Speaker & Language Adaptive Training for HMM-Based Polyglot Speech Synthesis

Heiga Zen

Toshiba Research Europe Ltd.
Cambridge Research Lab.

27/09/2010@Interspeech, Chiba, Japan
Outline

- Definition of polyglot speech synthesis
- Conventional approaches
  * Mixing mono-lingual corpora
- Speaker & language adaptive training (SLAT)
  * Concept
  * Definition
  * Parameter estimation
- Experiments
- Conclusions
Polyglot Speech Synthesis

Synthesize multiple languages with common voice

Applications
* Synthesize mix-lingual texts
* Speech-to-speech translators
* More efficient development of TTS for multiple languages
Conventional Approach

Mix mono-lingual corpus [Latorre;'06, Black;'06]

All languages & speakers are simply mixed to estimate model
→ Language & speaker variations are not well addressed
Speaker & Language Adaptive Training (SLAT)
Speaker & Language Adaptive Training (SLAT)

Speaker transform
- Speaker-specific characteristics
  * Vocal tract length & shape, F0 height & range, voicing
  * Speaking rate, speaker-specific speaking styles

⇒ Constrained MLLR [Gales;'98]
Speaker & Language Adaptive Training (SLAT)

Language transform
- Language-specific factors
  * Syntactic, morphological, & intonational factors

Canonical model
- Common factors across languages & speakers
  * Phonological & phonetic factors

⇒ CAT with cluster-dependent decision trees [Zen;'09]
Cluster Adaptive Training (CAT)

Language adaptation by CAT
- CAT ⇒ *Soft* version of speaker clustering
- Apply CAT idea to represent languages

![](image)

**Target language**
⇒ Weighted sum of underlying *prototype languages*
Cluster Adaptive Training (CAT)

Language adaptation by CAT

- CAT ⇒ *Soft* version of speaker clustering
- Apply CAT idea to represent languages

- Weight for bias cluster is always equal to 1
  ⇒ Represent *common* factor across *languages*
Cluster Adaptive Training (CAT)

Language adaptation by CAT
- CAT ⇒ *Soft* version of speaker clustering
- Apply CAT idea to represent languages

Prototype languages have their own context dependencies
⇒ CAT with cluster-dependent decision trees [Zen;'09]
Tree Intersection Interpretation

context space

$3 \times 3 \times 4 = 36$

#leaf nodes = 36
Tree Intersection Interpretation

Cluster 1

Cluster 2

Cluster P

Context space

$3 \times 3 \times 4 = 36$

#leaf nodes = 10
Speaker & Language Adaptive Training (SLAT)

Speaker transform
⇒ CMLLR

Language transform
⇒ CAT non-bias clusters & CAT interpolation weights

Canonical model
⇒ CAT bias cluster

Trees & params can be updated iteratively by EM
Definition of state output distributions of SLAT

\[ p(o(t) \mid m, s, l, \mathcal{M}) = \begin{vmatrix} A_r^{(s)}(m) \end{vmatrix} \mathcal{N} \left( \begin{array}{c} A_r^{(s)}(m) o(t) + b_r^{(s)}(m) \end{array} \right), \right. \right. \left. \left. \sum_{i=1}^{P} \lambda_{i,q(m)}^{(l)} \mu_{c(m,i)}, \Sigma_v(m) \right) \]

\( o(t) \): observation vector at frame \( t \)

\( m \): mixture component index

\( s \): speaker label associated with \( o(t) \)

\( l \): language label associated with \( o(t) \)

\( A, b \): CMLLR transforms

\( \lambda \): CAT interpolation weights

\( \mu \): CAT cluster mean vectors

\( \Sigma \): canonical covariance matrices

\( r(m) \): CMLLR regression class

\( q(m) \): CAT regression class

\( c(m,i) \): mean vector index

\( v(m) \): covariance matrix index
Experimental Conditions

Data
- German, French, Spanish, UK & US English
- 10 speakers per language (5 female & 5 male)
- 100 or 150 sentences per speaker
- Consistent microphone & recording condition

Data preparation
- IPA-like universal phone set
- Universal context-dependent label format
  * phone, syllable, word, phrase, & utterance-level contexts
Experimental Conditions

**Speech analysis / training setup**
- HTS-2008 (SAT system for BC08) setup [Yamagishi;'08]
- LI-SAT (language-independent) was trained
- Initialize SLAT model by LI-SAT model then reestimate
- LD-SAT (language-dependent) models were also trained
- #reg classes for CAT & CMLLR were 3 (dur) or 2 (others)
  * silence, short pause, or speech / silence or speech

**Synthesis setup**
- Speech parameter generation algorithm with GV [Toda;'07]

Please refer to paper for other details
### Number of Leaf Nodes

<table>
<thead>
<tr>
<th>Cluster</th>
<th>mel-cep</th>
<th>log F0</th>
<th>band ap</th>
<th>dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (bias)</td>
<td>4,537</td>
<td>12,894</td>
<td>1,866</td>
<td>1,724</td>
</tr>
<tr>
<td>2</td>
<td>165</td>
<td>1,954</td>
<td>306</td>
<td>65</td>
</tr>
<tr>
<td>3</td>
<td>244</td>
<td>1,970</td>
<td>173</td>
<td>59</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>1,940</td>
<td>226</td>
<td>127</td>
</tr>
<tr>
<td>5</td>
<td>208</td>
<td>1,119</td>
<td>227</td>
<td>52</td>
</tr>
<tr>
<td>6</td>
<td>161</td>
<td>1,421</td>
<td>261</td>
<td>94</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5,515</strong></td>
<td><strong>21,298</strong></td>
<td><strong>3,059</strong></td>
<td><strong>2,121</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>mel-cep</th>
<th>log F0</th>
<th>band ap</th>
<th>dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>LI-SAT</td>
<td>4,359</td>
<td>31,201</td>
<td>2,244</td>
<td>2,259</td>
</tr>
<tr>
<td>LD-SAT</td>
<td>5,895</td>
<td>37,847</td>
<td>3,205</td>
<td>1,898</td>
</tr>
</tbody>
</table>

Total sizes of trees were comparable
### Number of Leaf Nodes

<table>
<thead>
<tr>
<th>Cluster</th>
<th>mel-cep</th>
<th>log F0</th>
<th>band ap</th>
<th>dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (bias)</td>
<td>4,537</td>
<td>12,894</td>
<td>1,866</td>
<td>1,724</td>
</tr>
<tr>
<td>2</td>
<td>165</td>
<td>1,954</td>
<td>306</td>
<td>65</td>
</tr>
<tr>
<td>3</td>
<td>244</td>
<td>1,970</td>
<td>173</td>
<td>59</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>1,940</td>
<td>226</td>
<td>127</td>
</tr>
<tr>
<td>5</td>
<td>208</td>
<td>1,119</td>
<td>227</td>
<td>52</td>
</tr>
<tr>
<td>6</td>
<td>161</td>
<td>1,421</td>
<td>261</td>
<td>94</td>
</tr>
<tr>
<td>Total</td>
<td>5,515</td>
<td>21,298</td>
<td>3,059</td>
<td>2,121</td>
</tr>
</tbody>
</table>

Bias cluster was largest in all speech params

⇒ Common factor across languages was dominant
## Language-Dependent CAT Weight Vectors

### mel-cep

<table>
<thead>
<tr>
<th>Language</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>1</td>
<td>0.53</td>
<td>0.31</td>
<td>0.01</td>
<td>0.37</td>
<td>0.35</td>
</tr>
<tr>
<td>UK English</td>
<td>1</td>
<td>0.24</td>
<td>0.47</td>
<td>0.41</td>
<td>0.25</td>
<td>0.31</td>
</tr>
<tr>
<td>US English</td>
<td>1</td>
<td>0.25</td>
<td>0.37</td>
<td>0.70</td>
<td>0.26</td>
<td>0.33</td>
</tr>
<tr>
<td>Spanish</td>
<td>1</td>
<td>0.37</td>
<td>0.34</td>
<td>0.00</td>
<td>0.52</td>
<td>0.39</td>
</tr>
<tr>
<td>French</td>
<td>1</td>
<td>0.38</td>
<td>0.24</td>
<td>-0.05</td>
<td>0.36</td>
<td>0.56</td>
</tr>
</tbody>
</table>

### log F0

<table>
<thead>
<tr>
<th>Language</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>1</td>
<td>0.90</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>UK English</td>
<td>1</td>
<td>0.02</td>
<td>0.91</td>
<td>0.10</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>US English</td>
<td>1</td>
<td>0.05</td>
<td>0.10</td>
<td>0.90</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>Spanish</td>
<td>1</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.90</td>
<td>0.07</td>
</tr>
<tr>
<td>French</td>
<td>1</td>
<td>0.03</td>
<td>0.04</td>
<td>0.07</td>
<td>0.08</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Paired Comparison Test

Preference test comparing LD-SAT, LI-SAT, & SLAT
- 250 test sentences excluded from training data
  * 50 sentences per language
- 14 subjects, only evaluate native or near native languages
- 15 sentences per subject

Results

<table>
<thead>
<tr>
<th>LI-SAT</th>
<th>LD-SAT</th>
<th>SLAT</th>
<th>No pref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>33.3</td>
<td>36.2</td>
<td>–</td>
<td>30.5</td>
</tr>
<tr>
<td>24.2</td>
<td>–</td>
<td>37.6</td>
<td>38.2</td>
</tr>
<tr>
<td>–</td>
<td>26.7</td>
<td>45.6</td>
<td>27.7</td>
</tr>
</tbody>
</table>

Significant improvements ($p<0.01$) by SLAT
Conclusions

**Speaker & language adaptive training**
- Combine 2 adaptive training schemes
  * CMLLR for speaker adaptive training
  * CAT w/ cluster-dependent trees for lang adaptive training
- Multi-language / -speaker adaptive training
- Achieved significant improvements

**Future plans**
- Evaluate language adaptation
- Increase data & speakers per language
- Add non-European languages (e.g., Japanese)