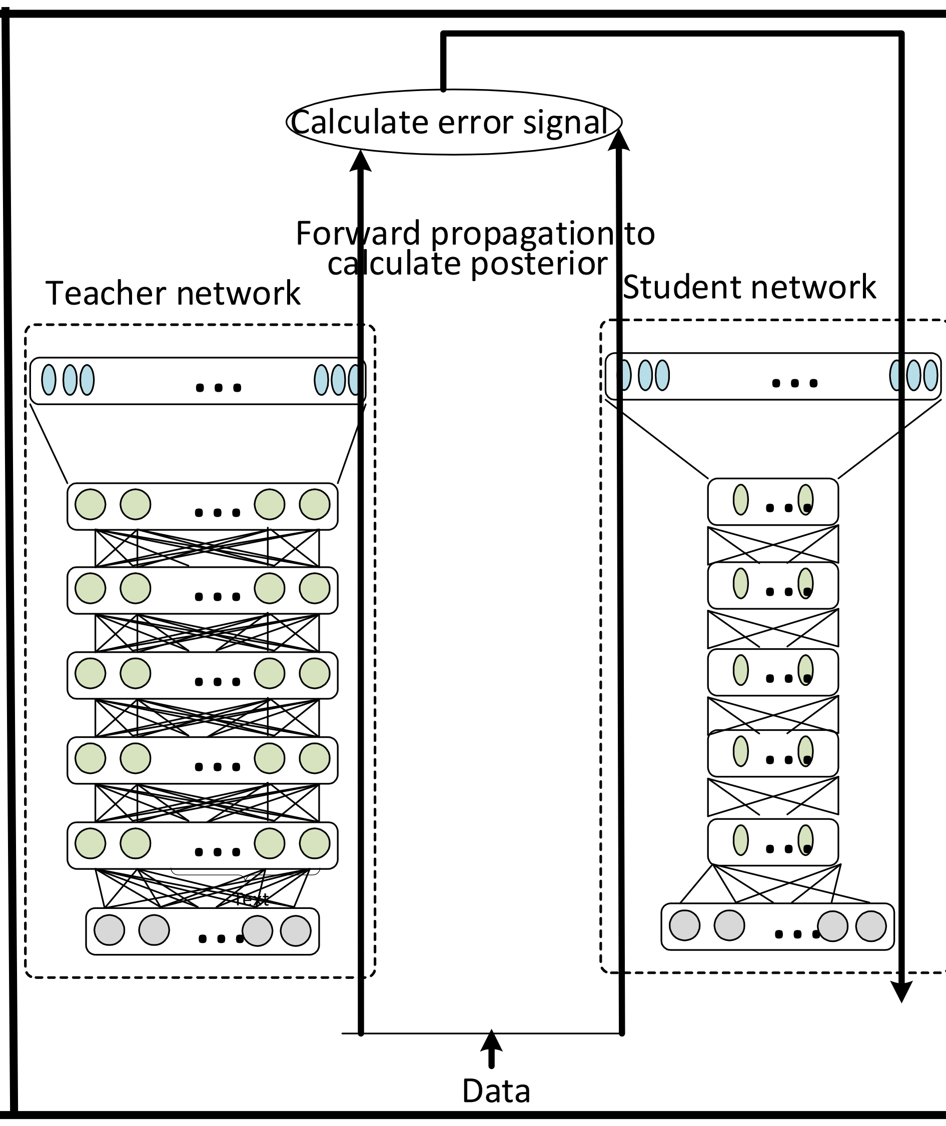
DEVELOPING FAR-FIELD SPEAKER SYSTEM VIA TEACHER-STUDENT LEARNING

Jinyu Li, Rui Zhao, Zhuo Chen, Changliang Liu, Xiong Xiao, Guoli Ye, and Yifan Gong
Microsoft AI and Research, USA

1. Introduction

We develop keyword spotting (KWS) and acoustic model (AM) components in a **far-field speaker system**.

- Use teacher-student (T/S) learning to adapt a *close-talk* well-trained production AM to *far-field* by using parallel close-talk and **simulated** far-field data.
- Use T/S learning to compress a *large-size* KWS model into a *small-size* one to fit the **device** computational cost requirement.
- Utilize unlabeled data to boost the model performance in both scenarios.



2. Teacher-Student (T/S) Learning

- T/S model compression

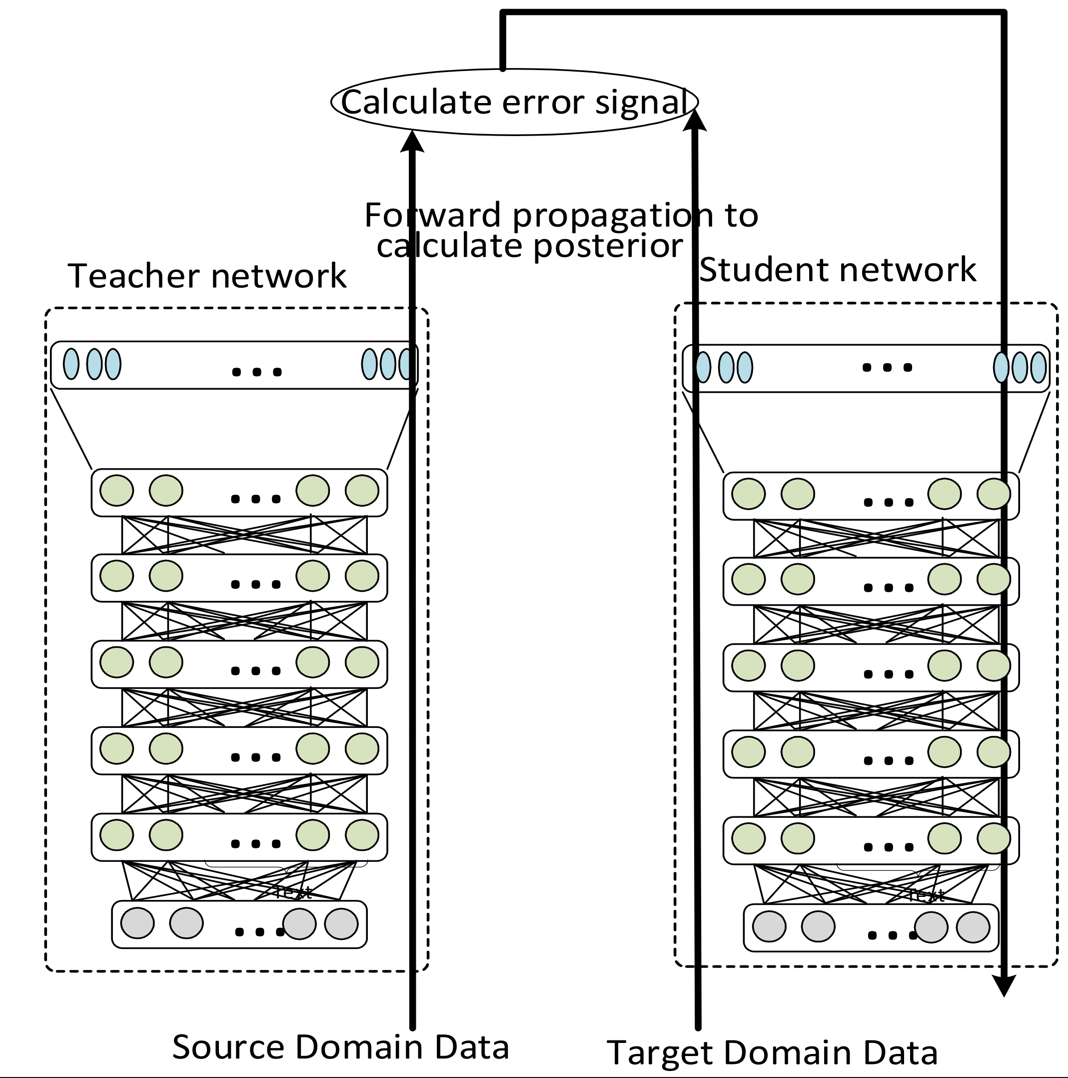
J. Li, R. Zhao, etc. "Learning small-size DNN with output-distribution-based criteria," In Proc. *Interspeech*, 2014.

$$-\sum_f \sum_i P_T(s_i|x_f) log P_S(s_i|x_f)$$

- T/S domain adaptation

J. Li, M. Seltzer, etc. "Large-scale domain adaptation via teacher student learning," in Proc. *Interspeech*, 2017.

$$-\sum_f \sum_i P_T(s_i|x_{src,f}) log P_S(s_i|x_{tgt,f})$$



3. ASR Experiments

- Source training data:
  - 3.4k hours of labeled US-English close-talk Cortana audio.
  - 25k hours of unlabeled US-English close-talk Cortana audio.
  - 300 hours labeled live far-field audio.
- Teacher Model:
  - LSTM-RNN: 4-layer uni-LSTM-P, 1024 memory units and projection layer with 512 nodes. Output layer has 9404 nodes, modeling senones.
  - Singular value decomposition (SVD) and frame skipping are used to reduce cost.
  - Trained with labeled data with CE and then sequence discriminative training.

Model	WER (%)	
	Playback	**Live**
Close-talk	47.34	23.81
CE (3.4k hours single channel **simulation**)	21.22	14.30
T/S (3.4k hours single channel simulation)	18.79	14.19
T/S (**25k** hours **unlabeled** single channel simulation)	16.61	12.98
T/S (25k hours unlabeled **beamformed** simulation)	15.26	11.96
T/S (25k hours unlabeled beamformed simulation) + 3.4k hours simulation **sequence training**	12.97	11.20
T/S (25k hours unlabeled beamformed simulation) + 3.4k hours simulation + **300 hours live** sequence training	13.38	10.20

4. KWS Experiments

- Source training data:
  - 760 hours labeled close-talk Cortana audio, half with "Hey Cortana" and half without.
  - 600 hours labeled far-field live data or 940 hours unlabeled far-field live data.
- Large-size model used as teacher (24M parameters):
  - LSTM-RNN-CTC: 5-layer uni-LSTM-P, 1024 memory units and projection layer with 512 nodes. Output layer has 5 nodes, modeling Hey, Cortana, silence, garbage, and blank.
- Small-size model (0.9M parameters):
  - 3-layer uni-LSTM-P, 256 memory units and projection layer with 128 nodes, with SVD.

Data / Model	simulation	simulation + **600-hour live labeled**	simulation + **940-hour live unlabeled**
large-size CTC	5.39	1.60	-
small-size CTC	11.28	1.94	-
small-size CTC with **T/S**	7.61	1.73	1.59

The FA rates (%) of **KWS models** operating at the 96% CA rate.

5. Conclusions

- **Simulating** far-field data, especially the beamformed one, is very helpful to improving the accuracy of ***real test data***.
- **T/S learning** effectively used ***unlabeled*** data to improve the student model.
- The final **AM** improves the baseline by with **72.60%** and **57.16%** relative WER reduction on play-back and live far-field data.
- The **small-size CTC KWS** model trained with unlabeled data using **T/S learning** has the ***same*** performance as the large-size CTC KWS model, but with only ***1/27 foot-print***.