Automatic Detection of Prosodic Focus in American English

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

[1]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA
[2]Cheongju University, Cheongju, South Korea

csunghye@sas.upenn.edu, myl@cis.upenn.edu, soongdora@gmail.com

Abstract

Focus, which is usually modulated by prosodic prominence, highlights a particular element within a sentence for emphasis or contrast. Despite its importance in communication, it has received little attention in the field of speech recognition. This paper developed an automatic detection system of prosodic focus in American English, using telephone-number strings. Our data were 100 10-digit phone number strings read by 5 speakers (3 females and 2 males). We extracted 18 prosodic features from each digit within the strings and one categorical variable and trained a Random Forest model to detect where the focused digit is within a given string. We also compared the model performance to human judgment rates from a perception experiment with 67 native speakers of American English. Our final model shows 92% of accuracy in detecting the location of prosodic focus, which is slightly lower than the human perception (97.2%) but much better than the chance level (10%). We discuss the predictive features in our model and potential features to add in the future study.

Index Terms: focus, prosody, machine learning, speech recognition, American English

1. Introduction

The main goal of communication is to deliver appropriate information to interlocutors. The information a speaker wants to convey needs to be structured systematically to facilitate communication. Consider the following brief dialogue:

A: Is it May 6th today?

B: No, today is the **7th**.

In B, *today is* is old information and *7th* is new and corrected information that speaker B wants to convey. In this dialogue, *7th* is the most informative part and thus receives a focus, a discourse function highlighting a particular element in a sentence [1], [2]. Given the importance of focus in communication, a focused element normally triggers prosodic prominence accompanied by concomitantly increased duration, intensity, and pitch. It, thus, becomes prosodically distinct from its adjacent words [2], [3], [4], [5], and becomes highly identifiable in perception [6]. Although prosodic focus has been studied extensively for decades (e.g., [3], [7]), it has received little attention in the field of speech recognition. This study aims to build and evaluate an automatic detection system of focus because automatic detection of focus is expected to facilitate human-machine interaction.

The success of previous studies on emotion recognition or speaker state and trait recognition has laid the foundation of this project. For example, [8] use hidden Markov models to classify seven emotions drawn from speech samples of five speakers. They extract pitch- and energy-related features from acted and spontaneous emotions and show that their model with global features correctly identifies emotions 86.8% of the time, which is higher than the human judgment (81.3%). [9] classifies speech emotions in two different corpora (one in Swedish, the other in English), using MFCC and pitch features in Gaussian mixture models, and show that the model trained with all features combined performed the best. [10] also classifies six emotion categories, using a hidden Markov model as the classifier and short time log frequency power coefficients (LFPC) as a feature. Their model correctly identifies 79.9% of Burmese utterances and 76.4% of Mandarin utterances, where the chance level is 16.67% (one out of six categories). Also, the challenge series on the emotion recognition, paralinguistics, and speaker traits at INTERSPEECH [11], [12], [13] (and subsequent challenges) promoted research on the field, showing that emotions and paralinguistic functions can be automatically detected.

The success of previous studies motivated us to develop an automatic detection system of prosodic focus. Despite the huge success and progress that have taken place in speech recognition, machines have not yet been trained to recognize focused information within a sentence or a discourse, leaving room for improvement in human-machine communications. Since acoustic features and machine learning models have been effective in predicting emotions and other paralinguistic functions from speech signals, it is reasonable to believe that prosodic focus can also be automatically detected using a machine-learning technique. In pursuit of this goal, we investigate prosodic features and develop a classifier that automatically detects a prosodic focus within a sentence.

2. Objectives

We chose prosodic focus on phone number strings as our training data for the following reasons: (i) numbers are important in human-machine interactions, such as in dialogues between voice assistants and users (for example, consider this common voice command usage scenario: VA: “Timer for 13 minutes, is this correct?” User: “No, timer for **30 minutes**.”), (ii) syntactic and morphological strategies are ruled out when focusing a digit within a phone number string so that only prosodic modulation can be used, and (iii) all positions within a string are equally susceptible to focus, which enables us to examine if a model can predict a focus regardless of variable focus positions.

To our best knowledge, this study is the first trial of building an automatic detection system of focus. Our objectives are to (i) extract and identify prosodic features that are most predictive of focus, (ii) train and evaluate a predictive model using those extracted features, and (iii) compare the performance of the trained model to humans' perception rate of focus in phone numbers.

3. Methods

3.1. Data

We collected the data set of prosodic focus in American English as a part of a larger project [14], [15], which aimed at investigating crosslinguistic commonalities and differences in focus. We elicited corrective focus, which corrects inaccurate information from the preceding utterance, using the following Q&A structure (the numbers are just for example):

A: Is Mary's number 887-412-4699?

B: No, the number is <u>**7**</u>87-412-4699.

After listening to a pre-recorded prompt question (speaker A in the above Q&A structure), five native speakers of American English (3 females, 2 males, mean age: 27.8 years) read 100 phone number strings, which were in the format of NNN-NNN-NNNN, that were different in only one digit from the preceding utterance, correcting the wrong information as if they were speaker B in the above dialogue. The participants were instructed to read the strings as naturally as possible.

The read phone number strings were created by a Python script so that every string position equally included 10 digits (from 0 to 9) and each digit in every string position was equally given a focus to counterbalance the distribution of focus. We also asked to read each digit separated (for example, '2156' as 'two one five six' instead of 'twenty-one fifty-six") and 0 as "O" instead of "zero" for consistency.

The recording session was carried out in a sound-attenuated recording booth with a Plantronics head-mounted microphone, and the recordings were saved into a laptop computer directly at a 44.1kHz sampling rate and with a 16-bit resolution.

3.2. Features

Each digit within the digit strings was manually aligned by one of the authors. We extracted 18 prosodic features from each digit using a Praat [16] script as described in Table 1.

Table 1: *Extracted prosodic features.*

Low level descriptors	Functionals
Fundamental frequency (F0)	Mean, median, minimum, maximum, Inter-quartile ratio (IQR), Difference between max and min, Standard deviation, Pitch slope, Excursion speed
Intensity	Mean, median, minimum, maximum, IQR, Difference between max and min, Standard deviation
Duration	Absolute duration, relative duration

When measuring pitch, we set the pitch range to 100Hz to 500Hz for female speakers and 75Hz to 300Hz for male speakers to reduce pitch-doubling or -halving errors. A relative duration of a digit was calculated as a proportion of a digit within a given phone number (= duration of a digit / total duration of the entire phone number string).

Besides the basic functionals such as mean, median, and standard deviation, we also measured the slope of the pitch contour and the excursion speed (Hz/sec) of each digit to capture the dynamic pitch patterns. When measuring the pitch slope, we implemented the method in [17] and as for the excursion speed, we implemented the method in [18].

We also had one categorical variable, which was the corrected digit. Since English digits vary in the number of syllables (e.g., seven vs. one), which directly affects the duration features, we hypothesized that having the corrected digit as a feature might improve the performance of the model. It is, however, important to note that the information about the corrected digit did not lead to data leakage, since the task was to identify the position of the focused digit, (e.g., the third position in 21<u>**5**</u>-123-4567), not to identify the focused digit itself (e.g., 5 in 21<u>**5**</u>-123-4567). We dummy-coded the digit information with a binary vector (1, 0) and used those values as categorical variables.

Since there were 10 digits in each phone number string, the number of acoustic features used was 180 with 500 examples (= 5 speakers x 100 phone number strings). To promote effective learning, we z-scored all acoustic features within each digit string. For example, we grouped the mean F0 values from all positions within a digit string together and z-scored the values. This is because prosodic features of focused positions are highly different from those of unfocused positions in American English (See Section 4), and the relative difference between the digits matters. We also imputed missing values in Python, where Praat failed to pitch-track due to too short duration or too much of air puff from a preceding consonant, with the median value of a feature within the given phone number string during this process, as imputing missing values is an important step for effective learning. The total number of features extracted was 190 (= 180 acoustic features + 10 categorical (from 0 to 9) features).

3.3. Model and feature selection

For better accuracy and easier model interpretability, we selected Random Forest classifier as our modeling framework. Since we had many features compared to the limited number of examples in our data and some features are likely to be highly correlated (such as the mean and median pitch values), it was important for us to select features that are informative enough. 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. This cross-validation technique was essential to prevent potential data leakage that could have been caused by random train, test splits of the examples produced by the same speaker. All processes in the pipeline were performed with *scikit-learn* [19] in Python.

4. Feature analysis

Figure 1 shows the prosodic differences between focused and non-focused digits. Focused digits have higher values than unfocused digits for all of the example features shown in Figure 1, except the relative duration. This means that focused digits were expressed with a higher pitch, intensity and a steeper pitch slope.

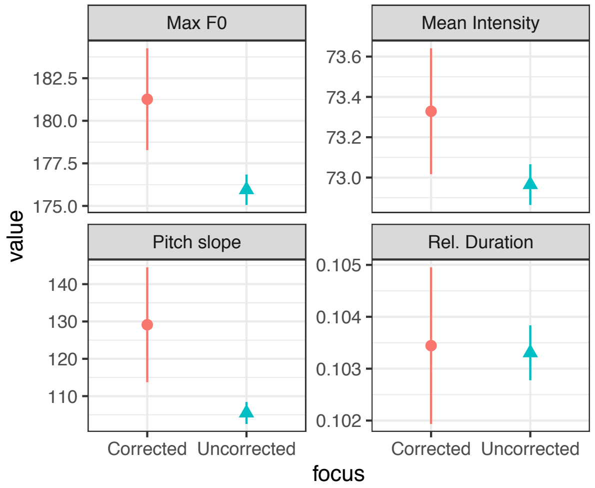

Figure 1: *Mean and standard error range of the example features. Rel. Duration is for relative duration.*

To examine if these differences are significant, we built linear mixed-effects models, with the feature values as a dependent variable, the focus condition as a fixed-effect predictor, and the speakers as a random effect, using *lmerTest* [20] in R. The models estimate that focused digits have higher max pitch values (Estimate coefficient = 4.92, t = 2.869, p = 0.004), higher mean intensity (Estimated coefficient = 0.36, t = 2.017, p = 0.044), and steeper pitch slopes (Estimated coefficient = 23.469, t = 2.316, p = 0.021), but do not have longer relative durations (Estimated coefficient = 0.0001, t = 0.084, p = 0.933). The reason for relative duration was not significant seems to be because string-final digits (NN**N**-NN**N**-NNN**N**) were subject to final lengthening. Since we only separated focused digits from the others in this analysis, string-final digits seemed to obscure the difference between focused and unfocused digits. Table 2 displays the models' random slopes by speaker, showing interspeaker variation in our data.

Table 2: *Random slopes of the models by speaker. F1-3 are female speakers, and M1-2 are male speakers; female speakers show higher values for max F0 and pitch slopes.*

Model	F1	F2	F3	M1	M2
Max F0 (Hz)	211.8	225.1	226.3	120.4	126.0
Mean Intensity (dB)	62.2	75.9	75.2	76.0	77.5
Pitch slope (Hz/sec)	154.1	132.5	150.9	90.2	118.4
Relative Duration	0.11	0.11	0.09	0.11	0.1

5. Human perception

5.1. Participants and procedure

The human perception data were adapted from [14]. 67 native speakers of American English (mean age: 19.5 years, standard deviation: 1.1) were recruited via Qualtrics, an Online platform for performing experiments. The participants were all undergraduate students who were studying at the University of Pennsylvania and their participation was compensated for a course credit.

We randomly selected 100 telephone digit strings produced by the five speakers (Section 3.1) and asked the listeners which digit sounds like corrected within a given phone number string. To make sure that the participants understood the purpose of the experiment, we provided a brief explanation about corrective focus before beginning the experiment. Only the decontextualized phone number strings were given to the participants, and the participants were able to select only one digit out of ten. They were able to listen to the stimuli as many times as they like.

5.2. Results

The listeners were able to correctly identify the focused digit 97.2% of the time. The accuracy slightly varied depending on where focus falls within a given string. The listeners were able to identify prosodic focus 99.1% of the time when it falls on the eighth digit, whereas they correctly identified focus on the fourth digit 93.8% of the time (See Table 5 in Section 6.3 for the confusion matrix). The listeners' individual scores varied from 89% to 100%, but in general, the human listeners' perception was highly accurate.

6. Classification results

6.1. Selected features

Table 3 shows the list of selected features in the order of feature importance in the model. Dropping the features that had a correlation higher than 0.5 left us 83 features, where 73 were acoustic features and 10 were the categorical feature for the corrected digit (from 0 to 9). Among 73 features, all 10 median F0, IQR F0, median Intensity, max Intensity, and IQR Intensity features (from all positions) were selected, and also included were one Max – Min F0 feature (from Digit 3), seven minimum Intensity features (from Digit 1, 2, 3, 5, 6, 7, 0, where 0 means the 10th position), six Max – Min Intensity features (from Digit 2, 4, 5, 7, 8, 0), four Duration features (from Digit 3, 6, 7, 9), two Relative Duration features (from Digit 1 and 5), and three Pitch slope features (from Digit 4, 5, 7). We summed the feature importance of a given feature of the selected positions and averaged the summed feature importance across the five cross-validation folds in Table 3.

Table 3: *The feature importance of selected features.*

Name	Mean feature importance across CV folds
Median F0	0.132
Median intensity	0.131
IQR intensity	0.129
Maximum intensity	0.127
IQR F0	0.125
Minimum intensity	0.094
Max – Min intensity	0.08
Duration	0.055
Corrected digit	0.05
Pitch slope	0.038
Relative duration	0.028
Max – Min F0	0.012

The selected features suggest that median F0 values of the digits were the most predictive feature, followed by three intensity-related values (median, IQR, and Maximum intensity). The only categorical variable, Corrected digit, was also important, but not as predictive as pitch or intensity.

6.2. Model performance

Table 4 summarizes the model performance for each CV fold.

Table 4: *The performance of the proposed model (macro-average values).*

Test CV	Accuracy	Precision	Recall	F1-score
F1	0.92	0.92	0.92	0.92
F2	0.90	0.91	0.90	0.90
F3	0.95	0.95	0.95	0.95
M1	0.95	0.95	0.95	0.95
M2	0.88	0.88	0.88	0.88
Mean	0.92	0.922	0.92	0.92

Our model could correctly classify focused digits about 92% of the time, which was lower than the human perception (97.2%) but was well above the chance level (10%, one out of the ten digits). The performance of our model is considered high, given that we had only 400 tokens for training per each CV fold. The model's performance varied depending on the test set (i.e., which speaker's tokens were presented as the test set) from 88% to 95%. It seems like the model performed relatively poorly when the test set was tokens produced by the second male speaker. This might suggest that this speaker's prosodic features were less similar to the ones of the other speakers in the training set and there is a between-speaker variation in marking prosodic focus. Since the goal of our project was to develop an automatic detection system of prosodic focus, not to investigate interspeaker variability in marking prosodic focus, we leave this observation for future study.

6.3. Comparison with human perception

In this section, we compare the model performance to human perception (Section 5.2). Table 5 displays the confusion matrices of corrective focus of the listeners and our model.

Table 5: *Confusion matrices of prosodic focus. Numbers in gray indicate correct identification rates (%). (Top: humans, bottom: machine) Correct answers are in the first column. For the machine performance, we calculated the rates from the sum of all CVs.*

Perceived/Predicted

	1	2	3	4	5	6	7	8	9	10
1	95.4	2.1	1.6	0.1	0.3	0.1	0.0	0.3	0.0	0.0
2	0.6	98.7	0.1	0.0	0.3	0.3	0.0	0.0	0.0	0.0
3	0.3	0.4	97.9	0.9	0.3	0.1	0.0	0.0	0.0	0.0
4	0.3	1.9	1.3	93.8	1.3	0.1	0.6	0.4	0.1	0.0
5	0.7	0.0	0.3	0.4	97.9	0.3	0.3	0.0	0.0	0.0
6	0.0	0.1	1.6	0.3	0.3	96.0	0.7	0.9	0.0	0.0
7	0.3	0.0	0.1	0.0	0.1	1.0	97.5	0.9	0.0	0.0
8	0.1	0.0	0.3	0.1	0.1	0.0	0.0	99.1	0.1	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	2.6	0.1	97.1	0.1
10	0.3	0.1	0.3	0.0	0.1	0.0	0.0	0.3	0.1	98.7
1	86.0	4.0	0.0	4.0	2.0	0.0	2.0	0.0	2.0	0.0
2	2.0	86.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0
3	2.0	0.0	94.0	0.0	2.0	0.0	0.0	0.0	2.0	0.0
4	0.0	0.0	0.0	94.0	0.0	2.0	0.0	2.0	2.0	0.0
5	0.0	0.0	2.0	2.0	94.0	0.0	0.0	0.0	0.0	2.0
6	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	4.0	0.0	88.0	2.0	2.0	0.0
8	4.0	2.0	0.0	0.0	0.0	0.0	0.0	96.0	0.0	2.0
9	0.0	0.0	4.0	4.0	2.0	0.0	0.0	4.0	88.0	2.0
10	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	94.0

The model performance is generally lower than the human perception, but the model performed better than the listeners in detecting prosodic focus in position 6 (human: 96%, machine: 100%) and comparable in position 4 (human: 93.8%, machine: 94%). In general, our model performed better for the boundary positions, such as positions 3, 6, and 10 than boundary-internal positions. This seems to be because the boundary digits were longer than boundary-internal digits due to final lengthening, making the duration-related features more robust for machine learning. However, when compared to the human listeners, the model performed poorly in detecting focus in the first digit group (**NNN**-NNN-NNNN), suggesting that the prosodic features for corrective focus might be weak in the first digit group.

7. Discussion and Conclusion

In this paper, we built an automatic detection system of prosodic focus and compared its performance to human listeners. We used simple and interpretable features for training, which could be adapted in developing a focus detection system in regular utterances and in large-scale speech corpora, and we rather unfolded the characteristics of focus in American English. 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%). Our model revealed that the median F0 value of each digit was the most predictive prosodic feature, followed by median intensity.

The fact that the listeners were able to correctly identify 97.2% of the time suggests that detecting prosodic focus in American English is a relatively easy task. Even though our model's performance was well above the chance level, our model's performance was 5% lower than the human accuracy. This might be because we did not have enough examples, compared to the complexity of our model. Given that the accuracy in the train sets was always 100% (high variance), adding more training examples may help to improve the model performance and increase the generalizability of the model. However, it might be also the case that prosodic features were not enough in detecting prosodic focus and native speakers might listen to other cues than prosodic features, for example, voice quality or spectral information. In particular, for the first digit group (**NNN**-NNN-NNNN), the listeners were able to correctly identify focus around 97% of the time, but our model's performance was around 89% (Table 5). This might indicate that there are other acoustic features that the native speakers are listening to. In this study, we only included prosodic features, but adding other features, such as phonation cues and spectral ones, and experimenting with them might also improve the model performance. We plan to examine both possibilities in the future study. We also plan to extend the project to regular sentences and natural conversations.

This study showed that prosodic focus could be automatically detected with a decent accuracy. We believe that automatic detection of focus would improve human-machine communication and speech recognition and help to better understand natural communication.

8. 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] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov Model-based speech emotion recognition,” *Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. II1–II4, 2003.

[9] D. Neiberg, K. Elenius, and K. Laskowski, “Emotion recognition in spontaneous speech using GMMs,” in *INTERSPEECH 2006 – Annual Conference of the International Speech Communication Association, September 17–21, Pittsburgh, Pennsylvania, Proceedings,* 2006, pp. 809–812.

[10] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” *Speech communications,* vol. 41, no. 4, pp. 603–623, 2003.

[11] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 Emotion Challenge,” in *INTERSPEECH 2009 – 10th Annual Conference of the International Speech Communication Association, September 6–10, Brighton, UK,* 2009, pp. 312–315.

[12] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, “The INTERSPEECH 2011 Speaker State Challenge”, in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 28–31, Florence, Italy,* 2011, pp. 3201–3204.

[13] B. Schuller, S. Steidl, A. Batliner, E. Noth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyber, T. Bocklet, G. Mohammadi, and B. Weiss, “The INTERSPEECH 2012 Speaker Trait Challenge,” in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, September 09–13, Portland, OR, USA, 2012.*

[14] Y.-C. Lee. “Prosodic focus within and across languages,” Doctoral dissertation, University of Pennsylvania, 2015.

[15] Y.-C. Lee, B. Wang, S. Chen, M. Adda-Decker, A. Amelot, S. Nambu, and M. Liberman, “A crosslinguistic study of prosodic focus,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015),* 2015, pp. 4754–4758.

[16] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer,” a software program, version 6.0.50, 2019, http://www.fon.hum.uva.nl/praat/

[17] 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.

[18] 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.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research,* vol. 12, pp. 2825–2830, 2011.

[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, “lmerTest Package: Tests in linear mixed effects models,” *Journal of Statistical Software,* vol. 82, no. 13, pp. 1–26, 2017.