Comparative Analysis of Prosodic Features of Native and Non-native Spontaneous Speech

Catherine Lai[1], Keelan Evanini[2] & Klaus Zechner[2], *University of Pennsylvania*[1], *Educational Testing Service*[2]
laic@ling.upenn.edu, KEvanini@ets.org, KZechner@ets.org

Introduction

- The frequency of prosodic events has an impact on the perception of nativeness and fluency:
  - Liscombe (2007): distances between high boundary tones correlates with higher pronunciation scores.
  - Rosenberg (2009): higher rate of pitch accenting for of Mandarin Chinese reading English segments
    ⇝ These stuides rely on ToBI annotations. How do these labels apply to non-native speech?
- Native/non-native speech also differs in terms of fine phonetic detail:
  - Levow (2009) found that native speakers employ larger changes in pitch to mark pitch accents than non-native speakers.

This study:
⇒ Detect distinctions between native and non-native speech using automatically extractable features.
⇒ Investigate aspects of native/non-native prosody that are gradient, such as relative pitch height of accents.

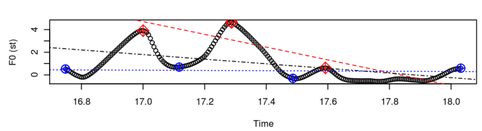
Data & Method

Corpora:

- Non-native speech: responses to the TOEFL Academic Speaking Test (TAST; 87 responses) and the TOEFL Practice Online (TPO; 90 responses).
- Native speech: responses to TOEFL iBT^{TM} items (TOEFL; 182 responses).
- Response duration: 45 - 60 seconds.

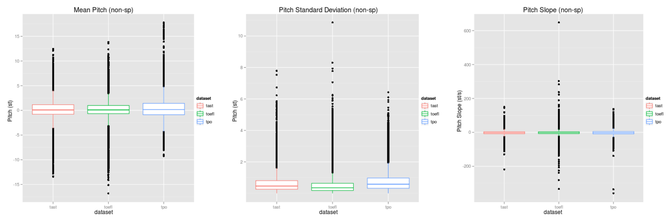
Feature extraction:

- Timing data, e.g. syllable boundaries, was determined using the Penn Phonetics Lab forced aligner.
- F0 data was extracted via Praat.
  - Pre-processing: Input parameter values for Praat were set based on estimated speaker pitch range (Evanini and Lai, 2010).
  - Post-processing: Conversion to semitones (based on speaker F0 median), removal of implausible F0 jumps, interpolated over unvoiced regions (excluding detected pauses), smoothing (Butterworth filter with a normalized cut off frequency of 0.1).
- Points of inflections in the F0 contour were detected using Mermelstein's syllabification algorithm (Yuan and Liberman, 2010) over chunks of speech (contiguous segments between aligner detected pauses).
- For each contour/chunk determine three 'declination' type lines:
  - High line: linear fit through top line points, i.e. local maxima.
  - Low line: linear fit through non-top line points,
  - Grand line: linear fit through all points in the chunk.

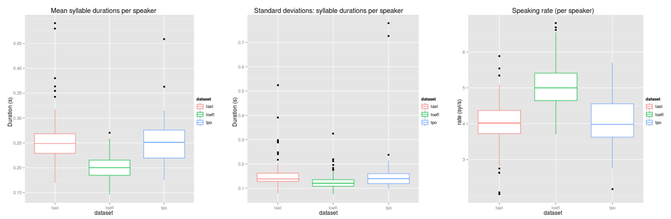


Syllable based aggregates

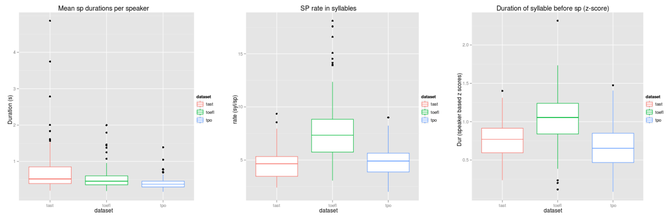
- *F0:* The differences between means for the corpora are small, e.g. differences of less than 0.3 semitones for F0 mean and standard deviations.



- *Duration:* Non-native speakers speak slower, in terms of syllables per second, and have more variable syllable durations.

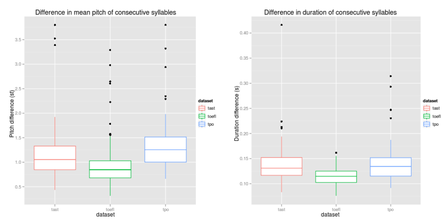


- *Pauses:* Non-native corpora (TAST, TPO) have a greater pause rate. Pre-pausal syllables are relatively longer for the TOEFL data than the TAST/TPO data (z-scores).
  ⇝ More pauses that do not express prosodic structure? i.e. disfluent pauses.



Syllable-to-syllable differences

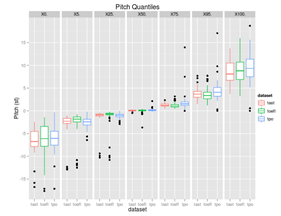
- *Syllable-to-syllable differences:* Non-native speakers are more variable locally in terms of F0 and duration.
  ⇝ Non-native speech is less monotone.



- At the syllable level, duration/pause features distinguish native/non-native speech better than F0 features.
- Looking beyond the syllable, non-native speech seems more variable in terms of F0 and duration.

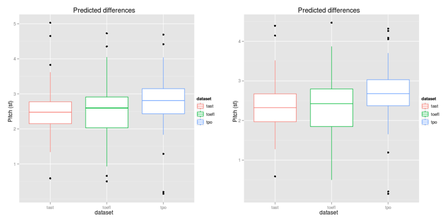
Pitch range

- *Pitch range by quantile:* TPO/TAST data is higher than the TOEFL data for the upper quantiles and lower in the bottom quantiles.
  ⇝ Non-native speakers used greater pitch range than the native speakers.

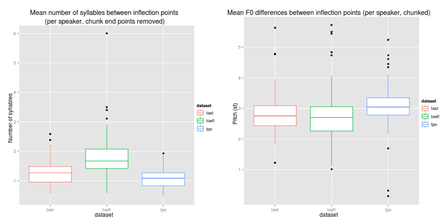


F0 Contour inflection points

- The distance between declination lines provides another way of looking at pitch range and excursion size.
- Differences between actual high points predicted low line (similarly predicted high line to low line, etc.) don't show greater excursions for native speech.
  ⇝ The greater difference in mean pitch between syllables for non-native corpora is due to greater frequency of inflection points rather than larger excursion size.



- Differences between inflection points: Inflection points are sparser in native-speech, i.e. it is more monotone.
- Mean differences in F0 between consecutive inflection points don't a significant difference between TOEFL and the TAST data (t-test, $p > 0.9$), although the TPO difference is larger ($p < 0.001$, 0.01 resp), so this does seem to be a native/non-native distinction.



Conclusion

- We are able to detect differences in the prosodic features of native and non-native speech without annotations of prosodic events.
- Non-native pitch appears more variable than that of native speech.
- The relationship between the inflection points found in our data and ToBI pitch accents remains to be investigated.
  ⇝ This approach should help illuminate the relationship between native ToBI labels and non-native prosody.

References

Evanini, K. and Lai, C. (2010). The importance of optimal parameter setting for pitch extraction. In *Presented at the 2nd PanAmerican/Iberian Meeting on Acoustics, Cancun, Mexico, 15-19 November 2010*.

Levow, G. (2009). Investigating pitch accent recognition in non-native speech. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 269–272. Association for Computational Linguistics.

Liscombe, J. (2007). *Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency*. PhD thesis, Columbia University.

Rosenberg, A. (2009). *Automatic Detection and Classification of Prosodic Events*. PhD thesis, Columbia University.

Yuan, J. and Liberman, M. (2010). F0 declination in English and Mandarin broadcast news speech. In *Proceedings of Interspeech 2010*.