RESPITE progress report

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

Outline

1. Hybrid AURORA system
2. Using hybrid results with HTK
3. Multifeature design
4. Multistream pronunciation modeling
Hybrid AURORA system

- AURORA noisy digits task
  - TIDIGITS + 4 kinds of noise x 7 SNR levels
  - standard HTK back-end provided
  - objective: standard features for mobile phones

- ICSI’s small-vocab techniques
  - modulation-filtered spectrogram (MSG) features
  - posterior probability combination (multistream)

- Can we combine them?
  - hybrid NN-HMM baseline system for AURORA
  - use a TIDIGITS lexicon & phone models
  - bootstrap labels from NUMBERS95 network
  - use 480 hidden-unit net as N95
Baseline AURORA results

- AURORA test has 28 numbers...
- ...report just a few
  - mean WER % for $\infty$, 15, 5, -5 dB SNR
  + overall mean ratio to HTK MFCC baseline

<table>
<thead>
<tr>
<th>System</th>
<th>Feature</th>
<th>Clean</th>
<th>SNR15</th>
<th>SNR5</th>
<th>SNR-5</th>
<th>Avg. ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTK</td>
<td>MFCC+d</td>
<td>1.4%</td>
<td>3.7%</td>
<td>15.9%</td>
<td>68.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Hybrid</td>
<td>MFCC+d</td>
<td>2.2%</td>
<td>2.6%</td>
<td>9.9%</td>
<td>49.1%</td>
<td>82.1%</td>
</tr>
<tr>
<td>Hybrid</td>
<td>plp12N+d</td>
<td>2.6%</td>
<td>2.8%</td>
<td>10.6%</td>
<td>47.9%</td>
<td>89.6%</td>
</tr>
<tr>
<td>Hybrid</td>
<td>msg3N</td>
<td>2.1%</td>
<td>2.9%</td>
<td>11.6%</td>
<td>49.2%</td>
<td>87.1%</td>
</tr>
<tr>
<td>HTK</td>
<td>msg3NKG</td>
<td>5.6%</td>
<td>6.4%</td>
<td>21.5%</td>
<td>66.8%</td>
<td>184.5%</td>
</tr>
</tbody>
</table>
Combination systems

- Posterior combination has worked well

\[ P(q_i|X_1,X_2) \propto P(q_i|X_1) \cdot P(q_i|X_2) / P(q_i) \quad \text{...if} \quad X_1 \perp X_2|q \]

- But it depends on features

<table>
<thead>
<tr>
<th>Features</th>
<th>Clean</th>
<th>SNR15</th>
<th>SNR5</th>
<th>SNR-5</th>
<th>Avg. ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>plp12Nd</td>
<td>2.6%</td>
<td>2.8%</td>
<td>10.6%</td>
<td>47.9%</td>
<td>89.6%</td>
</tr>
<tr>
<td>msg3N</td>
<td>2.1%</td>
<td>2.9%</td>
<td>11.6%</td>
<td>49.2%</td>
<td>87.1%</td>
</tr>
<tr>
<td>plp12Nd-msg3N</td>
<td>1.7%</td>
<td>2.4%</td>
<td>9.5%</td>
<td>47.3%</td>
<td>74.1%</td>
</tr>
<tr>
<td>plp12N-msg3aN</td>
<td>1.7%</td>
<td>2.1%</td>
<td>8.8%</td>
<td>46.9%</td>
<td>70.1%</td>
</tr>
<tr>
<td>• dplp12N-msg3bN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>plp12Nd • msg3N</td>
<td>1.5%</td>
<td>1.9%</td>
<td>8.2%</td>
<td>43.0%</td>
<td>63.0%</td>
</tr>
</tbody>
</table>
Using hybrid results with HTK

- AURORA specification: use HTK recognizer
- How to put combinations into HTK
  - feature combination (with LDA?)
  - posteriors as features (only 24 phone classes)

HTK handles it!

<table>
<thead>
<tr>
<th>System</th>
<th>Feature</th>
<th>Clean</th>
<th>SNR15</th>
<th>SNR5</th>
<th>SNR-5</th>
<th>Avg. ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td>plp • msg</td>
<td>1.5%</td>
<td>1.9%</td>
<td>8.2%</td>
<td>43.0%</td>
<td>63.0%</td>
</tr>
<tr>
<td>HTK</td>
<td>posteriors</td>
<td>1.1%</td>
<td>1.9%</td>
<td>8.2%</td>
<td>46.1%</td>
<td>59.1%</td>
</tr>
</tbody>
</table>
Tailoring posteriors for HTK

- Posteriors are very un-Gaussian
  - log-transform doesn’t help much

- A linear output layer helps a lot
  - remove softmax: \( y_i = \exp(x_i)/\sum_j(\exp(x_j)) \)

- Do combinations by summing linear outputs

<table>
<thead>
<tr>
<th>System</th>
<th>Feature</th>
<th>Clean</th>
<th>SNR15</th>
<th>SNR5</th>
<th>SNR-5</th>
<th>Avg. ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTK</td>
<td>posteriors</td>
<td>1.1%</td>
<td>1.9%</td>
<td>8.2%</td>
<td>46.1%</td>
<td>59.1%</td>
</tr>
<tr>
<td>HTK</td>
<td>log(p)</td>
<td>0.9%</td>
<td>1.8%</td>
<td>8.9%</td>
<td>48.8%</td>
<td>58.6%</td>
</tr>
<tr>
<td>HTK</td>
<td>( \Sigma(\text{lin. o/p}) )</td>
<td>0.9%</td>
<td>1.6%</td>
<td>7.7%</td>
<td>44.1%</td>
<td>51.6%</td>
</tr>
</tbody>
</table>
Multifeature design

(Mike Shire)

- ‘Optimal’ features for different conditions
  - subband envelope domain
  - linear-discriminant analysis (LDA) for filter coeffs

- Modulation-frequency domain responses for clean, reverb, mixture:

![Graph showing modulation-frequency domain responses for clean, light reverb, severe reverb, and clean-severe reverb mixtures.](image)
Multistream pronunciation models

(Barry Chen)

- Combine streams in the decoder
  - ‘HMM combination’
  - separate state assignment for each stream
  - constrain (disallow?) asynchrony

- Are particular asynchronies important?
  - between certain bands?
  - between certain sounds?
  - in particular directions?

- Re-estimate transition probabilities in 1-state asynchrony 4-band models
  - no improvement yet