Results

• Varimax and MPE training

• Local Trajectory Normalisation and Pitch

• Tonal Decision Tree Questions

• Mandarin Phone Sets

• Acoustic and Language Model Training Data

Mandarin CTS 2003 System

Bin Jia, Khe Chai Sim, Mark Gales, Thomas Hain, Andrew Liu, Phil Woodland, Kai Yu & the HTK STT Team

Cambridge University Engineering Department

May 19th 2003

CU-HTK RT03 Mandarin CTS System
Acoustic Training Set-Up

Acoustic/Training Test Data:
- **training data**: 34.9 hours, 379 sides, from LDC CallHome (22.4hrs) and CallFriend (12.5hrs), 451K Words (+7K English word), 628K Characters
- **development data**: dev02 1.94 hours from CallFriend

Front-end
- Reduced bandwidth 125{3800 Hz
- 12 PLP cepstral parameters + C0 and 1st/2nd derivatives
- Optional vocal tract length normalisation in training and test
- Optional vocal tract length normalisation from ESPS
- Side-biased cepstral mean and variance normalisation
- Approximate 3000 distinct states
- Decision tree state clustered, context dependent triphones
- Gender independent models
- Optional pitch (and derivatives) obtained from ESPS

Acoustic Models
- Class-based LM - 75 classes trained on acoustic transcriptions
- Word LMs - IJK vocabulary, 0.17% OOV on dev02
- Sources of data (using LDC character-to-word segmentor)

Language Model

Sources of data (using LDC character-to-word segmentor)
- News corpora: TDT[2,3,4], China Radio, People's Daily, Xinhua (good).
- Acoustic training data (mother tongue
- Development data: dev02 1.94 hours from CallFriend
- CallFriend (12.5hrs), 45K words (+/7K English worlds), 628K characters
- Training data: 34.9 hours, 379 sides, from LDC CallHome (22.4hrs) and LDC CallFriend
- Acoustic/Training Test Data:
Mandarin Phone Sets

Two phone sets considered:
{
59-phone set, start with LDC 60 phone set, remove tone markers.

46-phone set, start with 59-phone set and split long final phones, e.g.,

[u:e] \rightarrow [u:e]
[u:ng] \rightarrow [u:ng]
[au:ng] \rightarrow [au:ng]

\%

Mapping reduced CER by 1.1% absolute.

Tonal Decision Tree Questions

- Tonal questions were used for all further experiments.
- Tonal questions were used for all decision trees.
- 3% of possible questions were total.
- Tonal questions incorporated into decision tree process (without pitch features).

Tonal Decision Tree Questions

\%

Mapping reduced CER by 1.1% absolute.

Mandarin Phone Sets

\%

46 phone set was used for all further experiments.
Bin Jia, Khe Chai Sim et al: CU-HTK RT03 Mandarin CTS system

VTLN/Pitch Results

- **VTLN**
- **HLDA**
- **Pitch**

### VTLN

- **Pitch**
- **HLDA**
- **Pitch**

### HLDA

- **Pitch**
- **HLDA**

### Pitch

- **VTLN**
- **Pitch**
- **HLDA**

#### Results

- **Small variation in performance with dimensionality**
- **Pitch**: HLDA projection from baseline frontend and pitch
- **+Pitch**: baseline system with pitch added after HLDA
- **+Pitch**: baseline frontend with no pitch

#### Three systems examined:

<table>
<thead>
<tr>
<th>%CER for dev02 using MLE trained systems and word trigram LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>%CER for dev02 using MLE trained systems and word trigram LM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Vector Dimensionality</th>
<th># Dim</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP + Pitch + HLDA</td>
<td>39</td>
</tr>
<tr>
<td>PLP + Pitch</td>
<td>42</td>
</tr>
<tr>
<td>PLP</td>
<td>45</td>
</tr>
<tr>
<td>PLP + Pitch + HLDA + Pitch</td>
<td>48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th># Dim</th>
</tr>
</thead>
<tbody>
<tr>
<td>53.1</td>
<td>48</td>
</tr>
<tr>
<td>53.2</td>
<td>45</td>
</tr>
<tr>
<td>53.3</td>
<td>42</td>
</tr>
<tr>
<td>53.4</td>
<td>39</td>
</tr>
</tbody>
</table>

#### Feature Vector Dimensionality

- **VTLN** was used for all further experiments
- **Pitch** generally useful
- **HLDA** yields 0.8%-1.9% absolute reduction in CER
- **VTLN** yields 1.5%-1.8% absolute reduction in CER

#### Results

- **Small variation in performance with dimensionality**
- **VTLN/Pitch**: baseline frontend with no pitch
- **VTLN/Pitch + HLDA**: baseline system with pitch added after HLDA
- **VTLN/Pitch + HLDA + Pitch**: HLDA projection from baseline frontend and pitch

<table>
<thead>
<tr>
<th>%CER for dev02 using MLE trained systems and word trigram LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>%CER for dev02 using MLE trained systems and word trigram LM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VTLN/Pitch Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTLN + Pitch + HLDA</td>
</tr>
<tr>
<td>VTLN + Pitch</td>
</tr>
<tr>
<td>VTLN</td>
</tr>
<tr>
<td>VTLN + Pitch</td>
</tr>
<tr>
<td>VTLN + Pitch + HLDA</td>
</tr>
</tbody>
</table>
Automatic Segmentation degraded CER by 1% absolute.

- Manual segmentation error dominated by additional silence
- Untranscribed portions not scored (Manual M/S score attribute of smoothing)
- Reference derived from forced alignment of transcription portions

Diarisation score (% frame error) missed speech (M/S), false alarm (F/A): 64 components for speech, 1024 components for silence.

- PLP with energy and channel energy difference plus 1st/2nd derivatives

Automatic Segmentation

<table>
<thead>
<tr>
<th>CER (%)</th>
<th>Automatic</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.8</td>
<td>3.7</td>
<td>1.2</td>
</tr>
<tr>
<td>49.8</td>
<td>2.5</td>
<td>3.0</td>
</tr>
<tr>
<td>8.3</td>
<td>4.9</td>
<td>2.5</td>
</tr>
<tr>
<td>11.9</td>
<td>4.9</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Too many Gaussians per hour for 16 comp MPE system

- 16 component system MLE systems better and MPE system about same
- MPE yields 2.4% absolute for 16 component system
- Varmix yields 0.6%-0.8% absolute reduction in error rate

Additional Mixture Components/Varmix/MPE

<table>
<thead>
<tr>
<th># Comp</th>
<th>MLE +Varmix +MPE</th>
<th>12</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.2</td>
<td>52.3</td>
<td>51.7</td>
<td>49.9</td>
</tr>
</tbody>
</table>
Final confidence scores have NCE 0.190 on eval03

<table>
<thead>
<tr>
<th></th>
<th>CER (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dev02</td>
<td>47.9</td>
<td>48.5</td>
</tr>
<tr>
<td>eval03</td>
<td>54.7</td>
<td>51.3</td>
</tr>
<tr>
<td>P1 trans for VTLN</td>
<td>49.3</td>
<td>50.8</td>
</tr>
<tr>
<td>P2 trans for MLLR</td>
<td>50.5</td>
<td>51.3</td>
</tr>
<tr>
<td>P3 lat gen (bg)</td>
<td>49.8</td>
<td>50.5</td>
</tr>
<tr>
<td>P4 lat MLLR</td>
<td>48.6</td>
<td>49.5</td>
</tr>
<tr>
<td>P4 lat MLLR</td>
<td>48.6</td>
<td>49.5</td>
</tr>
<tr>
<td>LatMLLR</td>
<td>49.3</td>
<td>49.8</td>
</tr>
<tr>
<td>MPE triphones, HLDA, 11k, bgint03</td>
<td>49.3</td>
<td>49.8</td>
</tr>
<tr>
<td>MPE triphones, non-pitch, 11k, tgint03</td>
<td>49.3</td>
<td>49.8</td>
</tr>
</tbody>
</table>

Single system - currently no system combination.

Complete System Results

Cambridge University Engineering Department Rich Transcription Workshop 2003
Bin Jia, Khe Chai Sim et al: CU-HTK RT03 Mandarin CTS system

Absolute Gains:

<table>
<thead>
<tr>
<th>Phone Set</th>
<th>dev02 vs eval03</th>
</tr>
</thead>
<tbody>
<tr>
<td>59-phone</td>
<td>-1.1</td>
</tr>
<tr>
<td>46-phone</td>
<td>-1.0</td>
</tr>
<tr>
<td>non-Tonal</td>
<td>-1.3</td>
</tr>
<tr>
<td>Tonal</td>
<td>-1.7</td>
</tr>
<tr>
<td>non-VTLN</td>
<td>-1.8</td>
</tr>
<tr>
<td>VTLN</td>
<td>-1.9</td>
</tr>
<tr>
<td>non-pitch</td>
<td>-1.1</td>
</tr>
<tr>
<td>pitch</td>
<td>-0.1</td>
</tr>
<tr>
<td>non-HLDA</td>
<td>-1.9</td>
</tr>
<tr>
<td>HLDA</td>
<td>-0.9</td>
</tr>
</tbody>
</table>

Comparison of dev02 and eval03 gains:

- Standard techniques yield gains (but consistently less than expected)
- VTLN, pitch, MPE and linear adaptation
- 3 emitting states per phone model
- 46 phone set, with tonal decision tree questions

Future work:

- investigate limited gains from standard schemes
- all design choices gave improvements on both test sets
- all design choices gave improvements on both test sets
- train/test speaker overlap?
- investigate limited gains (but consistently less than expected)
- train/test speaker overlap?

Conclusions

- standard techniques yield gains (but consistently less than expected)
- all design choices gave improvements on both test sets

Current system:

- 46 phone set, with tonal decision tree questions
- pitch, MPE and linear adaptation
- 3 emitting states per phone model
- VTLN, pitch, MPE and linear adaptation
- VTLN, pitch, MPE and linear adaptation

Future work:

- investigate limited gains from standard schemes
- additional systems, SAT etc, and system combination
- alternative phone sets
- additional systems, SAT etc, and system combination
- alternative phone sets
- standard techniques yield gains (but consistently less than expected)

Conclusions

Current system:

- 46 phone set, with tonal decision tree questions
- pitch, MPE and linear adaptation
- 3 emitting states per phone model
- VTLN, pitch, MPE and linear adaptation

Future work:

- investi