Experiments with Meeting Data

Ramana Rao Gadde†, Dave Gelbart‡, Thilo Pfau‡,
Andreas Stolcke†, Chuck Wooters‡

†Speech Technology Research Laboratory, SRI International, Menlo Park,
CA, USA
‡International Computer Science Institute, Berkeley, CA, USA
Introduction

ASR on meeting data is a new task. To better understand this task, we explored the following areas:

- Language Modelling
- Noise reduction
- Automatic segmentation
- Automatic speaker clustering
**Problem:** Lack of public LM training data for meetings, so our RT-02 meeting recognizer used the Hub-5 LM.

**Question:** How does this affect performance?
**LM Approach**

Method: Train LM on in-domain data.

- Train LM on 270k words from 28 ICSI meetings (excluding the 4 RT-02 meetings)
- Include all words from these meetings in recognizer vocabulary (1200 new words)
- Interpolate meeting LM with SWB recognizer LM, minimizing perplexity on 2 RT-02 training meetings
- Run 1st recognition pass (recognize, N-best rescore, decode sausages)
## LM Results

WER on 2 RT-02 ICSI eval meetings (personal mics)

<table>
<thead>
<tr>
<th></th>
<th>Swbd LM</th>
<th>Meeting LM</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-best</td>
<td>34.6%</td>
<td>31.2%</td>
<td>3.4%</td>
</tr>
<tr>
<td>rescoped</td>
<td>30.6%</td>
<td>28.4%</td>
<td>2.2%</td>
</tr>
</tbody>
</table>

**Note:** OOV with Swbd LM is 1.5%, with Meeting LM it is 0.5%
Noise Reduction

**Problem:** Error rates on tabletop mics are significantly higher than on personal mics

**Question:** Can noise-reduction improve tabletop mic performance?
Noise Reduction Approach

Method: Use components from Qualcomm-ICSI-OGI Aurora system* (applied to test data only)

1. Apply voice-activity detection to find non-speech frames
2. Perform Wiener filtering using noise estimates obtained from the non-speech frames
3. Use overlap-add resynthesis to create a noise-reduced version of the original waveform

*Details can be found in the system description.
Noise Reduction Results

WER on 10-min dev (ICSI/LDC/CMU) meeting segments (with knowledge of “true” speakers)

- Original: 64.1%
- Noise-reduced: 61.7%
- Improvement: 2.4%
Automatic Segmentation

**Problem:** Tabletop mic data is unsegmented (no knowledge of speech or speaker boundaries)

**Question:** How does this affect performance?
Method:

- For eval system, we used a simple GMM-based speech/non-speech detector.
- Trained on one ICSI and one CMU meeting.
- Couldn’t use our “standard”* meeting segmenter, as it relies on info from personal mic channels. Could have tried it on personal mic unsegmented condition, but no time and probably not enough training data.

Automatic Segmentation Results

WER on 10-min, noise-reduced, dev (ICSI/LDC/CMU) meeting segments (includes overlapping segments)

“true” speaker segments 61.7%
auto segmentation 76.2%
Degredation 14.5%

Note that because we did not exclude overlapping speech, the ref transcripts contain the words from ALL speakers thus artifically increasing deletions.
Auto. Speaker Clustering

**Problem:** Switchboard system relies on speaker identity for feature normalization and acoustic model adaptation. However no speaker info for tabletop mic condition.

**Solution:** Cluster meeting waveform segments into "pseudo-speakers".
Auto. Speaker Clustering Approach

Method:

1. Build Gaussian mixture model from all segments.

2. Cluster segments based on mixture weight similarity. Distance metric: entropy increase due to cluster merging.

3. Stop when "expected" number of clusters is reached (5 for our system).
Auto. Speaker Clustering Results

WER on tabletop mic waveforms, dev (ICSI/LDC/CMU) data (non-overlapping segments).

- True speaker clusters: 64.6%
- Automatic speaker clusters: 65.6%
- Degredation: 1.0%
Conclusion

- With certain constraints, recognizer performance on meeting data seems to behave similarly to switchboard data.

- The level of difficulty of the meeting data task can be varied, by removing one or more of these constraints.

- The core meeting task (tabletop mics, unsegmented) is challenging and requires further research.