Student-Teacher Learning for BLSTM Mask-based Speech Enhancement
Aswin Shanmugam Subramanian, Szu-Jui Chen, Shinji Watanabe
Center for Language and Speech Processing, Johns Hopkins University

Abstract
- Neural network mask-based beamforming techniques have improved the performance of multichannel noise robust ASR significantly.
- Spectral masks have not been helpful in the single-channel case.
- We propose a student-teacher learning paradigm for mask estimation to fill out the gap between single-channel and multichannel speech enhancement.

BLSTM Masking Network
[Heymann*, 2016]

<table>
<thead>
<tr>
<th>Layer</th>
<th>Activation</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>BLSTM</td>
<td>Tanh</td>
<td>513</td>
</tr>
<tr>
<td>Feedforward 1</td>
<td>ReLU</td>
<td>513</td>
</tr>
<tr>
<td>Feedforward 2</td>
<td>clipped ReLU</td>
<td>513</td>
</tr>
</tbody>
</table>

Table 1: Masking Network Architecture

- $Y = (t, b)_{t=1 \cdots T}$: sequence of $T$-length noisy speech magnitude spectra
- $\text{IBM}_x(t, b) \in \{0, 1\}$ and $\text{IBM}_s(t, b) \in \{0, 1\}$ at each time-frequency bin $(t, b)$: ideal binary speech and noise mask target respectively
- $\phi_x(t, b) \in \{0, 1\}$ and $\phi_s(t, b) \in \{0, 1\}$ at each time-frequency bin $(t, b)$: predicted speech and noise mask respectively
- $\text{loss} = \text{loss}_\phi + \text{loss}_\psi$
- $\lambda_1 \chi \phi + \lambda_2 \chi \phi$ for real
- $\lambda_3 \chi \phi$ for simulation

Mask-Based Beamformer Cnt’d
- $\Phi_x(b) = \sum_{t \in (X, N)}^{n_{max}(b)} \phi_x(t, b)y(t, b)_{v \in M}$, where $v \in [X, N]$,
- $\phi_x(t, b)\in \mathbb{C}^M$ and $\Phi_x(b)\in \mathbb{C}^M$, $\Phi_x(b) = \text{argmax}_b \Phi_x(b)$

Student Model (Additional Loss Term):

$$\text{loss}_\text{st} = \frac{1}{T + B} \sum_{t \in [T \cup B]} \text{CE}(\text{IBM}_s(t, b), \phi_s(t, b))$$

Teacher Model:

$$\text{loss} = \frac{1}{T + B} \sum_{t \in [T \cup B]} \text{CE}(\text{IBM}_x(t, b), \phi_x(t, b))$$

Student Model with Real Data:

$$\text{loss} = \frac{1}{T + B} \sum_{t \in [T \cup B]} \text{CE}(\text{IBM}_s(t, b), \phi_s(t, b))$$

Experiments
- Dataset: 1 channel track in CHiME-4
- Training: 1000 (real) + 738 (simulated) - use all 6 ch data
- Dev & Test: 3280 & 2640 respectively - equal real and simulated noisy utterances
- HMM-GMM ASR system - Kaldi CHiME4 recipe

Figure: WER vs Epoch for Different Parameter Combinations

- Best validation loss - not necessarily gives best word error rate (WER)
- Choosing the epoch based on the WER of the development data seems to be a better criterion.

Table 2: WER of HMM-GMM ASR System

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Train data (ASR)</th>
<th>BLSTM Mask</th>
<th>WER Dev (%)</th>
<th>WER Test (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st channel</td>
<td>all 6 ch noisy</td>
<td>Baseline</td>
<td>21.40</td>
<td>23.22</td>
</tr>
<tr>
<td>2nd channel</td>
<td>all 6 ch noisy</td>
<td>Baseline</td>
<td>28.99</td>
<td>30.41</td>
</tr>
<tr>
<td>3rd channel</td>
<td>all 6 ch noisy</td>
<td>Teacher</td>
<td>24.91</td>
<td>26.00</td>
</tr>
<tr>
<td>4th channel</td>
<td>all 6 ch noisy</td>
<td>Student</td>
<td>23.34</td>
<td>23.11</td>
</tr>
<tr>
<td>5th channel</td>
<td>all 6 ch noisy</td>
<td>Student</td>
<td>23.42</td>
<td>23.55</td>
</tr>
<tr>
<td>6th channel</td>
<td>all 6 ch noisy</td>
<td>Student</td>
<td>32.64</td>
<td>32.88</td>
</tr>
</tbody>
</table>

Table 2 (rows 4-7):
- The training data for ASR not enhanced
- Student models performed better than both the teacher model and baseline
- Student models don’t perform better than the non-enhanced noisy speech

Table 2 (rows 8-10):
- 5th channel data enhanced using the baseline masking model is included as part of ASR training
- The performance improved significantly compared to using the original noisy data in the all conditions when the development and evaluation data was enhanced using our best student models (rows 9 and 10).
- WER improvement for the real test set was observed when real training data was included while training the mask (row 10).

Speech Enhancement Scores
- Masking gives significantly better scores in all four metrics.
- No considerable difference in the scores amongst the masking models.

Conclusion
- The proposed student-teacher paradigm improved the performance of a GMM-HMM ASR system from both original noisy speech and the baseline masking.
- Our preliminary experiments on a strong ASR backend improved performance over the baseline masking but not the original noisy data.