NATIONAL ENGINEERING LABORATORY
FOR SPEECH AND LANGUAGE INFORMATION PROCESSING

A Speaker-Dependent Deep Learning Approach to Joint Speech Separation and Acoustic Modeling for Multi-Talker Automatic Speech Recognition

Tu Yan-Hui[1] , Du Jun[1], Dai Li-Rong[1] and Lee Chin-Hui[2]

[1]University of Science and Technology of China

[2]Georgia Institute of Technology, USA

University of Science and Technology of China
USTC iFLYTEK CO.,LTD.

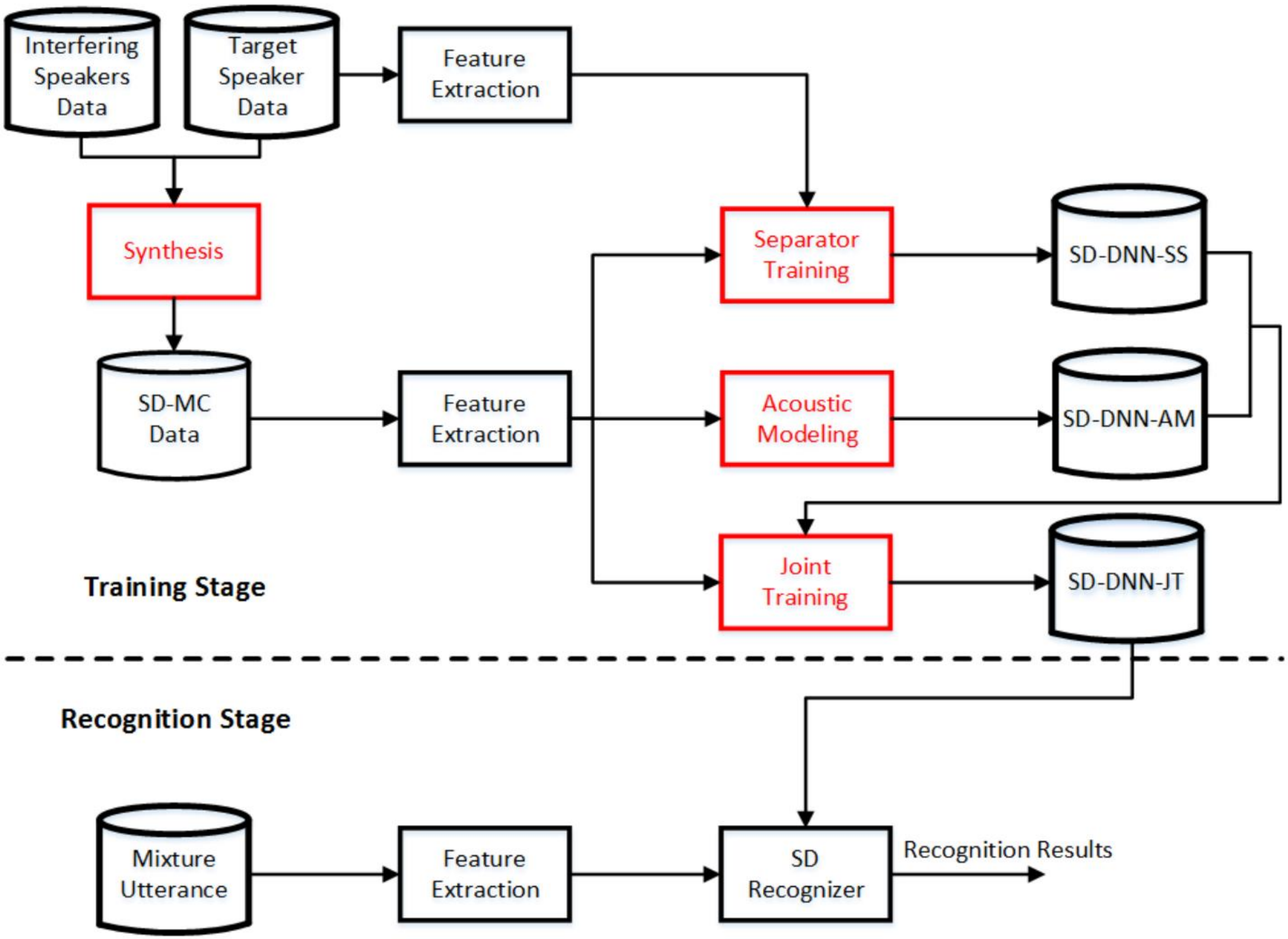
Outline

- Motivation
- Proposed approach
- Experiments
- Conclusions

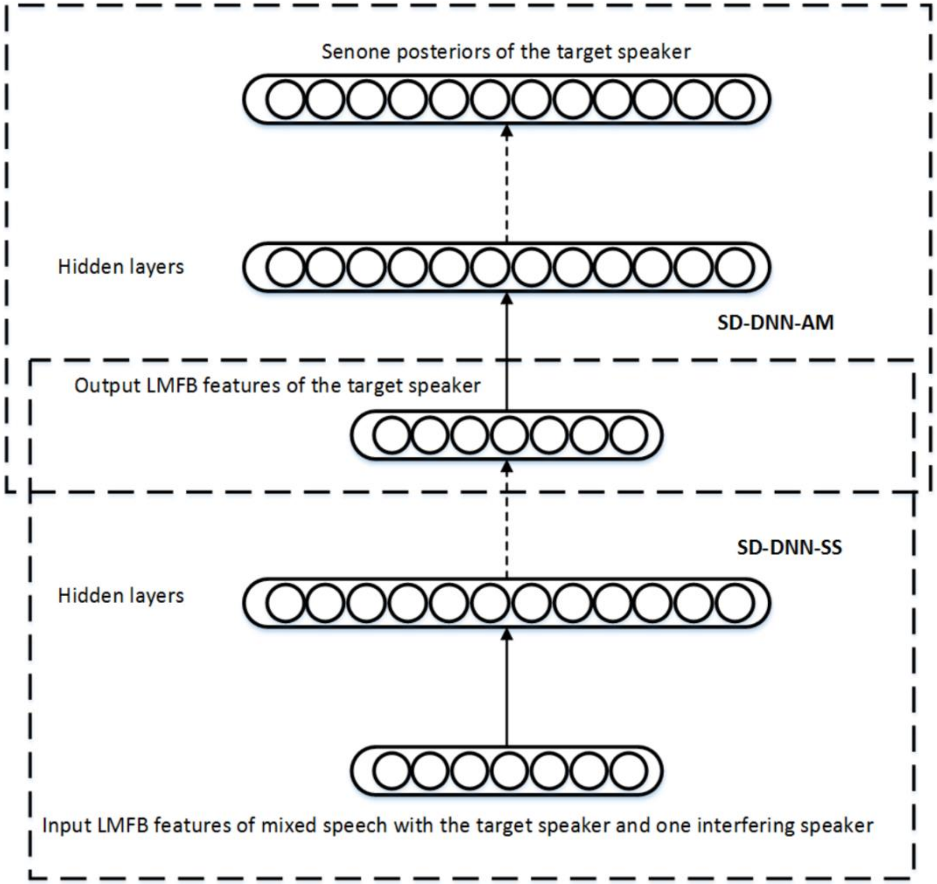
Motivation

- DNN-based separation method is better than GMM
  - Jun Du, Yan-Hui Tu, Yong Xu, Li-Rong Dai and Chin-Hui Lee, "Speech Separation of A Target Speaker Based on Deep Neural Networks.", ICSP(2014)
- The separated signals can improve SI ASR system performance
  - Yan-Hui Tu, Jun Du, Li-Rong Dai and Chin-Hui Lee, "Speech Separation based on signal-noise-dependent deep neural networks for robust speech recognition.", ICASSP(2015).
- SD recognition system in multi-talker scenarios
  - The proposed speaker-dependent approach is quite robust to the interference of a competing speaker even in low target-to-masker ratio (TMR) conditions

SD Recognition: System Overview



Joint training for SD ASR



Joint training

Step 1: Train a SD-DNN-SS to eliminate the interferences of other speakers.

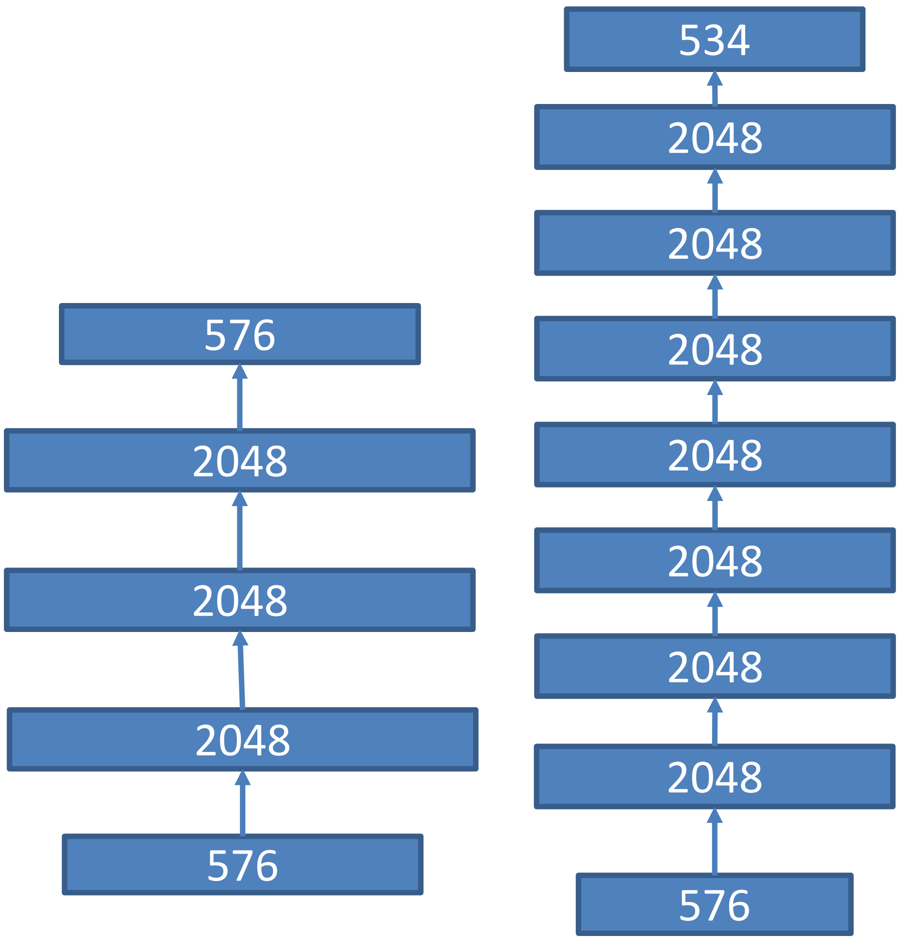
Step 2: Train a SD-DNN-AM with the SD-MC training set as an initial model.

Step 3: Concatenate SD-DNN-SS and SD-DNN-AM as one SD-DNN-JT and fine-tune all the parameters of SD-DNN-JT via the CE criterion.

Experimental Setup

- **SSC corpus**
  - **training set**: 34 speakers(18 males and 16 females), 500 utterances for each speaker
  - **test set**: two-speaker mixtures at a range of signal-to-noise ratios (SNR) from -9dB to 6dB with an increment of 3dB
- **Train set**
  - 500 utterances for each speaker were as our target speech
  - The interfering speakers for each speaker were randomly selected from the 34 speakers except the target speaker
- **Fixed grammar(six parts)**
  - Command, color, preposition, letter, number, adverb

DNN Configurations



SD-DNN-SS SD-DNN-AM

Sampling rate : 16 kHz
LMFB : 64 dimensions

SD-DNN-SS:
576=64*9
9 frames input context expansion
2048 for three hidden layers

SD-DNN-AM:
2048 for seven hidden layers
soft-max output layer： 534

Experimental Results and Analysis (1/4)

- **Experiments under Clean-condition Training**

Table 1: WER comparison of SI and SD DNN-HMM systems under clean-condition training on the test set of all 34 target speakers with different TMRs.

System	6dB	3dB	0dB	-3dB	-6dB	-9dB
SI	32.8	47.1	63.3	76.9	84.2	90.9
SD	31.5	45.6	59.1	72.8	82.3	89.8

Training set
34 target speaker : 18 male and 16 woman
Size of SI system : 17000utterances(500*34)
Size of SD systems : 500 utterance / per model

Conclusion:
Although the SD system slightly outperformed the SI system, both systems yielded very poor performance, especially under low TMRs, which implied the necessity of multi-condition training.

Experimental Results and Analysis (2/4)

- **Experiments under Multi-condition Training**

Table 2: WER comparison of SD DNN-HMM systems under clean-condition (Clean) and multi-condition (Multi) training on the test set of 6 selected target speakers with different TMRs.

System	6dB	3dB	0dB	-3dB	-6dB	-9dB	Avg.
Clean	32.3	47.2	61.9	78.3	85.2	92.3	66.2
Multi	19.7	23.9	25.4	28.2	31.7	39.4	28.1

Training set
6 target speakers: 3 male and 3 woman
33 interfering speakers for each target
TMR : -9 dB to 6 dB with an increment of 3 dB
Size : 3000(500*6) utterances for each speaker

Conclusion:
SD multi-condition training significantly reduced the average WER from 66.2% in clean-condition training to 28.1%, yielding a relative WER reduction of 57.6%.

Experimental Results and Analysis (3/4)

- **Experiments under Multi-condition Training**

Table 3: WER comparisons of SD DNN-HMM systems on the test set of 6 selected target speakers under multi-condition training with different amounts of training data (3000, 102000, and 357000 training utterances for S1, S2 and S3, respectively).

System	6dB	3dB	0dB	-3dB	-6dB	-9dB	Avg.
S1	19.7	23.9	25.4	28.2	31.7	39.4	28.1
S2	6.3	7.1	9.1	9.8	10.6	11.2	9.1
S3	2.1	2.8	3.5	3.5	4.3	6.3	3.8

Training set

S1:

TMR : -9 dB to 6 dB with an increment of 3 dB

3000(500*6) utterances for each speaker

S2:

Each clean utterance of the target speaker was repeatedly 34 times corresponding to all 34 speakers

102000(500*34*6) utterances for each speaker

S3:

TMR : -10 dB to 10 dB with an increment of 1 dB

357000(500*34*21) utterances for each speaker

Conclusion:

WERs for all TMRs were significantly reduced with the increase of training data amounts.

Experimental Results and Analysis (4/4)

- **Experiments with Jointly Trained DNN Models**

Table 4: WER comparison of the multi-condition trained SD-DNN-AM system (Multi) and the jointly trained SD-DNN-JT system (Joint) on the test set of 6 selected target speakers.

System	6dB	3dB	0dB	-3dB	-6dB	-9dB	Avg.
Multi	2.1	2.8	3.5	3.5	4.3	6.3	3.8
Joint	2.1	2.1	2.8	3.5	3.5	5.6	3.3
[1]	7	8.5	9.2	11.3	12.7	16.9	10.9

Conclusion:

In comparison to a WER of 10.9% obtained with the proposed pre-processing DNN approach in [1], a relative WER reduction of 69.7% could be observed.

Conclusion

- We have proposed a novel speaker-dependent approach for single-channel automatic speech recognition of mixture speech in a multi-talker scenario.

- The feasibility of designing a SD recognizer on portable devices will also be explored in the mobile internet era.

Thank you!

Q&A