An overview of methods for articulatory feature detection

Mahir Morshed

26 October 2020

Outline

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
    - /ʈ/ in South Asian languages vs /t/ elsewhere (place)
    - /r/ vs /ɹ/ (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

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

Figure: Articulatory feature set used in [1]

- 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?)

[1] Rajamanohar and Fosler-Lussier, "An evaluation of hierarchical articulatory feature detectors".

Identifying articulatory cues

Liu, "Landmark detection for distinctive feature-based speech recognition"

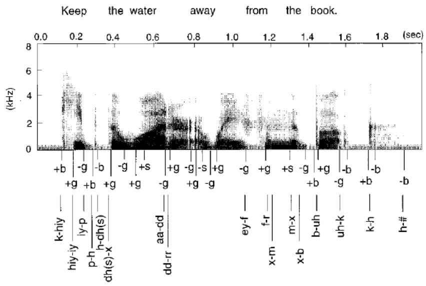

Figure: Landmarks identified using detectors for glottal vibration, sonorant closure/release, and stop bursts.

- 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%)

Recurrent 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

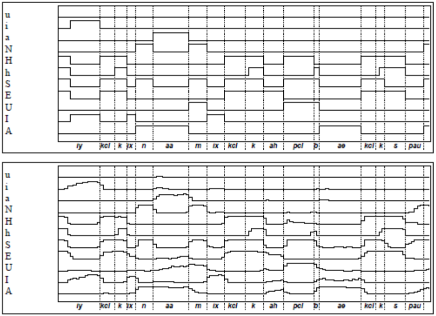

Figure: Comparison of ground truths for the phrase "economic cutbacks" and outputs appertaining from the network trained using GP.

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"

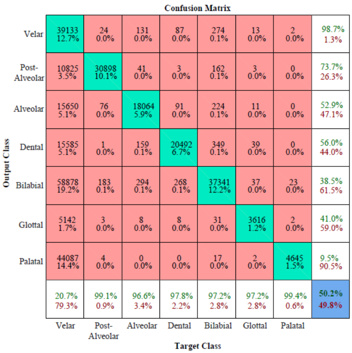

Figure: Confusion matrix for the place of articulation classifier.

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

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

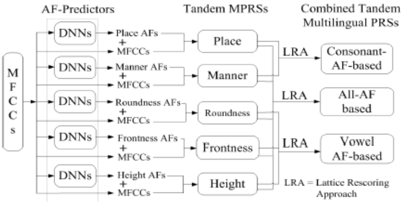

Figure: Multilingual phone recognition system information flow.

Convolutional classification

Merkx and Scharenborg, "Articulatory Feature Classification Using Convolutional Neural Networks"

- $\sim$ 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

Softmax output layer
4 x FC layer 2048
Max pooling 2x3
2x Convolational layer 3x2, 256
Max pooling 2x2
2x Convoluational layer 3x3, 256
Max pooling 2x2
2x Convolutational layer 3x3, 128
Max pooling 1x2
2x Convolutional layer 3x3, 128
Max pooling 1x2
2x Convolutional layer 3x3, 64
Mel Fbank layer
Input layer

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"

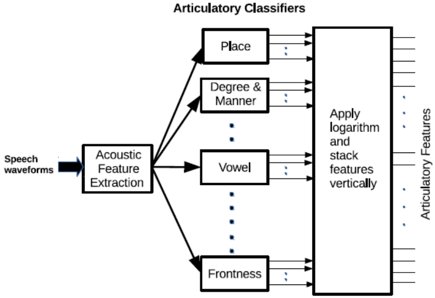

Figure: Articulatory feature extraction information flow

- 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

Aiding bottleneck features

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

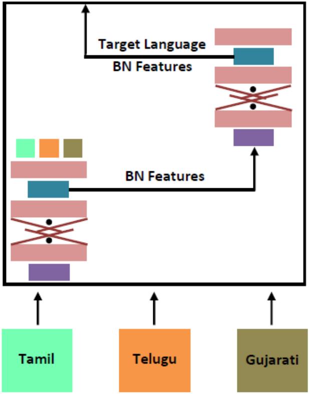

Figure: Stacked bottleneck architecture for multilingual phone recognition

- 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

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

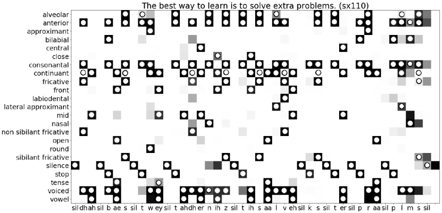

Figure: Ground truths compared with outputs from the decoder

Attributes from transformers

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

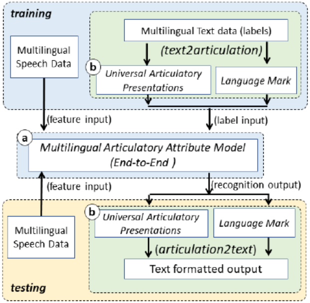

Figure: Overall architecture of the speech recognizer showing intermediate inputs and outputs thereof

- 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]
- Language model fusion [2]
- Articulograph readings as supplements [3]

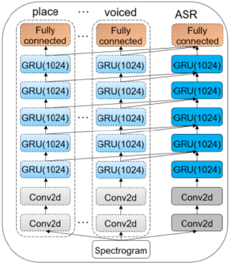

Figure: Progressive network architecture using articulatory feature detectors.

[1] 1 Qu et al., "Combining Articulatory Features with End-to-End Learning in Speech Recognition".

[1] 2 Inaguma et al., "Transfer Learning of Language-independent End-to-end ASR with Language Model Fusion".

[1] 3 Dash et al., "Automatic Speech Recognition with Articulatory Information and a Unified Dictionary for Hindi, Marathi, Bengali and Oriya".

Transfer learning elsewhere

such as with variations between same-language speakers

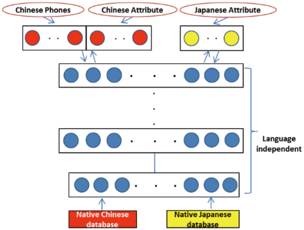

Figure: Multi-task, multilingual enhancement of a fully-connected phone recognizer.

- Accounting for differences between native- and second-language speakers [1] [2]
- Handling differences arising in pathological speech [3]

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

[1]2 Jenne and Vu, "Multimodal Articulation-Based Pronunciation Error Detection with Spectrogram and Acoustic Features".

[1]3 Yılmaz et al., "Articulatory Features for ASR of Pathological Speech".

Thank you!