An overview of methods for articulatory feature detection

Mahir Morshed

26 October 2020
Outline

1. The features themselves
   - Overview

2. Before deep learning

3. Earlier deep methods
   - Fully connected
   - Convolutional
   - Recurrent

4. More recent methods
   - Attention

5. The road ahead
Articulatory features

(what are they?)

- Facets of phone production by which differences between such phones may be characterized
- Although two languages may lack a common phone, close equivalents may exist which differ in a single characteristic
  - /t/ in South Asian languages vs /t/ elsewhere (place)
  - /r/ vs /ɾ/ (manner)
  - /p/ vs /b/ (voicing)
  - /e/ vs /ɛ/ (height)
  - /a/ vs /ɑ/ (frontness)
  - /ɯ/ vs /u/ (roundedness)
Modeling differences

or considerations in choosing and using an articulatory feature model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sonority</td>
<td>Vowel, Obstruent, Sonorant, Syllabic, Silence</td>
</tr>
<tr>
<td>Voicing</td>
<td>Voiced, Voiceless, Not Applicable</td>
</tr>
<tr>
<td><strong>Consonantal features</strong></td>
<td></td>
</tr>
<tr>
<td>Manner</td>
<td>Fricative (FR), Stop (STP), Flap (FLA), Nasal (NAS), Approximant (APP), Nasal Flap (NF), Not Applicable (NA)</td>
</tr>
<tr>
<td>Place</td>
<td>Labial (LAB), Dental (DEN), Alveolar (ALV), Palatal (PAL), Velar (VEL), Glottal (GLO), Lateral (LAT), Rhotic (RHO), Not Applicable (NA)</td>
</tr>
<tr>
<td><strong>Vowel features</strong></td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>High, Mid, Low, Lowhigh, Midhigh, Not Applicable</td>
</tr>
<tr>
<td>Frontness</td>
<td>Front, Back, Central, Backfront, Not Applicable</td>
</tr>
<tr>
<td>Roundness</td>
<td>Round, Non-round, Round-Non-round, Non-round-Round, Not Applicable</td>
</tr>
<tr>
<td>Tense</td>
<td>Tense, Lax, Not Applicable</td>
</tr>
</tbody>
</table>

- Binary/unary features ([+sonorant], [+round], [nasalized])?
- Features on a spectrum (e.g. for place, [bilabial]-[glottal])?
- Separate detectors per class, or a single detector for all features?
- (Direct detection of features, or translation from phones?)

**Figure:** Articulatory feature set used in

---

1Rajamanohar and Fosler-Lussier, “An evaluation of hierarchical articulatory feature detectors”.
Identifying articulatory cues

Liu, “Landmark detection for distinctive feature-based speech recognition”

- Separate coarse and fine preprocessors of broad frequency bands in original signal (using energy and deltas)
- Fine tailoring of detectors to distinctive features based on precomputed measurement thresholds
- Considerably greater error with sonorant detection (57%) versus for glottal vibration and bursts (5%/14%)

**Figure:** Landmarks identified using detectors for glottal vibration, sonorant closure/release, and stop bursts.
Recurrence binary detection
King and Taylor, “Detection of phonological features in continuous speech using neural networks”

- SPE, n-ary, and government phonology based feature sets examined
- Two-layer, 250 hidden unit, fully recurrent network detecting all SPE and GP-based features (multiple detectors in the n-ary case)
- \( \sim 90\% + \) accuracy for most features individually, but closer to \( \sim 50\% \) when taken together

**Figure:** Comparison of ground truths for the phrase "economic cutbacks" and outputs appertaining from the network trained using GP.
Fully connected detection/classification

Bhowmik, Chowdhury, and Das Mandal, “Deep Neural Network based Place and Manner of Articulation Detection and Classification for Bengali Continuous Speech”

<table>
<thead>
<tr>
<th>Output Class</th>
<th>Velar</th>
<th>Post-Alveolar</th>
<th>Alveolar</th>
<th>Dental</th>
<th>Bilabial</th>
<th>Glottal</th>
<th>Palatal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velar</td>
<td>10133</td>
<td>24</td>
<td>131</td>
<td>87</td>
<td>274</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Post-Alveolar</td>
<td>10825</td>
<td>5.3%</td>
<td>41</td>
<td>3</td>
<td>162</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Alveolar</td>
<td>30898</td>
<td>10.1%</td>
<td>41</td>
<td>3</td>
<td>162</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Dental</td>
<td>15650</td>
<td>5.1%</td>
<td>91</td>
<td>224</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bilabial</td>
<td>18064</td>
<td>5.9%</td>
<td>224</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Glottal</td>
<td>20492</td>
<td>6.7%</td>
<td>348</td>
<td>39</td>
<td>37</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>Palatal</td>
<td>33744</td>
<td>12.2%</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>Target Class</td>
<td>Veslar</td>
<td>Post-Alveolar</td>
<td>Alveolar</td>
<td>Dental</td>
<td>Bilabial</td>
<td>Glottal</td>
<td>Palatal</td>
</tr>
<tr>
<td>20.7%</td>
<td>99.1%</td>
<td>96.6%</td>
<td>97.8%</td>
<td>97.2%</td>
<td>97.2%</td>
<td>99.4%</td>
<td>50.2%</td>
</tr>
<tr>
<td>79.3%</td>
<td>0.9%</td>
<td>3.4%</td>
<td>2.2%</td>
<td>2.8%</td>
<td>2.8%</td>
<td>0.6%</td>
<td>49.8%</td>
</tr>
</tbody>
</table>

- 4-layer fully connected feature detectors
- ”Manner” groupings rather broad, covering voicing and aspiration
- ~ 90% accuracy for detection, but degraded to 50% for place classification

**Figure:** Confusion matrix for the place of articulation classifier.
Articulatory feature supplements

Manjunath et al., “Indian Languages ASR: A Multilingual Phone Recognition Framework with IPA Based Common Phone-set, Predicted Articulatory Features and Feature fusion”

- Comparisons between deep (5-layers) and shallow (1-layer) fully connected networks for detectors
- $\sim 85\%$ accurate feature classifiers in the deep case, with mixed improvements in overall phone recognition among tandem combinations

Figure: Multilingual phone recognition system information flow.
Convolutional classification

Merkx and Scharenborg, “Articulatory Feature Classification Using Convolutional Neural Networks”

- ~90% accuracy across feature classes using spectrograms without Mel filtering
- Compared to multi-layer perceptrons, major improvements to place classification, minor ones to manner classification

Figure: CNN-based articulatory feature detector architecture.
CTC-based feature extractors

Abraham, Umesh, and Joy, “Articulatory Feature Extraction Using CTC to Build Articulatory Classifiers Without Forced Frame Alignments for Speech Recognition”

- Fully connected, convolutional, and hybrid thereof architectures examined, alongside varied acoustic models
- $\sim 30\%$—word error rates using BiLSTMs with CTC loss and, as input, articulatory features appended to MFCCs

**Figure:** Articulatory feature extraction information flow
Aiding bottleneck features

Shetty et al., “Articulatory and Stacked Bottleneck Features for Low Resource Speech Recognition”

- Features, whether phones or articulations (concatenated, if necessary) fed into time-delayed neural network
- Slight accuracy improvements across languages compared to MFCCs or either articulatory or bottleneck features alone

**Figure:** Stacked bottleneck architecture for multilingual phone recognition
More recent methods

Listening and attending to articulation

Karaulov and Tkanov, “Attention Model for Articulatory Features Detection”

- Multi-task learning setups (cross-training with phone outputs) considered
- $\sim 20 - 25\%$ phone error rates using models in which LAS decoder inputs were mapped directly to features

**Figure:** Ground truths compared with outputs from the decoder
Attributes from transformers

Li et al., “End-to-End Articulatory Attribute Modeling for Low-Resource Multilingual Speech Recognition”

**Figure**: Overall architecture of the speech recognizer showing intermediate inputs and outputs thereof

- Grapheme inputs converted to sequences of attributes (that is, not as separate streams)
- Slightly reduced character error rates compared to multilingual models based on words, characters, or phones
Transfer learning for languages using recurrent networks as a basis

- The progressive network format
  1 Qu et al., “Combining Articulatory Features with End-to-End Learning in Speech Recognition”.

- Language model fusion
  2 Inaguma et al., “Transfer Learning of Language-independent End-to-end ASR with Language Model Fusion”.

- Articulograph readings as supplements
  3 Dash et al., “Automatic Speech Recognition with Articulatory Information and a Unified Dictionary for Hindi, Marathi, Bengali and Oriya”.

Figure: Progressive network architecture using articulatory feature detectors.
Transfer learning elsewhere
such as with variations between same-language speakers

Accounting for differences between native- and second-language speakers
Handling differences arising in pathological speech

Figure: Multi-task, multilingual enhancement of a fully-connected phone recognizer.

---

1 Duan et al., “Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning”.

2 Jenne and Vu, “Multimodal Articulation-Based Pronunciation Error Detection with Spectrogram and Acoustic Features”.

3 Yilmaz et al., “Articulatory Features for ASR of Pathological Speech”.
The road ahead

Thank you!