Integrated speaker-adaptive speech synthesis[1]
Moquan Wan, Gilles Degottex, Mark J.F. Gales
mw545@cam.ac.uk, gad27@cam.ac.uk, mjfg@eng.cam.ac.uk
University of Cambridge

Speech Synthesis Pipeline

Integrated iVector Extraction with
Attention Mechanism (IIEA)

Objective Measures

<table>
<thead>
<tr>
<th>Model</th>
<th>MCD (dB)</th>
<th>BAP (dB)</th>
<th>RMSE</th>
<th>F0 CORR</th>
<th>VUV error</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>6.703</td>
<td>2.229</td>
<td>30.463</td>
<td>0.836</td>
<td>5.924%</td>
</tr>
<tr>
<td>IIE</td>
<td>6.601</td>
<td>2.220</td>
<td>30.017</td>
<td>0.831</td>
<td>5.770%</td>
</tr>
<tr>
<td>IIEA</td>
<td>6.615</td>
<td>2.222</td>
<td>29.889</td>
<td>0.833</td>
<td>5.799%</td>
</tr>
</tbody>
</table>

Subjective Evaluations

Figure 6: naturalness preference test results, iVector size 128; 32 listeners took the test properly

Figure 7: similarity preference test results, iVector size 128; 33 listeners took the test properly

Conclusions

- Improved performance
- Reasonable Attention trajectory
  - Mostly in the voiced regions
  - Mostly in the vowel regions
- More selective than a binary of either

Challenges

- Attention mechanism is hard to train
  - Initialisation: with a pre-trained IIE, bootstrap
  - Also initialise IIE with a pre-trained baseline
- Additional Depth, vanishing gradient
  - Hyper-parameter tuning
  - Balance of adaptive data and acoustic data

Problem:

- Independent and different optimisation criteria
- Uniform weight on adaptive data
- i-vector extraction from acoustic data only

Auxiliary Model

Acoustic Model

• Learn a speaker space, each speaker corresponds to an iVector
• Previous iVectors: one-hot speaker code; d-vector, GMM-based i-vector

Problems:

- Contextual-Dependent Labels
- State-Alignment
- \(/\text{f}/, /\text{v}/, /\text{f}/, /\text{v}/, /\text{f}/, /\text{v}/, /\text{f}/, /\text{v}/, /\text{f}/, /\text{v}/, ...

iVector-based Adaptive Speech Synthesis

Integrated iVector Extraction with
Attention Mechanism (IIEA)

Auxiliary Model

Acoustic Model

- Integrate the iVector extraction process in training
- Apply attention mechanism to adaptive data
- Learn attentions from contextual labels
- “Flat Attention” (IIE): remove the Attention Mechanism (top right) and use a uniform Attention of \(1/T\).

Attention Trajectory

Phone sequence (black with green boundaries), VUV trajectory (red) and Attention trajectory (blue): “We also need a small plastic snake”