Feature-based Pronunciation Modeling for Automatic Speech Recognition

Karen Livescu and James Glass
Spoken Language Systems Group, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge MA 02139 USA

1. Main ideas

• **Motivation:**
  - From linguistics: Speech consists of multiple streams of semi-independent phonological features, rather than phones.
  - From speech recognition research:
    - Phone-based pronunciation modeling has had limited success.
    - Spontaneous speech is difficult to describe phonetically (see 2).

• **Previous work:**
  - Much research on recognition with feature classifiers.
  - But pronunciation model is still typically phone-based.

• **Contributions:**
  - Introduction of a general, flexible feature-based pronunciation model, accounting for feature asynchrony and feature substitutions.
  - Implementation using dynamic Bayesian networks.
  - Initial experiments on spontaneous pronunciations.

2. Examples

• warm/h [w aw r m p th] : phone insertion?
• wants/h [w aw n st] : phone deletion??
• several/h [s el h r v lx] : exchange of phones???
• instruments/h [j h n s ch em i n n s l] : ????
• everybody/h [e h r w ay] : ???
• All explainable via feature asynchrony + some feature value substitutions (‘undershooting’).

3. The model

**baseform dictionary**

To produce a given utterance, each feature should proceed through a particular sequence of values

- **asynchrony**
  - Different features may proceed through their respective sequences at different rates

**feature substitutions**

Surface (actual) feature values may differ from underlying (dictionary) values

**References**


4. Implementation

- The model is implemented as a dynamic Bayesian network (DBN): A representation, via a directed graph, of a distribution over a set of variables that evolve through time.
- Example DBN with three features:

5. Experiments

**Task:** Classify isolated words from the Switchboard corpus, given a detailed phonetic transcription (from ICSI)

- Convert transcription into feature vectors $S_0$, one per 10ms
- Find the word $w^* = \text{argmax}_w P(w|S_0)$
- Feature set based on vocal tract variables of articulatory phonology [1]: lip opening, tongue tip location/opening, tongue body location/opening, velum, glottis
- Experiments implemented in GMTK [2]
- Parameters initialized by hand and trained using EM
- See [3],[4] for additional details

6. Ongoing/future work

- Integration with landmark-based feature classifiers
- Context-dependent feature distributions
- More complex tasks (multisyllables, larger vocabularies)
- Use of articulatory databases
- Possible uses of such a model to learn about speech

---

**Feature-based pronunciation model**

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev set (165 words)</th>
<th>Test set (256 words)</th>
<th>Error rate</th>
<th>Feature rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseform-only</td>
<td>62.0</td>
<td>57.5</td>
<td>66.4</td>
<td>62.5</td>
</tr>
<tr>
<td>baseform + phone</td>
<td>57.0</td>
<td>54.5</td>
<td>62.1</td>
<td>56.5</td>
</tr>
<tr>
<td>baseform + articulations</td>
<td>35.2</td>
<td>33.9</td>
<td>44.5</td>
<td>25.7</td>
</tr>
<tr>
<td>baseform + values</td>
<td>28.5</td>
<td>16.8</td>
<td>40.7</td>
<td>24.6</td>
</tr>
<tr>
<td>baseform + values + EM</td>
<td>27.9</td>
<td>16.8</td>
<td>40.7</td>
<td>24.6</td>
</tr>
</tbody>
</table>

- Qualitative examination of alignments produced by the model showed expected asynchrony effects
- EM training has minor effect on error rates, but improves rank/score distributions [4]