Pronunciation Variability

Emily Coogan

Pronunciation Variability

- Clarity of pronunciation varies greatly

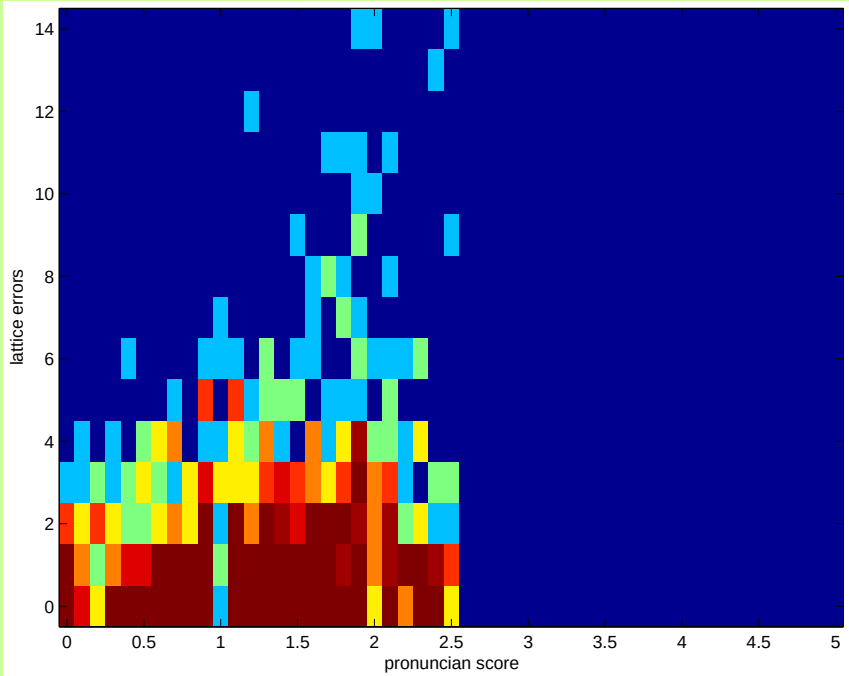
- Can we identify difficult-to-recognize speakers and utterances?
- A recognizer could make different decisions based on whether it was dealing with good or bad pronunciation, and possibly achieve greater accuracy

Preliminary Experiments

- No data labeled for pronunciation quality, so we need a computable metric
- Initially, used phone edit distance between the phonetic transcription and a dictionary of canonical pronunciations
  - Example: variety v ax r ay ih t iy
    observed: v r ay d iy
    2 deletions + 1 substitution = 3
- Currently using a weighted edit distance, depending on how many features (manner, place, or voicing) are changed
  - Same example: t→d = 1 (voicing)
    deletions worth 3 each

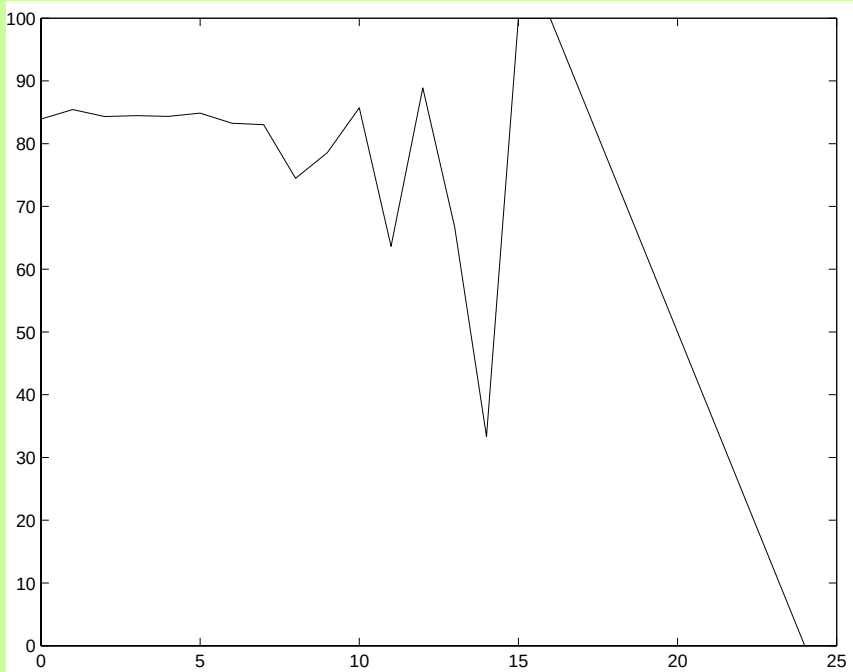
Preliminary Experiments

- Took the mean word pronunciation score over the entire utterance (x axis)
- Compared with number of word errors in the lattice output (y axis)
- Color represents density of data points at those coordinates
- Normalized correlation coefficient of 0.2105



Preliminary Experiments

- Compared score of individual words (x axis) with lattice output
- Calculated what percentage of words with that score the lattice got right (y axis)
- No evidence of correlation



Planned Experiments

- Investigate other metrics of pronunciation quality for transcribed data
  - Fix word-phone alignment
  - More precise phonetic distances
- Investigate metrics of pronunciation quality for untranscribed (test) data
  - Bushiness of the lattice
  - Difference in score of the best hypothesis and the second best
  - Pronunciation distance between the best lattice hypothesis and best landmark hypothesis

Planned Experiments

- Machine learning to score new utterances
- Speaker classification
  - Similar experiments but this time score over speaker rather than word or utterance
  - Try to identify good and bad speakers early in the conversation (after only a few utterances)
- Train separate DBN pronunciation models for good and bad speakers (or utterances) and use them in lattice rescoring

Summary

- Preliminary experiments have shown some correlation between pronunciation quality and recognizer error
- I want to identify good and bad utterances and speakers so that separate pronunciation models can be used to better recognize them