Current work at ICSI

Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1. Broadcast News MLP recognizer
2. Topic modeling
3. Acoustic segment classification
4. Thisl demonstrator front-end
The modulation-filtered spectrogram
(Brian Kingsbury)

- **Goal:** invariance to variable acoustics
  - filter out irrelevant modulations
  - channel adaptation (on-line auto. gain control)
  - multiple representations

- **Results (small vocabulary):**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Clean test WER</th>
<th>Reverb test WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>plp</td>
<td>5.9%</td>
<td>22.2%</td>
</tr>
<tr>
<td>msg</td>
<td>6.1%</td>
<td>13.8%</td>
</tr>
</tbody>
</table>
Broadcast News recognizer

- 1998 evaluation - RNN + MLP
- 8000 HU nets trained for MLP-only system:

<table>
<thead>
<tr>
<th>combo</th>
<th>RNN98</th>
<th>MSG-8kHz</th>
<th>PLP-16kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN98</td>
<td>27.2</td>
<td>24.9</td>
<td>24.5</td>
</tr>
<tr>
<td>MSG-8kHz</td>
<td></td>
<td>29.7</td>
<td>24.4</td>
</tr>
<tr>
<td>PLP-16Khz</td>
<td></td>
<td></td>
<td>25.5</td>
</tr>
</tbody>
</table>

- RNN+MSG+PLP: 23.7%
- plp 8000HU forward-pass ~0.7x real time (spert)

- Gender-dependent versions:

<table>
<thead>
<tr>
<th>net set</th>
<th>WER_F%</th>
<th>WER_M%</th>
<th>WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>plp-GD</td>
<td>20.3</td>
<td>27.2</td>
<td>24.6</td>
</tr>
<tr>
<td>msg-GD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plp+msg-GD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Broadcast News: ongoing

- **Dynamic pronunciations** (Eric Fosler)
  - data-derived rules for context-dependent pronunciations:
    phones, syllables, words, rate ...
  - rescoring N-best output from 1st pass
  - ~ 3% RER improvement

- **Multiband** (Adam Janin / Nikki Mirghafori)
  - 20% RER for small-vocabulary (Numbers)
  - no significant improvement yet for BN
  - features: MSG, cepstra, KLT, plp
  - all-way possible combinations & weights
Multiband for Broadcast News
(Adam Janin / Nikki Mirghafori)

- **Scheme that worked best for small vocab:**
  - 4-way frequency split
  - plp cepstra+deltas within each band
  - MLP classifier for each band + MLP combiner

- **Weighted average of all possible combos**
  - \( p(q \mid a,b,c,d) = \sum_S p(q \mid S,a,b,c,d) \cdot p(S) \)
  - \( S \) ranges over 16 possible combinations
  - \( p(S) \) from? constant, local feature (entropy)
  - oracle best \( p(S) \) → WER=19% (25%RER)
**Topic modeling**  
(Dan Gildea & Thomas Hofmann)

- **Bayesian model:**
  - $p(\text{word} \mid \text{doc}) = \sum_t p(\text{word} \mid \text{topic}) \ p(\text{topic} \mid \text{doc})$
  - EM modeling of $p(\text{word} \mid \text{topic})$ & $p(\text{topic} \mid \text{doc})$ over training set
  - $p(\text{topic} \mid \text{doc})$ estimated from context in recognition

- **Use to modify language model weights**
  - $p(\text{word}) \propto p_{\text{tri}}(\text{word}) \ p_{\text{top}}(\text{word}) / p_{\text{uni}}(\text{word})$
  - WSJ: trigram perplexity of 109 reduced 17%
  - use for BN recognition?

- **Use for topic segmentation?**
Acoustic Segment Classification
(Gethin Williams (SU) & Dan Ellis)

- Features from posteriors show utterance type:
  - average per-frame entropy
  - ‘dynamism’ - mean squared 1st-order difference
  - average energy of ‘silence’ label
  - covariance matrix distance to clean speech

- 100% on Scheirer/Slaney speech-music testset

- Use for acoustic segmentation?
ThisI demo development

- Stand-alone Tcl/Tk implementation
  - doesn’t require httpd
  - speech-input ready