Predicting Focus through Prominence Structure

Sasha Calhoun
Centre for Speech Technology Research
University of Edinburgh, UK
Sasha.Calhoun@ed.ac.uk

Abstract
Focus is central to our control of information flow in dialogue. Spoken language understanding systems therefore need to be able to detect focus automatically. It is well known that prominence is a key marker of focus in English, however, the relationship is not straightforward. We present focus prediction models built using the NXT Switchboard corpus. We claim that a focus is more likely if a word is more prominent than expected given its syntactic, semantic and discourse properties. Crucially, the perception of prominence arises not only from acoustic cues, but also the position in prosodic structure. Our focus prediction results, along with a study showing the acoustic properties of focal accents vary by structural position, support our claims. As a largely novel task, these results are an important first step in detecting focus for spoken language applications.

Index Terms: prosodic structure, spoken language understanding, phonology and phonetics, information structure

1. Introduction

In English, speakers use prosodic prominence and phrasing to convey which parts of each utterance are salient. To improve spoken language understanding systems, we need to harness this. Following [1], we define focus as implying a set of alternatives to the focussed word in the context. For instance, in (1) the accent \( F \) (focus)-marks \( F \) the answer (opposed to other \( F \) cuts), and in (2) the restrictor of even (opposed to the movie):

(1) Q: What would be the first thing you’d cut [in the budget]? defence?
   A: I would cut the PRISON SYSTEMS.
   \*A: I would CUT the prison systems.

(2) A: I like Michael J Fox, though I thought he was cummy in ‘The Hard Way’.
   B: I didn’t even like the PREVIEWS on that.
   \*B: I didn’t even like the PREVIEWS on THAT.

However, the relationship between accenting and focus, and between the acoustic cues to prominence and its perception, are complex (see [2]). Foci are not simply the most acoustically prominent words. In this study we automatically identify foci in an annotated corpus (NXT Switchboard [3, 4, 5]), as well as the acoustic properties of focal and non-focal accents. Our thesis is that a focus is more likely when a word is more prominent than expected in the prosodic structure given its syntactic, semantic and discourse properties.

Automatic focus detection in unrestricted speech is a largely novel task, although key for spoken language understanding in dialogue. Previous studies have looked at predicting acceptable prominence patterns for different focus structures in constructed dialogues, rather than predicting focus itself in spontaneous speech (e.g. [6, 7]). Close to our work is [8], who automatically detected focus kernels (the novel part of each utterance) and contrasts (a syntactically parallel explicit contrast) in a corpus of tutorials between children and a talking head. They report impressive accuracy figures of 84% and 93% respectively. However, their corpus of short sentences in a limited domain was more constrained than ours. Our definition is also broader: both focus kernels and contrasts are subtypes of our focus (see section 6).

2. Focus and Prosodic Structure

In the examples above, accents marked focus. However, a direct relationship between accents and focus is problematic given the existence of ‘secondary’ accents, e.g. an accent on cut in (1) or didn’t in (2) does not seem to affect the focus (see [2]).

In [9] we claim that, rather, focus is signalled through the alignment of words with prosodic structure: foci align with nuclear accents, and focal units with prosodic phrases (see also [2, 10]). Accenting and phrasing form an organic structure [2]. In each phrase, syllables align with a binary-branching structure of relative prominence, with the most prominent node being the nuclear accent [11]. Acoustic correlates of prominence, e.g. pitch, loudness and duration, are key perceptual cues to this structure. However, in English it is by default right-branching, so the last strong node is usually heard as more structurally prominent, even if it is not the loudest, highest and/or longest [11, 12, 13]. Acoustic prominence can also vary gradually, but an early accent needs to be much stronger to override expectation and be perceived as nuclear.

Importantly, we also claim that alignment between words and prosody is probabilistic. It is affected by multiple factors, including focus, syntax, and part-of-speech; as well as constraints on prosody itself, including rhythm, phrase length and emphasis. Focus status is judged from the prominence of a word in prosodic structure, given all these constraints. Non-nuclear accents can signal focus where prominence is not expected, e.g. on a pronoun. Likewise, not all nuclear accents signal focus, e.g. in a very long phrase. It is also affected by acoustic prominence: as an accent gets stronger a focus becomes more likely.

3. Highly Annotated Switchboard

Our data set was 18 NXT Switchboard conversations with integrated syntax, disfluency, focus, information status and prosody annotations [4, 5]. A Focus was a word made salient, implying a set of alternatives in the context [1]. These foci were identified via discourse triggers, e.g. contrasts and answers; all other words were Background. Only ‘contentful’ words (nouns,
verbs, adjectives, etc.) were marked, at either the word or NP level (see [5]). Since Switchboard is a lot noisier than the usual input to dialogue systems, we excluded words not annotated for focus status, and in clauses without at least one focus, as being irrelevant to the theory being tested. Each NP was also marked for its information status in the discourse, i.e. old, mediated or new (see [5]). Syntactic features such as clause and constituent type, and position in the clause/constituent were also extracted.

The prosodic annotation included phrase breaks, and accents, marked as nuclear or plain. Disfluent phrases (ToBI - p or X) were excluded. Nuclear accents were defined as the structurally, and not phonetically, most prominent in the phrase, normally the right-most. Phrasal features were extracted, e.g. speech rate, position in the phrase and mean pitch phrase. Pitch was normalised as a percentage of the speaker’s logged range, with outliers excluded. Intensity was relative to the speaker mean, and duration relative to the syllables in the word. Finally, we used an acoustic prominence measure prom, as this was more consistent than including its correlates separately:

\[
\text{prom} = \frac{(2 \times \text{duration}) + \text{quartile}_\text{pitch} + \text{range} + \text{mean}_\text{pitch} + (\text{intensity} - 5))}{10} \tag{3}
\]

### 4. Focus Prediction

Our claim is that focus is signalled through the alignment of words with prominence structure. A focus is more likely if the word is more prominent, both structurally and acoustically, than expected given its other syntactic, semantic and discourse features. Here, we build focus prediction models to test this claim.

#### 4.1. Method

We used logistic regression models to predict whether a word was a focus or background. The data set, described above, included 9289 words from 33 speakers. Each set of models included different feature groups: Semantic/Syntactic (Sem/Syn), using syntactic features and information status; Prosodic, using accent type (plain/nuclear) and phrasal features; and a combined model plus the prom measure (S+Pros+Prom). For each set, non-significant features were excluded and the models rerun.

#### 4.2. Results and Discussion

Table 1: Focus prediction using different groups of features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Baseline</th>
<th>Focus</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Backgd</td>
<td>Focus</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>100</td>
<td>0</td>
<td>60.0</td>
</tr>
<tr>
<td>Prosodic</td>
<td>80.9</td>
<td>60.1</td>
<td>72.6</td>
</tr>
<tr>
<td>Sem/Syn</td>
<td>81.4</td>
<td>65.1</td>
<td>74.8</td>
</tr>
<tr>
<td>S+Pros+Prom</td>
<td>80.0</td>
<td>73.6</td>
<td>77.4</td>
</tr>
</tbody>
</table>

Table 2: Factors which significantly affect the likelihood of a focus in the full regression model (S+Pros+Prom). The odds ratio (Exp(B)) and the % change in likelihood (P diff) are given.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Exp(B)</th>
<th>P diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mediated</td>
<td>2.26</td>
<td>20.1%</td>
</tr>
<tr>
<td>JJ</td>
<td>8.31</td>
<td>44.7%</td>
</tr>
<tr>
<td>VB</td>
<td>2.66</td>
<td>24.0%</td>
</tr>
<tr>
<td>NN</td>
<td>6.70</td>
<td>41.7%</td>
</tr>
<tr>
<td>DT</td>
<td>2.47</td>
<td>22.2%</td>
</tr>
<tr>
<td>positionWd_clause</td>
<td>3.10</td>
<td>27.4%</td>
</tr>
<tr>
<td>new by propSyl_ph</td>
<td>2.31</td>
<td>20.7%</td>
</tr>
<tr>
<td>constituent_head by propSyl_ph</td>
<td>2.72</td>
<td>24.5%</td>
</tr>
<tr>
<td>accent by JJ</td>
<td>10.44</td>
<td>47.4%</td>
</tr>
<tr>
<td>accent by PR</td>
<td>9.67</td>
<td>46.6%</td>
</tr>
<tr>
<td>nuclear by PR</td>
<td>21.36</td>
<td>53.4%</td>
</tr>
<tr>
<td>nuclear by VB</td>
<td>6.53</td>
<td>41.3%</td>
</tr>
<tr>
<td>adjunct by prom</td>
<td>1.19</td>
<td>4.3%</td>
</tr>
<tr>
<td>object by prom</td>
<td>1.16</td>
<td>3.5%</td>
</tr>
<tr>
<td>nuclear by RB by prom</td>
<td>3.35</td>
<td>29.0%</td>
</tr>
<tr>
<td>nuclear by NN by prom</td>
<td>1.89</td>
<td>15.8%</td>
</tr>
</tbody>
</table>

| Decrease                        |        |        |
| constituent_head                | 0.29   | -23.6% |
| adjusted by positionWd_constituent | 0.62 | -10.7% |
| object by positionWd_constituent| 0.62   | -10.8% |
| mediated by positionSyl_ph      | 0.47   | -16.0% |
| Constant                        | 0.12   | -      |

We can see how our theory works more clearly in a feature analysis. Table 2 shows features in the S+Pros+Prom model which significantly (p <= 0.5) affected the probability of a focus (only significant levels of categorical variables shown). For each variable, the odds ratio (Exp(B)) and the percentage difference in probability with that variable (P diff) is given. (For continuous variables, this is a one unit increase).

The types of semantic/syntactic factors which make a focus more likely are generally as expected. A focus is substantially more likely on a noun (NN) or an adjective (JJ). There is also, curiously, a moderate increase on a verb (VB) or determiner (DT), showing there may be substantial variation within these classes, e.g. with demonstratives. As expected, the likelihood increases toward the end of a clause (positionWd_clause), and is lower for constituent heads (constituent_head). Interestingly, it also higher for mediated words (mediated). This may be because its common sub-types, e.g. set or situation (see [5]), are often focussed as they pick out the part of a set being discussed. Another common subtype, general, covers commonly known entities that may nevertheless behave similarly to new entities. The interaction between info status and info type was not significant, though this may be due to data sparsity in some subtypes.

We can see that there are no absolute prosodic constraints on focus interpretation. Rather, the effect of prosody is through the interaction with semantic/syntactic features. Generally pronouns are unaccented and backgrounded. Therefore, if accent, a focus is much more likely (accent by PR and nuclear by PR). On the other hand, verbs are likely to be accented, therefore a verb must be nuclear (nuclear by VB) to increase the likelihood of a focus. Structural and acoustic prominence also interact: nouns must be nuclear and have increased prominence to increase this likelihood (nuclear by NN by prom).

There is also interaction with phrasal structure, which is intimately linked to prominence perception because of the right-branching bias (see section 2). While overall a constituent head...
is less likely to be focussed, this increases considerably toward the end of the phrase (constituent head by positionSyL_ph). The likelihood for new entities also only increases toward the end of the phrase (new by positionSyL_ph), entering the usual position of focus. Conversely, adjuncts and objects are less likely to be focussed toward the end of the constituent (adjunct by positionWd_consit and object by positionWd_consit). This could be because these elements tend to be at the end of sentences, and therefore the end of phrases. In this position a nuclear accent is expected, and therefore gives less information about focus. In order to increase the likelihood, acoustic prominence must be increased, which indeed does make a focus more likely for these elements (adjunct by prom and object by prom).

5. Phonetic Features of Focal Accents

We have shown that both acoustic and structural prominence are manipulated to convey focus. However, acoustic prominence is also a principal cue to prosodic structure (see section 2). Therefore, we need to know how these interact. We predict that nuclear accents tend to be more prominent than plain accents; and focal accents more prominent than background accents. However, the way prominence is manifested differs by accent type.

5.1. Method

A series of MANCOVAs tested the effect of Focus Status: focus marked at word level (focword), focus marked at NP level (focnp) or background on correlates of prominence for each word, i.e. maximum pitch (npmax_wd), inter-quartile pitch range (nqprange_wd), intensity (nimean_wd) and duration (dur_relSyl). We also tested the effect of Focus Status on the location and height of the accent peak relative to the stressed syllable (naccH_time and naccH). These values were not available for all accents, so the data set for these peak models was smaller. In all models covariates were included to control for the effect of phrasal position and prominence: proportion of the phrase so far, mean phrase pitch and intensity. Pre-nuclear and nuclear accent models were built using the same data set as above, excluding unaccented and post-nuclear accented words. There were 1927 pre-nuclear accented words in the general model, and 1437 in the peak model; and 2827 nuclear accented words in the general model, and 1994 in the peak model.

5.2. Results and Discussion

For pre-nuclear accents, a one factor MANCOVA showed a highly significant main effect of Focus Status (F(4,3836) = 18.0, p < 0.0001), and highly significant main effects for all five covariates (p < 0.0001). Post-hoc univariate tests (Bonferroni correction) showed that, of the four dependents tested, there was only a significant effect of Focus Status on npmax_wd (F(2,1919) = 5.4, p < 0.005) and dur_relSyl (F(2,1919) = 24.1, p < 0.0001). There was no significant effect of Focus Status on naccH_time and naccH for pre-nuclear accents. For nuclear accents, a one factor MANCOVA showed a highly significant main effect of Focus Status (F(6,5634) = 8.4, p < 0.0001), as well as for the covariates (p < 0.05). Post-hoc univariate tests (Bonferroni correction) showed significant effects of Focus Status on npmax_wd (F(2,2819) = 8.65, p < 0.0001), dur_relSyl (F(2,2819) = 7.97, p < 0.0001), and nqprange_wd (F(2,2819) = 14.1, p < 0.0001). A separate MANCOVA showed a significant main effect of Focus Status on naccH_time and naccH (F(4,3972) = 8.1, p < 0.0001).

![Figure 1](image-url) Figure 1: Acoustic Features of Pre-Nuclear and Nuclear Accents by Focus Status (y-axis estimated marginal means in normalised units) as well as for the covariates (p < 0.01). Post-hoc univariate tests showed significant effects of Focus Status on each dependent: naccH_time (F(2,1897) = 5.9, p < 0.003) and naccH (F(2,1897) = 11.7, p < 0.0001).

Figure 1 shows the estimated marginal means for significant dependents by Focus Status for each accent type, controlling for covariates. Note acoustic measures are reported in normalised units (see section 3). As expected, focwords are more prominent than background words. focnp words are in between since they include words both emphasised as the head of the focussed NP, and less emphasised words in the scope of the head. We can begin to see how prominence is manipulated to show both focus status and prosodic structure. For both pre-nuclear and nuclear accents, focussed words have higher maximum pitch. However, pre-nuclear accents are higher overall, not apparently affecting perception of prosodic structure. Conversely, while in general focal accents are longer than backgrounded accents, nuclear accents are longer overall. It seems that while both pitch and duration increase gradiente to convey focus, duration is a better cue than pitch to prosodic structure.

Figure 1 also shows the estimated marginal means for naccH_time on nuclear accents. For both focwords and focnp the peak is ‘delayed’ relative to background accents. Interestingly, despite the models being highly significant overall, nqprange_wd and naccH_time were only significant for nuclear, not pre-nuclear accents. The nuclear accent is the ‘perceptual centre’ of the phrase, therefore the most likely place for expressive variation in accent shape. In particular, below we argue that ‘delayed peaks’ may be correlated with ‘contrastive’ focus
readings. For pre-nuclear accents, merely increasing the prominence is enough to signal focus, as pre-nuclear accents are not expected to align with foci.

6. Focus Interpretation

In section 1, we defined focus as implying alternatives to the focussed word. In (1), a focus on prison systems implies not cutting defence, education, etc; while in (2), the focus on preview implies not liking the movie. In (4), the foci show the contrast between the alternatives, backyard and frontyard:

(4) I have got some [flowers] in the BACKYARD$_F$ ...
... that I would have liked in the FRONT$_F$

A key insight of alternative semantics is that every focus is theoretically contrastive, i.e. with its alternatives [1]. However, in contexts like (4) the alternatives are more available. This pragmatic contrastiveness is also cued by prominence, with conflicting evidence for distinct ‘contrastive’ accents such as L+H* (e.g. [14, 15]). We argue that, rather, the more prominent a focus, the more available its alternatives, and so the more likely a contrastive reading, e.g. an emphatic accent on pre-focus in (2) highlights the implied contrast, emphasising the speaker’s dislike. Emphatic accents (and L+H*) tend to have delayed peaks [16], explaining the finding above. Our theory thus offers the beginnings of a unified explanation of the semantic and pragmatic interpretation of focus.

7. Conclusion

In this paper, we have shown how prominence affects the perception of focus. Focus is important to spoken language understanding, and prosody is one of its primary cues. Our theory is that focus acts as a strong probabilistic constraint on the alignment of words with prosodic structure. Therefore, a focus is more likely if a word is more prominent, both structurally and acoustically, than expected given its syntactic, semantic and discourse properties. In support of this claim, we presented focus prediction models using the NIXT Switchboard corpus. We showed that the factors that most increased the likelihood of a focus involved the interaction of structural and acoustic prominence with part-of-speech, syntactic position, etc. We went on to show that the acoustic properties of focal accents differ by accent type; further supporting our contention that structural, as well as acoustic, prominence affect the perception of focus. Finally, we showed the relationship between the semantic and pragmatic interpretation of focus and increased prominence in a discussion of the status of ‘contrastive accents’.

As we discussed, automatic focus prediction is a relatively new task. It is also clear from our results that it is a hard task, with plenty of scope for development. For instance, examples like (4) show that focus interpretation is also strongly affected by the availability of alternatives in the context. We plan to try to model this using features such as word dissimilarity and syntactic parallelism (as in [8]). Further, at the moment the ‘inherent properties’ of words are modelled using linguistic features like part-of-speech. Recent work suggests this may not be sensitive enough to account for differences in behaviour within such classes [17], lexicalised features such as accent ratio may be better. Lastly, prominence has discourse functions which are not related to focus, e.g. discourse markers like Okay! which are usually accented [17]. The relationship between prominence structure and meaning remains complex, yet important to unravel for spoken language understanding.

Acknowledgements
Thanks to Scottish Enterprise Edinburgh-Stanford Link for support, and B. Ladd, M. Steedman, J. Carletta, D. Jurafsky, D. Beaver and J. Brenier for discussion and help with the corpus.

8. References