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

Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Kate Knill, Sacha Krstulovic, Javier Latorre

Toshiba Research Europe Ltd.
Cambridge Research Lab.

24/09/2010@ISCA SSW7, Kyoto, Japan
Outline

- Definition of polyglot speech synthesis
- Conventional approaches
  * Polyglot speaker
  * Mixing mono-lingual corpora
  * Cross-lingual speaker adaptation
- Speaker & language adaptive training (SLAT)
  * Concept
  * Details
- 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
Finding good polyglot speakers is very difficult
→ Hardly expandable
Conventional Approaches (2)

Mix mono-lingual corpus [Latorre;'06, Black;'06]
Conventional Approaches (2)

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

All languages & speakers are simply mixed to estimate model
→ Language & speaker variations are not well addressed
Conventional Approaches (3)

Cross-language speaker adaptation [Chen;'09, Wu;'09]

adaptive training  mapping  adaptation
Conventional Approaches (3)

Cross-language speaker adaptation [Chen;'09, Wu;'09]

Language-dependent SAT models are estimated independently
→ Mismatch between language-dependent SAT models
→ Degrade adaptation & synthesis [Liang;'10]
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)

Speaker adaptation by CAT [Gales;00]
"Soft" version of speaker clustering

Cluster 1: mean 1
Cluster 2: mean 2
(bias) Cluster P: mean P

\[ \text{Target speaker} \Rightarrow \text{Weighted sum of underlying } \text{prototype} \text{ speakers} \]
Cluster Adaptive Training (CAT)

Speaker adaptation by CAT [Gales;00] "Soft" version of speaker clustering

Prototype speakers are fixed across all speakers
Interpolation weights change speaker-by-speaker
Cluster Adaptive Training (CAT)

Speaker adaptation by CAT [Gales;00]
"Soft" version of speaker clustering

cluster 1  mean 1  \( \lambda_1 \)  +  \( \lambda_2 \)

cluster 2  mean 2

(bias) cluster \( P \)  mean \( P \)  1

Variance

Mean

Mix weights

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

Language adaptation by CAT
Extend CAT idea to represent languages

cluster 1
mean 1
\(\lambda_1\)
Variance

cluster 2
mean 2
\(\lambda_2\)
Mean

(bias) cluster \(P\)
mean \(P\)
Mix weights

Target language
\(\Rightarrow\) Weighted sum of underlying prototype languages
Cluster Adaptive Training (CAT)

Language adaptation by CAT
Extend CAT idea to represent languages

cluster 1
mean 1
\( \lambda_1 \)
Variance

cluster 2
mean 2
\( \lambda_2 \)
Mean

(bias) cluster \( P \)
mean \( P \)
1
Mix weights

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

Language adaptation by CAT
Extend CAT idea to represent languages

Tonal langs
European langs
(bias) cluster $P$

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

Cluster P

Cluster 2

Cluster 1

Context space
Tree Intersection Interpretation

cluster 1

context space for lang 1

cluster 2

context space for lang 2
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
Experimental Conditions

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

Data preparation
- IPA-like universal phone set
- Universal context-dependent label format
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

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>
<tr>
<td><strong>LI-SAT</strong></td>
<td><strong>4,359</strong></td>
<td><strong>31,201</strong></td>
<td><strong>2,244</strong></td>
<td><strong>2,259</strong></td>
</tr>
<tr>
<td><strong>LD-SAT</strong></td>
<td><strong>5,895</strong></td>
<td><strong>37,847</strong></td>
<td><strong>3,205</strong></td>
<td><strong>1,898</strong></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
# CAT Weight Vector After Reestimation

## Mel-Cep.

<table>
<thead>
<tr>
<th></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 0.53, 0.31, 0.01, 0.37, 0.35]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK English</td>
<td>[1 .24, 0.47, 0.41, 0.25, 0.31]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US English</td>
<td>[1 .25, 0.37, 0.70, 0.26, 0.33]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>[1 .37, 0.34, 0.00, 0.52, 0.39]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>[1 .38, 0.24, -0.05, 0.36, 0.56]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## log F0

<table>
<thead>
<tr>
<th></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 0.90, 0.06, 0.07, 0.08, 0.08]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK English</td>
<td>[1 .02, 0.91, 0.10, 0.02, 0.06]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US English</td>
<td>[1 .05, 0.10, 0.90, 0.01, 0.09]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>[1 .04, 0.04, 0.05, 0.90, 0.07]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>[1 .03, 0.04, 0.07, 0.08, 0.89]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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 evaluated native or near native langs
- 15 sentences per subject

Results

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

Significant improvements \((p<0.01)\) by SLAT
Evaluate Similarity of Cross-Lingual Adaptation

**DMOS test setup**

- Target speaker: German male speaker
- Target language: US English
- 10 subjects
- 7 systems * 5 sentences per subject
- Play reference, then candidate to be evaluated
- 5-scale similarity score (1: very dissimilar - 5: very similar)
- Subjects were asked to ignore other factors (e.g., intelligibility or naturalness)
Evaluate Similarity of Cross-Lingual Adaptation

**Evaluated systems**

1) US English LD-SAT w/o adaptation (AVM)
2) US English LD-SAT w/ CMLLR from a training speaker who sounded similar to target speaker (TRAIN)
3) US English LD-SAT w/ CMLLR for target speaker estimated by cross-lingual speaker adaptation (CROSS)
4) LI-SAT w/ CMLLR for target speaker (LI-SAT)
5) SLAT w/ CMLLR for target speaker (SLAT)
6) German LD-SAT w/ CMLLR for target speaker (INTRA)
7) Vocoded natural speech (VOCOD)

Reference, 6), & 7) were in German but others were in English
Evaluate Similarity of Cross-Lingual Adaptation

Adaptation from SLAT

⇒ Reduces mismatch between languages
⇒ Improved speaker similarity

Differential Mean Opinion Score

AVM: 1.1
TRAIN: 1.6
CROSS: 1.5
LI-SAT: 2.2
SLAT: 2.1
INTRA: 2.7
VOCOD: 4.6
Evaluate Similarity of Cross-Lingual Adaptation

Significant gap between SLAT & INTRA
⇒ Still some impact by language difference
Evaluate Similarity of Cross-Lingual Adaptation

Large gap between INTRA & VOCOD
⇒ Statistical modeling had the largest impact
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)