Introduction

Motivation
- accuracy of multi-talker distant conversational ASR is still poor
- problems: competing speakers, reverberation, background noise, speech disfluency etc.

Context
- speech enhancement improves word error rate (WER), but is typically applied on the test data only
- it is generally agreed upon that enhancement in ASR training would reduce the acoustic variability
- training data is often artificially increased by adding more degraded speech to it

CHiME-5 Challenge
- distant multi-microphone conversational speech recognition challenge in everyday home environments [1]
- corpus description:
  - 20 dinner party recordings (aprox. 2 hours each)
  - 4 participants and 3 locations (kitchen, dining, and living room)
  - 6 x 4-channel distant recording devices (‘U’ set)
  - in-ear binaural microphones (‘W’ set)
- recording devices not time synchronized
- single (reference) U device track and multiple U device track
- baseline CHiME-5 system achieved roughly 80% WER

Contributions of this work
- study on the effectiveness of acoustic enhancement in ASR training and test for CHiME-5
- state-of-the-art single-system for CHiME-5

Guided Source Separation (GSS)
- blind source separation method adapted to CHiME-5 [2]
- spatial mixture model:
  - complex Angular Central Gaussian Mixture Model (cACGMM)
- cACGMM parameters and posterior probabilities of each speaker being active estimated by EM algorithm
- mask based beamforming (Fig. 1)

Experiments & Results
- CHiME-5 corpus was used for ASR training and test (Table 1)
- GMM-HMM alignment model
- acoustic model topology:
  - 6 x CNNs + 9 x TDNNFs
  - speed perturbation (3x), 40-dim MFCCs + 100-dim i-vectors
- Lattice-Free Maximum Mutual Information criterion, 3-G LM

Table 1: Naming of the speech enhancement methods.

<table>
<thead>
<tr>
<th>Enhancement</th>
<th>Array</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td>Single/Multi</td>
<td>None</td>
</tr>
<tr>
<td>BeamformIt</td>
<td>Single</td>
<td>BFIt</td>
</tr>
<tr>
<td>WPE + GSS1 + BF w/ Context</td>
<td>Single</td>
<td>GSS1</td>
</tr>
<tr>
<td>WPE + GSS6 + BF w/ Context</td>
<td>Multi</td>
<td>GSS6</td>
</tr>
</tbody>
</table>

Effect of Acoustic Enhancement in ASR Training and Test

Table 2: WER results on the DEV (EVAL) set and various combinations of speech enhancement for ASR training and test. Amount of training data (hrs) is also specified.

<table>
<thead>
<tr>
<th>Enh. in trng (hrs)</th>
<th>Enhancement in test</th>
<th>Emb. in trng</th>
<th>Emb. in test</th>
<th>DT</th>
<th>RNN-LM</th>
<th>WER in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (2006)</td>
<td>GSS1</td>
<td>68.3 (65.9)</td>
<td>68.3 (65.9)</td>
<td>68.3 (65.9)</td>
<td>59.8 (57.3)</td>
<td>48.8 (49.2)</td>
</tr>
<tr>
<td>BFIt (680)</td>
<td>GSS1 + GSS6</td>
<td>73.1 (67.5)</td>
<td>73.1 (67.5)</td>
<td>73.1 (67.5)</td>
<td>53.0 (49.6)</td>
<td>48.0 (47.5)</td>
</tr>
<tr>
<td>GSS6 (308)</td>
<td>GSS1 + GSS6</td>
<td>78.5 (73.1)</td>
<td>78.5 (73.1)</td>
<td>78.5 (73.1)</td>
<td>58.0 (56.1)</td>
<td>45.4 (45.7)</td>
</tr>
</tbody>
</table>

State-of-the-art Single-system for CHiME-5

Table 3: Comparison of the reference [3] and proposed systems in terms of amount of training data.

<table>
<thead>
<tr>
<th>Track System</th>
<th>Amount trng data (hrs)</th>
<th>WER in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single H/UPB [3]</td>
<td>4500</td>
<td>58.3 (53.1)</td>
</tr>
<tr>
<td>Proposed</td>
<td>791</td>
<td>48.6 (46.7)</td>
</tr>
<tr>
<td>Multiple</td>
<td>4508</td>
<td>41.1 (43.2)</td>
</tr>
<tr>
<td>Proposed</td>
<td>308</td>
<td>41.6 (43.2)</td>
</tr>
</tbody>
</table>

Conclusions
- cleaning up training data can lead to substantial WER reduction
- enhancement in training is advisable as long as enhancement in test is at least as strong as in training
- top single-system performance for CHiME-5: 41.6 (43.2%) WER

References