A Unified Deep Modeling Approach to Simultaneous Speech Dereverberation and Recognition for The REVERB Challenge

Bo Wu, Kehuang Li, Zhen Huang, Sabato Marco Siniscalchi, Minglei Yang, Chin-Hui Lee
School of ECE, Georgia Tech
chl@ece.gatech.edu
Outline and Talk Agenda

- A deep learning and big data perspective on DSP
- DNN speech enhancement (SE): T-SALP, 01/14
- DNN source separation (SS), T-SALP, 08/16
- DNN de-reverberation: T-SALP, 01/17
  - Reverberant-Time-Aware (RTA-DNN)
  - Simulation data can be very powerful
- Robust automatic speech recognition (ASR)
  - Separate and integrated pre- and post-processing
  - Best ASR performance in REVERB Challenge
  - Good speech quality and top ASR performance
- Conclusion and future work: “A New Hope”
DNN Based Spectral Mapping

\[ X = F(Y) + E \]

1. Pre-training

- \( W_1 + \varepsilon_1 \)
- \( W_2 + \varepsilon_2 \)
- \( W_3 + \varepsilon_3 \)

2. Fine Tuning

- \( W_4 \)

Y

(Input with multiple frames of noisy speech features)

(Output with a single frame of clean speech features)

Dereverb
Baseline & Reverberant-Time-Aware (RTA) DNN Based Speech Dereverberation

The RT60 Estimator was adopted from Keshavarz et al, T-ASLP, 2012

R: size of the overlap in framing, and N: the no. of context frames

Baseline DNN: in black
• RTA-DNN by adding red
**PESQ Comparisons for Three Levels of Environmental Awareness**

<table>
<thead>
<tr>
<th>RT60 (s)</th>
<th>0.10</th>
<th>0.20</th>
<th>0.30</th>
<th>0.40</th>
<th>0.50</th>
<th>0.60</th>
<th>0.70</th>
<th>0.80</th>
<th>0.90</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSA-DNN-oracle</td>
<td>3.75</td>
<td>3.39</td>
<td>3.16</td>
<td>3.03</td>
<td>2.91</td>
<td>2.82</td>
<td>2.74</td>
<td>2.66</td>
<td>2.59</td>
<td>2.52</td>
</tr>
<tr>
<td>ACA-DNN-oracle</td>
<td>3.52</td>
<td>3.25</td>
<td>3.06</td>
<td>2.94</td>
<td>2.84</td>
<td>2.78</td>
<td>2.72</td>
<td>2.67</td>
<td>2.62</td>
<td>2.56</td>
</tr>
<tr>
<td>RTA-DNN-oracle</td>
<td>3.77</td>
<td>3.41</td>
<td>3.17</td>
<td>3.05</td>
<td>2.95</td>
<td>2.87</td>
<td>2.80</td>
<td>2.73</td>
<td>2.66</td>
<td>2.61</td>
</tr>
</tbody>
</table>

By combining FSA and ACA parameters, RTA-DNN gets best PESQ

<table>
<thead>
<tr>
<th>RT60 (s)</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>R (ms)</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>N</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

R=16 msec (half frame size) and N=5 in he baseline DNN system
Even with 4-hr training, RTA-DNN performs well at all RT60s
2014 REVERB Challenge: SE TEST

Cepstrum Distance \textit{lower is better}

Log Likelihood Ration Distance - \textit{The lower, the better}

fwSegSNR – \textit{higher is better}

SRMR– The higher, the better (only target signal is needed)
Good DSP will Lead to Accurate ASR

Reverberant speech, No pre-processing, ASR errors

Mismatched dereverberation, less ASR errors, worst PESQ

Matched dereverberation, no ASR errors, Best PESQ
Separate versus Integrated Modeling

Integrated

Separated
## Pre-Processing Followed by Post-Processing (Single-Channel)

<table>
<thead>
<tr>
<th>Training Scheme</th>
<th>Sys.</th>
<th>Training Data</th>
<th>Eval. Data</th>
<th>Room 1</th>
<th>Room 2</th>
<th>Room 3</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cln. Rev. Base RTA</td>
<td>Near</td>
<td>Far</td>
<td>Near</td>
<td>Far</td>
<td>Near</td>
</tr>
<tr>
<td>CE</td>
<td>S1</td>
<td>X</td>
<td>Rev.</td>
<td>6.95</td>
<td>8.11</td>
<td>8.68</td>
<td>14.61</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>X X X</td>
<td>Base</td>
<td>6.15</td>
<td>6.98</td>
<td>7.83</td>
<td>12.91</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>X X X X X</td>
<td>RTA</td>
<td>6.25</td>
<td>6.45</td>
<td>7.62</td>
<td>12.46</td>
</tr>
<tr>
<td>sMBR</td>
<td>S4</td>
<td>X X X</td>
<td>Base</td>
<td>6.10</td>
<td>6.56</td>
<td>7.46</td>
<td>11.36</td>
</tr>
<tr>
<td></td>
<td>S5</td>
<td>X X X X X</td>
<td>RTA</td>
<td>5.70</td>
<td>6.00</td>
<td>7.10</td>
<td>11.50</td>
</tr>
</tbody>
</table>

Baseline REVERB: 42.81%, Best system in 06/14: 37.88%

Base without RTA-DNN: 27.56%, with RTA-DNN: 25.70%

Best single system performance with separate training: 8.75%
## Integrated Pre-Processing and Post-Processing (Single-Channel)

<table>
<thead>
<tr>
<th>System Configurations</th>
<th>Room 1</th>
<th>Room 2</th>
<th>Room 3</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acoustic Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SMBR CD-DNN-HMM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>Joint Training + 3-gram</td>
<td>5.05</td>
<td>5.86</td>
<td>6.00</td>
</tr>
<tr>
<td>S7</td>
<td>Joint Training + RNN-LM</td>
<td>3.11</td>
<td>4.01</td>
<td>4.42</td>
</tr>
<tr>
<td>S10</td>
<td>Joint Training + LSTM-LM</td>
<td>2.49</td>
<td>2.80</td>
<td>3.18</td>
</tr>
<tr>
<td>S8</td>
<td>S4 + RNN-LM</td>
<td>4.30</td>
<td>4.71</td>
<td>5.45</td>
</tr>
<tr>
<td>S11</td>
<td>S4 + LSTM-LM</td>
<td>2.90</td>
<td>3.39</td>
<td>4.29</td>
</tr>
<tr>
<td>S9</td>
<td>S5 + RNN-LM</td>
<td>3.74</td>
<td>4.40</td>
<td>5.48</td>
</tr>
<tr>
<td>S12</td>
<td>S5 + LSTM-LM</td>
<td>2.71</td>
<td>3.15</td>
<td>3.95</td>
</tr>
<tr>
<td><strong>System Combination [26]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S7 + S8 + S9</td>
<td>3.29</td>
<td>3.71</td>
<td>4.18</td>
<td>6.43</td>
</tr>
<tr>
<td>S10 + S11 + S12</td>
<td>2.57</td>
<td>2.71</td>
<td>3.46</td>
<td>5.50</td>
</tr>
</tbody>
</table>

NTT in 01/16 (More Data): 5.2% (1-ch), 4.40% (2-ch), 4.20% (8-ch)
Result Summary: REVERB Challenge

- Clean Training: WSJ0, 17.5 hours, 7861 utterances
- MC Training: 24 measured RIRs, SNR at 20dB
- Testing SimData: corrupted WSJCAM0 training set
  - 3 rooms (1, 2, 3) with different volumes (small, medium, and large), at two distances between a speaker and a microphone array (near = 50 cm and far = 200 cm).
  - Measured stationary ambient noise signals with a signal-to-noise ratio (SNR) of 20 dB has been added, 16KHz
  - 4.8 hours (2176 utterances)
- Baseline: WER at 42.81%, with clean CD-GMM-HMM
- Top Participant: 37.88%, clean CD-GMM-HMM (June 2014)
- Top NTT (01/16): 5.2% (1-ch), 4.40% (2-ch), 4.20% (8-ch) by adding lots of extra speech training data with simulation

GT: single channel at best 4.10% (reporting here)
Conclusion and Future Work

1. Combining deep learning and big data: a new machine learning paradigm to achieve new capabilities for classical signal processing problems

2. A large training set: learning rich regression structure
   - Even simulation data can be very useful

3. For dereverberation the results are amazingly good

4. A New Hope: with proper pre-processing followed by integrated post-processing leading to robust ASR!

5. Multi-channel extension and noise simulation with multi-objective learning are highly desirable