Speaker Diarization with Enhancing Speech for the First DIHARD Challenge

Lei Sun\textsuperscript{1}, Jun Du\textsuperscript{1}, Chao Jiang\textsuperscript{2}, Xueyang Zhang\textsuperscript{2}, Shan He\textsuperscript{2}, Bing Yin\textsuperscript{2}, and Chin-Hui Lee\textsuperscript{3}

\textsuperscript{1}University of Science and Technology of China
\textsuperscript{2}IFLYTEK Research
\textsuperscript{3}Georgia Institute of Technology
Outline

- Introduction
  - Background
  - Our previous work
- System description
- Results
- Conclusion and future work
Introduction

• Background
  – State-of-the-art diarization systems should be promoted to more challenging domains (JSALT 2017).
  – Current approaches performs poorly in some problems, such as dynamic environmental noise, overlapping speech, etc.

• Our previous work
  – Deep learning based denoising method has stronger potentials in coping with realistic noisy environments than traditional approaches. (A NOVEL LSTM-BASED SPEECH PREPROCESSOR FOR SPEAKER DIARIZATION IN REALISTIC MISMATCH CONDITIONS, ICASSP2018)
  – Limited conditions: oracle SAD, not the state-of-the-art diarization system, few training data.

• DIHADR Challenge
System pipeline:
1. Front-end speech preprocessing: speech denoising, SAD
2. Short-term diarization: BIC segmentation and clustering
3. Long-term diarization: UBM I-Vector, CNN I-Vector
4. Score fusion: PLDA score, Cosine score
5. Re-segmentation: Viterbi
Speech Denoising

DIHARD data sources (Red indicates noisy condition):
Seedlings, SupremeCourt, ADOS, DCIEM, YouthPoint, SLX, RT-04S, LibriSpeech, VAST

Denoising model: **LSTM-based densely connected progressive learning network**

Simulated data:

- **SNR of Input:** -5 dB, 0 dB and 5 dB
- **SNR of Target 1:** 5 dB, 10 dB and 15 dB
- **SNR of Target 2:** 15 dB, 20 dB and 25 dB
- **Final Target 3:** Clean Log Power Spectrum (LPS) + Ideal Ratio Mask (IRM)
Two-pass diarization

Short-term diarization (Given the speech segments from SAD):

1. Segmentation (change point detection):
   - Sliding window (2s);
   - Use a single gaussian model;
   - Hypothesis Testing:
     - H1: x, y are from the same speaker
     - H2: x, y are from different speakers
   - Metric: BIC distance

2. Clustering:
   - Hierarchical Agglomerative Clustering (HAC);
   - Use a single gaussian model;
   - Metric: BIC distance;
   - Stop condition: default MAX_class_number = (total_time)/10
Two-pass diarization

Long-term diarization (Given the long segments from above):

1. UBM i-vector
   - $M = m + Tw$
   - config: 1024 gaussians, 400 factors
   - Whitening; Length-normalization;
   - scoring: PLDA(100 factors)
   - data: VoxCeleb

2. CNN i-vector
   - input: 512 frames of 64-dim filterbank
   - output: 512 dimensional vector
   - Pre-train: softmax loss
   - Fine-tune: triplet loss
   - scoring: cosine distance
   - data: larger home-made corpus
Score fusion

Long-term diarization:

Score fusion function:

\[ \text{Score} = \text{PLDA}(U_{vec_i}, U_{vec_j}) + w \times \cos(C_{vec_i}, C_{vec_j}) \]

\[ \text{Score Fusion} \]
\[ \text{AHC} \]
\[ \text{Re-segmentation} \]
\[ \text{Diarization Results} \]
Database

Speech denoising model:
  Clean Speech: WSJ0(English), 863 Program(Chinese)
  Reading style
  Noise: 115 kinds
  Simulated data size: 400 hours

Speech Activity Detection model:
  Home-made realistic speech from complicated acoustic environments
  600 hours
  Human annotations

UBM based i-vector extractor:
  VoxCleb corpus
  100,000 utterances from 1,251 celebrities

CNN based i-vector extractor:
  Home-made corpus
  about 5,800 hours from 38,000 celebrities
Subsidiary details

Deep-learning based model training details:

**SAD**
- Feature: 39-dimensional PLP features
- Input: an input context of 5 neighbouring frames (±2)
- Target: binary classification
- Architecture: 195-256-128-2, DNN

**Denoising**
- Feature: 257-dimensional LPS features
- Input: an input context of 7 neighbouring frames (±3)
- Target: progressively designed in increasing SNRs
- Architecture: 1799-1024-257-1024-257-1024-257*2, LSTM

System design summary:
- Attach equal importance to both front-end and back-end
- Use appropriate features or strategy in different stages
- No adaptation to a specific corpus or environmental condition
Results

Track 1: Using oracle SAD

The DER results of UBM based i-vector system, which uses PLDA score for clustering:

<table>
<thead>
<tr>
<th>DER(%)</th>
<th>Track1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>Miss</td>
</tr>
<tr>
<td>Original</td>
<td>8.50</td>
</tr>
<tr>
<td>Denoised</td>
<td>8.50</td>
</tr>
<tr>
<td>Retrained</td>
<td>8.50</td>
</tr>
</tbody>
</table>

Performance on each specific corpus:

<table>
<thead>
<tr>
<th>DER(%)</th>
<th>Seedlings</th>
<th>SCOTUS</th>
<th>DCIEM</th>
<th>ADOS</th>
<th>YP</th>
<th>SLX</th>
<th>RT04S</th>
<th>LIBRIBOX</th>
<th>VAST</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>40.58</td>
<td>7.65</td>
<td>7.27</td>
<td>22.9</td>
<td>3.32</td>
<td>18.88</td>
<td>36.55</td>
<td>0.6</td>
<td>36.46</td>
<td>20.26</td>
</tr>
<tr>
<td>Denoised</td>
<td>37.86</td>
<td>7.10</td>
<td>7.58</td>
<td>21.73</td>
<td>3.12</td>
<td>16.94</td>
<td>35.93</td>
<td>1.9</td>
<td>36.08</td>
<td>19.68</td>
</tr>
</tbody>
</table>

1. Denoised speech can improve performance in all noisy conditions;
2. For clean scenes, it have performance loss, especially in LIBRIBOX, from 0.6 to 1.9 (34.17s to 107.53 among 5742.37 scored time).
3. It’s a trade-off decision in terms of overall results.
### Results

**Track 2: Diarization from scratch**

<table>
<thead>
<tr>
<th>DER(%)</th>
<th>Track2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Speech</td>
</tr>
<tr>
<td>Original</td>
<td>18.60</td>
</tr>
<tr>
<td>Denoised</td>
<td>16.50</td>
</tr>
<tr>
<td>Retrained</td>
<td>16.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DER(%)</th>
<th>Seedlings</th>
<th>SCOTUS</th>
<th>DCIEM</th>
<th>ADOS</th>
<th>YP</th>
<th>SLX</th>
<th>RT04S</th>
<th>LIBRIBOX</th>
<th>VAST</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>65.72</td>
<td>15.75</td>
<td>17.49</td>
<td>39.46</td>
<td>12.97</td>
<td>31.45</td>
<td>49.42</td>
<td>8.77</td>
<td>53.16</td>
<td>33.2</td>
</tr>
<tr>
<td>Denoised</td>
<td>61.89</td>
<td>15.25</td>
<td>20.02</td>
<td>34.55</td>
<td>12.06</td>
<td>27.81</td>
<td>45.00</td>
<td>9.07</td>
<td>44.64</td>
<td>30.4</td>
</tr>
</tbody>
</table>

**In track 2, denoised speech can bring bigger improvements after introducing SAD module;**

![Image of original speech](image1)

![Image of enhanced speech using 36h model](image2)

![Image of enhanced speech using 400h model](image3)
1. These two scores have some complementarity to enhance the final performance.
2. Score fusion is more effective when using oracle SAD in Track1.

**Final results on evaluation set:**

<table>
<thead>
<tr>
<th>DER (in %)</th>
<th>USTC-iFlytek (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track1</td>
<td>24.56</td>
</tr>
<tr>
<td>Track2</td>
<td>36.05</td>
</tr>
</tbody>
</table>
Conclusion and Future Work

• Front-end can improve overall results, especially in realistic data
  
  Future: We need a deeply connected front-end and back-end

• Overlapping speech processing is not involved here, and it accounts for the largest proportion of errors
  
  Future: detection and attribution
Thanks for ...

- Organizers
- Reviewers
- All the audience