CTS PROGRESS AT LIMSI

J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen

RT03 meeting
Boston, MA
May 19, 2003
TALK OUTLINE

- System overview
- Progress
- Acoustic models and decoding
- Language models and components for system combination
- Conclusions
# SYSTEM OVERVIEW

<table>
<thead>
<tr>
<th>Acoustic modeling</th>
<th>RT02</th>
<th>RT03</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP frontend</td>
<td></td>
<td>+ MFCC + PLP-S</td>
</tr>
<tr>
<td>Normalisations:</td>
<td>VTLN, CMN, CVN</td>
<td>+ gender-dependent VTLN</td>
</tr>
<tr>
<td>MLE/MAP trained</td>
<td>GD models</td>
<td>+ MMI training</td>
</tr>
<tr>
<td>triphones</td>
<td>28k tied-state triphones, 11k states</td>
<td>32k triphones</td>
</tr>
<tr>
<td>Cell models and</td>
<td></td>
<td>only one model set</td>
</tr>
<tr>
<td>cell switch</td>
<td></td>
<td>+ CTRAN SWB2 data</td>
</tr>
<tr>
<td>Training data:</td>
<td>SWB1, CallHome</td>
<td>+ reduced 35 phoneset</td>
</tr>
<tr>
<td>48 phone symbols</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**LIMSI-CNRS**

Spoken Language Processing Group - *May 2003, RT03 Meeting*
# SYSTEM OVERVIEW

<table>
<thead>
<tr>
<th>RT02</th>
<th>RT03</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language modeling</strong></td>
<td><strong>Language modeling</strong></td>
</tr>
<tr>
<td>42k vocabulary, ~300 compounds&lt;br&gt;4-gram backoff LM&lt;br&gt;Neural-net LM</td>
<td>50k vocabulary&lt;br&gt;Improved LM (smoothing, data, …)</td>
</tr>
<tr>
<td><strong>Decoding</strong></td>
<td><strong>Decoding</strong></td>
</tr>
<tr>
<td>3-gram lattices, 4g rescoring&lt;br&gt;3 passes decoding&lt;br&gt;2 and 5 phone class MLLR&lt;br&gt;Consensus decoding with pron probs&lt;br&gt;Confidence scores from CN</td>
<td>2-gram lattices, 4g rescoring&lt;br&gt;4 passes + 1 pass per component&lt;br&gt;+ 8 phone class MLLR</td>
</tr>
</tbody>
</table>
MAIN IMPROVEMENTS FOR RT03

- Gender-dependent VTLN (∼0.5%)
- MMI training of GD acoustic models (∼1.2%)
- Revised decoding (∼1.0%)
- CTRAN acoustic data (∼0.5%)
- Improved LM (∼1.0%)
- System combination (3 front-ends, 2 phone sets) (∼1.0%)
  Total gain ∼5.5% (eval01 and eval02)
- Integrated system with BBN

Total gain ∼5.5% (eval01 and eval02)
GENDER-DEPENDENT VTLN

- RT02 Method
  - Warp filter bank with a piecewise linear scaling function
  - GI ML estimation based on a 1st hypothesis with large models
  - Incremental search with 0.2 step

- RT03 Method
  - Gender-dependent warping (like 2 frontends)
  - ML estimation with single Gaussian models, Brent’s search
  - WER reduction $\sim 0.5\%$
  - Very fast (0.1xRT)
GD VTLN FOR MAP FM TRAINING
ACOUSTIC MODEL TRAINING

Training data
SW1, CH, CTRAN

F-VTLN

SI MLE training

F MAP adaptation

SI MMIE

F models

M-VTLN

SI MLE training

M MAP adaptation

SI MMIE

M models

Decoder

Lattices

F/M models
## RT03 DECODING

<table>
<thead>
<tr>
<th></th>
<th>VTLN</th>
<th>MLLR</th>
<th>LM</th>
<th>Eval01</th>
<th>Eval02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass 1 PLP MLE</td>
<td>n</td>
<td>-</td>
<td>3g</td>
<td>35.6</td>
<td>40.5</td>
</tr>
<tr>
<td>Pass 2 PLP MMI</td>
<td>y</td>
<td>-</td>
<td>4g</td>
<td>25.2</td>
<td>29.0</td>
</tr>
<tr>
<td>Pass 3 PLP MMI</td>
<td>y</td>
<td>2</td>
<td>4g</td>
<td>22.8</td>
<td>26.2</td>
</tr>
<tr>
<td>Pass 4 PLP MMI</td>
<td>y</td>
<td>5</td>
<td>NN  4g</td>
<td>21.9</td>
<td>25.1</td>
</tr>
<tr>
<td>4 system combination</td>
<td>y</td>
<td>5/8</td>
<td>NN  4g</td>
<td>20.9</td>
<td>24.0</td>
</tr>
</tbody>
</table>
LANGUAGE MODEL AND COMPONENTS FOR SYSTEM COMBINATION

Presented by H. Schwenk
LM TRAINING CORPORAS

RT02 system:
- SWB transcriptions from LDC (2.75M words) and from ISIP (2.93M words)
- CallHome corpus (229k words)
- SwitchBoard cellular transcriptions (217k words)
- BN commercial transcriptions (270M words)
- ”Switchboard-like” part of BN transcriptions (65M)

Additional corpora for RT03:
- CTRAN data from BBN [80h of fast SWB transcriptions] (1.1M words)
- WEB data from University of Washington (59M words)
## CONTRIBUTION OF NEW TEXTS

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Number of 2-gram</th>
<th>3-gram</th>
<th>4-gram</th>
<th>Perplexity Std</th>
<th>Decomp</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT03 dryrun</td>
<td>12M</td>
<td>21M</td>
<td>12M</td>
<td>82.8</td>
<td>60.2</td>
<td>22.94</td>
</tr>
<tr>
<td>+ CTRAN data</td>
<td>12M</td>
<td>21M</td>
<td>13M</td>
<td>80.8</td>
<td>58.8</td>
<td>22.76</td>
</tr>
<tr>
<td>+ improved smoothing</td>
<td>12M</td>
<td>22M</td>
<td>15M</td>
<td>80.3</td>
<td>58.4</td>
<td>22.47</td>
</tr>
<tr>
<td>+ WEB and CNN data</td>
<td>14M</td>
<td>24M</td>
<td>18M</td>
<td>79.3</td>
<td>57.8</td>
<td>22.21</td>
</tr>
</tbody>
</table>

Full decode with best acoustic models (without NN LM)

- Overall gain of 0.7%
- Important need for in-domain data
New Fisher data:

- Different epoch than previous CTS data
- New conversation topics
- No representative development data available

We tried to anticipate changes by updating the system vocabulary and the language models

- Added frequent words in recent BN data (mainly CNN)
- New wordlist: 51077 words (262 compounds), OOV 0.23% on eval01
  Eval03 LM has 16M 4-grams, 35M 3-grams and 22M 4-grams
- No change in px (55.6→55.3) and WER (21.92%→21.86%)
NEURAL NETWORK LANGUAGE MODEL

Characteristics:

- Performs n-gram probability estimation in a continuous space
- Trained only on the HUB5 corpora, interpolated with backoff LM
- Used for lattice rescoring during the last decoding pass

Performance comparison:

- Perplexity on eval01: 57.5 → 55.3 (68.8 → 63.5)

<table>
<thead>
<tr>
<th>WER (%)</th>
<th>Eval01</th>
<th>Eval02 (man)</th>
<th>Eval02 (auto)</th>
<th>Eval03 (auto)</th>
</tr>
</thead>
<tbody>
<tr>
<td>backoff LM</td>
<td>22.27</td>
<td>25.50</td>
<td>26.03</td>
<td>24.78</td>
</tr>
<tr>
<td>neural LM</td>
<td>21.86</td>
<td>25.09</td>
<td>25.71</td>
<td>24.43</td>
</tr>
</tbody>
</table>

Consistent gains of about 0.4%
COMPONENT SYSTEMS FOR COMBINATION

Four different systems were developed:

- **PLP** baseline system
- **PLP-S** short term cepstral mean and variance normalization.
- **PLP-R** reduced phone set (35 instead 45)
- **MFCC** front-end

**Characteristics:**

- All models are gender-dependent and MMI trained
- The alternate systems are built on top of the baseline system using 5-class MLLR adaptation
PERFORMANCE OF COMPONENT SYSTEMS

<table>
<thead>
<tr>
<th></th>
<th>Eval01</th>
<th>Eval02 (man)</th>
<th>its Eval02 (auto)</th>
<th>Eval03 (auto)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP</td>
<td>21.9</td>
<td>25.1</td>
<td>25.7</td>
<td>24.4</td>
</tr>
<tr>
<td>PLP-S</td>
<td>21.8</td>
<td>25.0</td>
<td>25.6</td>
<td>24.3</td>
</tr>
<tr>
<td>MFCC</td>
<td>21.8</td>
<td>25.0</td>
<td>25.6</td>
<td>24.3</td>
</tr>
<tr>
<td>PLP-R</td>
<td>21.9</td>
<td>24.9</td>
<td>25.6</td>
<td>24.4</td>
</tr>
</tbody>
</table>

System combination with BBN:

- Analysis of several combinations of 8 systems (4 LIMSI, 4 BBN)
- Selected 2 LIMSI systems: **PLP-S** and **PLP-R**
- Rover followed by 8-class transatlantic MLLR adaptation of the individual systems
- Experimental details will be given in BBN’s talk
CONCLUSIONS

• Significant improvements compared to RT02
• Main changes: VTLN, MMI, front-ends, phone set, LM, decoding
• Total gain of about 4.5% on each component system
• Selected components for system integration with BBN
• Large gain with BBN+LIMSI integrated system (… coming soon in the BBN talk)