Automatic Detection of Prosodic Focus in American English

Sunghye Cho[1], Mark Liberman[1], Yong-cheol Lee[2]

[1]Linguistic Data Consortium, University of Pennsylvania, USA

[2]Department of English language and literature, Cheongju University, South Korea

1. INTRODUCTION

- Focus highlights the most informative element in a sentence [1, 2].
- A focused element triggers prosodic prominence accompanied by increased duration, intensity, and pitch.
- It becomes prosodically distinct from its adjacent words [2, 3, 4, 5].
- Although prosodic focus has been studied extensively (e.g., [3, 7]), it has received little attention in the field of speech recognition.
- We aim to build and evaluate an automatic detection system of focus, hoping to facilitate human-machine interaction.

2. DATA

- We elicited corrective focus in telephone numbers with a Q&A structure:
  - A: Is Mary's number 887-412-4699?
  - B: No, the number is **7**87-412-4699.
- After listening to a pre-recorded question (A), 5 native speakers of American English (3F, 2M, mean age=27.8) read 100 phone numbers, correcting one wrong digit from the preceding utterance (B).
- The stimuli phone numbers (NNN-NNN-NNNN) were created to include10 digits in every string position equally frequently.

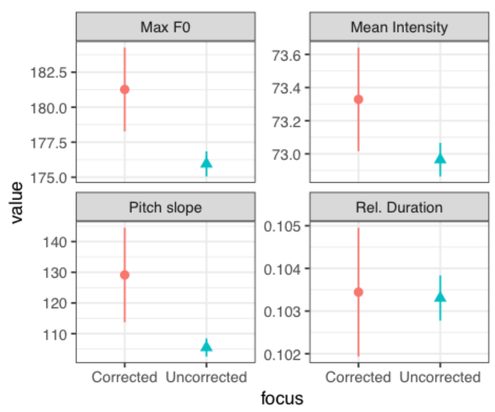
3. FEATURES

- We extracted 18 prosodic features from each digit using Praat:
  - Mean, median, min, max, IQR, max-min, sd of pitch
  - Mean, median, min, max, IQR, max-min, sd of intensity
  - Absolute and relative (= one digit / phone number string) duration
  - Pitch slope [8] and pitch excursion [9]
- One categorical variable, corrected digit, was also used.
- We z-scored all acoustic features within each digit string to capture relative differences among the digits within phone numbers.
- We imputed missing values in Python before training.
- The total number of features was 190 (= 19 features x 10 positions).

4. FEATURE & MODEL SELECTION

- We selected Random Forest classifier as our modeling framework, and trained the model to classify **the position of focused digit** within a 10-digit phone number string.
- As for feature selection, we measured the degree of correlation among the features using the basic correlation function in Python, and dropped features that had a correlation higher than 0.5 before training.
- To evaluate the generalizability of our model, we performed leave-one-group-out cross validation (CV), grouping all tokens produced by one speaker as one group.

5. EXAMPLES OF FEATURE DIFFERENCES



- Results of linear mixed-effects models:
- Focused digits have higher max pitch values (p = 0.004), higher mean intensity (p = 0.044), and steeper pitch slopes (p = 0.021).
- But relative duration does not differ by focus.

6. HUMAN PERCEPTION

- 67 native speakers of American English (mean age=19.5) participated in a perception study [10].
- We randomly selected 100 telephone digit strings produced by the five speakers and asked the listeners which digit sounds like corrected within a given phone number string.
- Participants were recruited via Qualtrics and a brief explanation about corrected focus was provided before the experiment.
- Listeners were able to correctly identify the focused digit **97.2%** of the time (range 89% to 100%).

7. CLASSIFICATION RESULTS

Feature	Mean feature importance
Median F0	0.132
Median intensity	0.131
IQR intensity	0.129
Max intensity	0.127
IQR F0	0.125

Table 1. Feature importance of selected features.

Test CV	F1-score
Female 1	0.92
Female 2	0.90
Female 3	0.95
Male 1	0.95
Male 2	0.88
Average	**0.92**

Table 2. Performance of our model (macro-average values).

8. SUMMARY & FUTURE DIRECTIONS

- We built an automatic detection system of prosodic focus and compared its performance to human listeners' performance.
- Our model correctly identified the focused position within a phone number string 92% of the time. This performance was slightly lower than the human performance (97.2%) but well above the chance level (10%).

- Future direction 1: to increase the number of examples to increase the model performance.
- Future direction 2: to add more features, such as phonation cues and spectral ones, and experiment with them
- Future direction 3: to take a frame-wise approach than a digit-wise one
- Future direction 4: to extend the project to regular sentences and natural conversations

REFERENCES

[1] D.R. Ladd, "English compound stress," in D. Gibbon and H. Richer (Eds.), Intonation, Accent and Rhythm, pp.253–266. Berlin: Walter de Gruyter, 1984. [2] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," Journal of Phonetics, vol. 33, no. 2, pp. 157–197, 2005. [3] M. S. Alzaidi, Y. Xu, and A. Xu. Prosodic encoding of focus in Hijazi Arabic. Speech Communication, vol. 106, pp. 127–149, 2019. [4] W. E. Cooper, S. J. Eady, and P. R. Mueller. Acoustical aspects of contrastive stress in question-answer contexts. The Journal of the Acoustical Society of America, vol. 77, no. 6, 2142–2156, 1985. [5] Y . Lee, and Y . Xu. Phonetic realization of contrastive focus in Korean. Proceedings of Speech Prosody, pp. 100033:1–4, 2010. [6] Y. Xu, S. Chen, and B. Wang. Prosodic focus with and without post-focus compression: A typological divide within the same language family? The Linguistic Review, vol. 29, no. 1, 131–147, 2012. [7] D. O'Shaughnessy. 1979. Linguistic features in fundamental frequency patterns. Journal of Phonetics, vol. 7, 119–145. [8] Z. Huang, L. Chen, and M. Harper, "An open source prosodic feature extraction tool," in Proceedings of the Language Resources and Evaluation Conference (LREC), 2006. [9] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," Journal of Acoustical Society of America, vol. 111, no. 3, pp. 1399–1413, 2002. [10] Y.-C. Lee. "Prosodic focus within and across languages," Doctoral dissertation, University of Pennsylvania, 2015.