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

Tu Yan-Hui¹, Du Jun¹, Dai Li-Rong¹ and Lee Chin-Hui²
¹University of Science and Technology of China
²Georgia Institute of Technology, USA
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

• 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

The system consists of two stages: Training Stage and Recognition Stage.

**Training Stage**:
- **Interfering Speakers Data** and **Target Speaker Data** lead to Feature Extraction.
- **Synthesis** is followed by **SD-MC Data**.
- Feature Extraction leads to **Separator Training** and **Acoustic Modeling**.
- **Joint Training** is connected to SD-DNN-JT.

**Recognition Stage**:
- **Mixture Utterance** undergoes Feature Extraction.
- SD Recognizer outputs Recognition Results.
**Joint training for SD ASR**

**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:**
- 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

**Sampling rate:** 16 kHz
**LMFB:** 64 dimensions

**Sampling Rate:** 16 kHz
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.

<table>
<thead>
<tr>
<th>System</th>
<th>6dB</th>
<th>3dB</th>
<th>0dB</th>
<th>-3dB</th>
<th>-6dB</th>
<th>-9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI</td>
<td>32.8</td>
<td>47.1</td>
<td>63.3</td>
<td>76.9</td>
<td>84.2</td>
<td>90.9</td>
</tr>
<tr>
<td>SD</td>
<td>31.5</td>
<td>45.6</td>
<td>59.1</td>
<td>72.8</td>
<td>82.3</td>
<td>89.8</td>
</tr>
</tbody>
</table>

Training set
34 target speaker: 18 male and 16 woman
Size of SI system: 17000 utterances (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.

<table>
<thead>
<tr>
<th>System</th>
<th>6dB</th>
<th>3dB</th>
<th>0dB</th>
<th>-3dB</th>
<th>-6dB</th>
<th>-9dB</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>32.3</td>
<td>47.2</td>
<td>61.9</td>
<td>78.3</td>
<td>85.2</td>
<td>92.3</td>
<td>66.2</td>
</tr>
<tr>
<td>Multi</td>
<td>19.7</td>
<td>23.9</td>
<td>25.4</td>
<td>28.2</td>
<td>31.7</td>
<td>39.4</td>
<td>28.1</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>System</th>
<th>6dB</th>
<th>3dB</th>
<th>0dB</th>
<th>-3dB</th>
<th>-6dB</th>
<th>-9dB</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>19.7</td>
<td>23.9</td>
<td>25.4</td>
<td>28.2</td>
<td>31.7</td>
<td>39.4</td>
<td>28.1</td>
</tr>
<tr>
<td>S2</td>
<td>6.3</td>
<td>7.1</td>
<td>9.1</td>
<td>9.8</td>
<td>10.6</td>
<td>11.2</td>
<td>9.1</td>
</tr>
<tr>
<td>S3</td>
<td>2.1</td>
<td>2.8</td>
<td>3.5</td>
<td>3.5</td>
<td>4.3</td>
<td>6.3</td>
<td>3.8</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>System</th>
<th>6dB</th>
<th>3dB</th>
<th>0dB</th>
<th>-3dB</th>
<th>-6dB</th>
<th>-9dB</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi</td>
<td>2.1</td>
<td>2.8</td>
<td>3.5</td>
<td>3.5</td>
<td>4.3</td>
<td>6.3</td>
<td>3.8</td>
</tr>
<tr>
<td>Joint</td>
<td>2.1</td>
<td>2.1</td>
<td>2.8</td>
<td>3.5</td>
<td>3.5</td>
<td>5.6</td>
<td>3.3</td>
</tr>
<tr>
<td>[1]</td>
<td>7</td>
<td>8.5</td>
<td>9.2</td>
<td>11.3</td>
<td>12.7</td>
<td>16.9</td>
<td>10.9</td>
</tr>
</tbody>
</table>

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