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.
- It becomes increased duration, intensity, and pitch.

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 287-412-4699.
- 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 include 10 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] relative differences among the digits within phone numbers.
- We imputed missing values in Python before training.
- We used 18 features to distinguish digits within a given phone number string.
- 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 × 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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean feature importance</th>
<th>Test CV F1-score</th>
<th>Average F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median F0</td>
<td>0.132</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Median intensity</td>
<td>0.131</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>IQR intensity</td>
<td>0.129</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Max intensity</td>
<td>0.127</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>IQR F0</td>
<td>0.125</td>
<td>0.88</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 1. Feature importance of selected features.

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