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

Catherine Lai[1], Keelan Evanini[2] and Klaus Zechner[2]
[1]University of Pennsylvania, [2]Educational Testing Service

Introduction. This paper investigates the prosody of native and non-native spontaneous speech using automatically derived features. We find that non-native speakers exhibit slower speaking rate and higher rate of short pauses. With respect to pitch we find that contours of non-native speakers were less monotone than those of native speakers. That is, they show a greater rate of local extrema. However, pitch excursion size does not seem to be divided on the native/non-native line. In general, this approach allows us to examine the relationship between local prosodic changes without requiring manually annotated corpora.

Background. Previous studies indicate that the frequency of prosodic events in certain prosodic contexts has an impact on the perception of nativeness and fluency. For example, Rosenberg (2009) finds a higher rate of pitch accenting in a production study of native speakers of Mandarin Chinese reading English segments, while Liscombe (2007) found that distance between high boundary tones correlates with higher pronunciation scores. These sorts of findings highlight the importance of prosodic context when evaluating prosody. However, this sort of tone grammar modeling relies on the availability of tone labels. Manually annotated labels, usually based on the ToBI framework, are often taken as the gold standard for such data sets. However, given that the ToBI framework encodes a particular theory of English phonology, it is not completely clear how such tone labels should map on to non-native speech and whether human annotation should be taken as gold standard for this data. Moreover, in a native/non-native classification task, Rosenberg's (2009) study found that accuracy increased when using machine hypothesized labels rather than manual annotations, and more pitch accents and intonational boundaries were hypothesized than were manually identified.

As such, we would like to see whether we can detect similar distinctions between native and non-native speech using features that do not rely on manual annotations. Moreover, we would like to understand aspects of native/non-native prosody that are gradient, such as relative pitch height of accents (Levow 2009).

Data. We examined three data sets related to the TOEFL iBT® test. The two sets of non-native speech were drawn from responses to the TOEFL Academic Speaking Test (TAST; 87 responses) and the TOEFL Practice Online (TPO; 90 responses). The native speech was drawn from a study in which native speakers responded to TOEFL questions (TOEFL; 182 responses). All responses from the three data sets have durations ranging between 45sec and 60sec. F0 data was extracted using Praat. We implemented a set of F0 pre-processing steps to reduce errors associated with F0 extraction: setting input parameter values for Praat based on estimated speaker pitch range (Evanini and Lai 2010) and removal of implausible F0 jumps. The data was interpolated over unvoiced regions (excluding detected pauses) and smoothed using a Butterworth filter with a normalized cut off frequency of 0.1. Syllable boundaries were derived using the Penn phonetics lab forced aligner.

Native/Non-native prosodic features. As expected from previous studies, we find duration/fluency features to show the greatest levels of separation between native and non-native speech: non-native speech has a slower speaking rate and a higher rate of pauses (as detected by the forced aligner). On the other hand, we see more lengthening in syllables before short pauses for native speech. The non-native corpora also exhibit greater F0 standard deviations and higher overall pitch quantiles. We also find significantly greater differences between consecutive

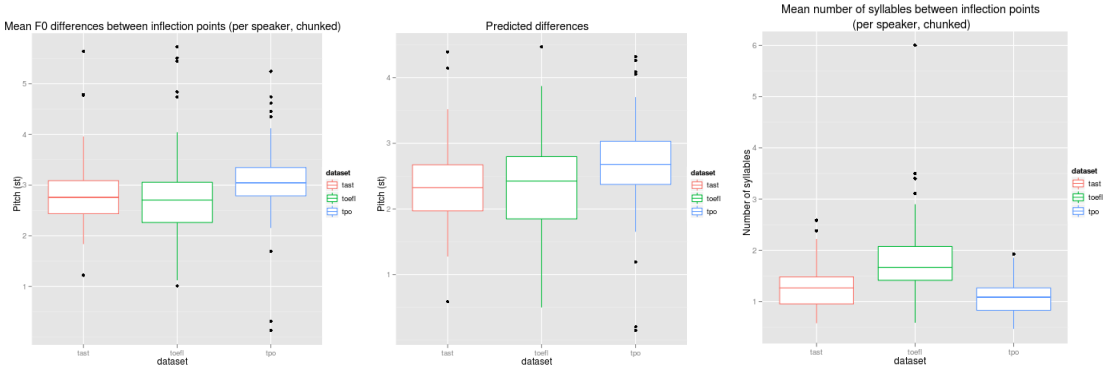
syllables for both duration and mean F0 (t-test: native vs non-native, $p < 0.001$).

Pitch extrema. To further examine pitch changes at a local level, we employed a pitch contour approximation technique based on Mermelstein's (1975) automatic syllabification algorithm (c.f. Yuan and Liberman 2010). This allows us to identify points of inflection in the pitch (i.e. local maxima, minima). This in turn gives us a novel way to look at how variable the contour is locally, as well as pitch range and pitch excursion size.

The figure below shows the mean number of syllables between inflection points per speaker where differences were calculated over chunks. Chunks were defined as contiguous intervals of speech between the short pauses determined by the aligner. Since non-native speakers produce more pauses than native-speakers, we discounted inflection points representing the start and end pitch points of the chunk to reduce inflation from frequent chunking. Even so, we found that inflection points are sparser in native-speech, i.e. it is more monotone.

We take values predicted by linear regression lines through maxima (high line) and minima (low line) to give an indication of the local pitch range. The predicted distances between high and low inflection points do not appear significantly different for the TAST and TOEFL dataset, although the difference is somewhat higher for the TPO data. Similarly, when we examine the F0 difference between consecutive inflection points there is no significant difference between TOEFL and the TAST data (t-test, $p > 0.9$), but again while the TPO difference is larger ($p < 0.001$, 0.01 resp), this does seem to be a native/non-native distinction. This suggests that the higher difference in pitch means of consecutive syllables in the non-native corpora is due to a greater frequency in inflection points, rather than the excursion sizes being bigger. This is somewhat at odds with studies finding that native speakers use larger relative pitch excursions on accented syllables than non-native speakers. However, this may be due to genre differences and requires further investigation.

Implications. We are able to detect differences in the prosodic features of native and non-native speech without annotations of prosodic events. In general, 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. G. Levow. 2009. *Investigating Pitch Accent Recognition in Non-native Speech*, Proceedings of the ACL-IJCNLP ● J. Liscombe. 2007. *Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency*, Ph.D. Thesis. Columbia University ● A. Rosenberg. 2009. A*utomatic Detection and Classification of Prosodic Events.* Ph.D. Thesis. Columbia University ● Yuan, J., Liberman, M., 2010. *F0 declination in English and Mandarin broadcast news speech*, Proceedings of Interspeech 2010.