Analysis and Modeling of "Focus" in Context

Dirk Hovy[1*], *Gopala Krishna Anumanchipalli*[2,3], *Alok Parlikar*[2], *Caroline Vaughn*[2], *Adam Lammert*[4], *Eduard Hovy*[2,4], *Alan W Black*[2]

[1]CLT, University of Copenhagen, Denmark
[2]Language Technologies Institute, CMU, Pittsburgh, PA
[3]INESC-ID/IST Lisboa, Portugal
[4]USC's Viterbi School of Engineering, Los Angeles, CA
{dirkhovy,lammert}@usc.edu, {gopalakr,aup,cvaughn,hovy,awb}@cs.cmu.edu

Abstract

This paper uses a crowd-sourced definition of a speech phenomenon we have called "focus". Given sentences, text and speech, in isolation and in context, we asked annotators to identify what we term the "focus" word. We present their consistency in identifying the focused word, when presented with text or speech stimuli. We then build models to show how well we predict that focus word from lexical (and higher) level features. Also, using spectral and prosodic information, we show the differences in these focus words when spoken with and without context. Finally, we show how we can improve speech synthesis of these utterances given focus information.

Index Terms: Focus, Context, Speech Synthesis, Prominence, Emphasis, Prosody

1. Introduction

Words in a spoken sentence are not all equally important [1]. Some content words, words conveying the sentiment of the sentence, or sometimes even function words perceptually stand out from rest of the words in that sentence. Even if presented only in text, rather than speech, people have opinions on what words are salient. While the agreement on these words may not be universal [2] among all speakers, there is still a good deal of systematic predictability about them that we investigate in this work. This paper presents the analysis and modeling of where and how "focus" appears on words to make them stand out. At the speech planning and production level given the text, a number of factors determine this relative importance among words, like the given/newness of information, the socio-linguistic background of the speaker, or the mere ordering of the words in some languages. In speech, this salience has both spectral and prosodic correlates that a listener perceives to realize the speaker's intent.

In this work, we have chosen the word "focus" to denote words that people think are distinguished in a sentence. However, these words go by many names in many fields, e.g., emphasis, prominence etc. The definitions span the full linguistic hierarchy: phonetic definitions (often closely resembling definitions of prosodic stress) are related to the perceptual salience of certain elements heightened by some combination of duration, pitch and intensity [3, 4, 5], semantic definitions are based on the notion that some words introduce critical or novel information into the sentence [6]). It's generally assumed that focus functions to provide lexical or syntactic disambiguation [7, 8], or information about the "novelty" of a particular token [9]. All these definitions are not completely disparate, however, and may in fact be complementary.

Our goal here is to investigate this phenomenon not by explicitly labeling or naming it, but by finding out what listeners can consistently label and to see how well we can model these labels. Importantly, this work is trying to identify a phenomenon whose usage we can predict, and one that is realized in speech through modification of spectral and prosodic features.

We crowd-sourced annotations of "focus" over a text corpus as well as speech recordings. The annotations were collected under two conditions: (i) stimulus presented in isolation, and (ii) stimulus presented in the context of a previous sentence. Our analysis reveals interesting findings into both, how people perceive "focus" in a sentence, as well as what types of words and word positions tend to assume "focus". The annotations are also used to determine how the agreement among subjects varies when more context about a sentence is provided. We have used these results to try to model the phenomenon to predict it on unseen test stimuli. Our results suggest that it is possible to determine what the focussed words are given only the text. Further, we demonstrate empirically that this predictability can be improved by additionally using their associated speech stimuli. While several previous modeling attempts were made in this general area [10] [11], the difference in this work is the use of context, here the previous sentence.

We also conducted analyses on the speech elicited from the voice talent collected in the two different recording conditions — within and without the context of a given previous sentence. The results from this study give some valuable insights into how knowledge of context causes a speaker to alter the same sentence to make it sound coherent with the context. This aspect is missing in current day text-to-speech (TTS) systems, where synthesis invariably happens at the sentence level with little or no knowledge of how a previous sentence was delivered. With increased recent interest in the use of audiobooks [12] [13] for building high quality speech synthesis, it is important to incorporate aspects that go beyond the sentence, such as discourse level pragmatics, for both creation of high-quality voices and appropriate synthesis of multi-sentence paragraph text inputs. We conclude the current paper with some preliminary attempts in this direction of conditioning the parameters of a synthesized sentence with information about the previous sentence. The goal here is to match the delivery of human speaker's contextual focus.

*Research conducted while author was still at USC

2. Corpus and Focus Annotations

We designed a corpus that consists of about 1000 sentence pairs: (i) a target sentence we want to elicit focus annotations for, and (ii) the previous sentence as its context. The source of this text is the Brown corpus[14]. It is balanced for genre, and our subset is built to be balanced as well. The particular version of the Brown corpus we used comes distributed with the NLTK toolkit[15]. The text is processed, tokenized, and marked with part-of-speech tags. The sentences we have chosen have a length between 10 and 15 words. We believe this is the optimal length to study the phenomenon of focus—shorter sentences may be sentence fragments or headlines, and longer sentences may be too long for people to annotate reliably. We also constrained the context sentences to have at least ten words. From the list of sentence pairs that satisfied these requirements, we manually chose a subset that was balanced for genre, followed by filtering for offensive or inappropriate material. See Figure 1 for an example.

C: I try to give him as many normal experiences as possible .
S: " What is **your** experience with autistic children ? "

Figure 1: Example context (C) and stimulus sentence (S). Predicted focus word underlined

A 100-sentence subset of these sentence pairs was randomly chosen to collect speech recordings from a voice talent. The sentences were recorded by a female graduate student who speaks standard American English. She was made aware of the purpose of the recordings as being the study of the phenomenon of focus, but was asked to deliver the sentences naturally. Recordings were performed in two settings. In the first session, we presented the sentence pair and recorded both the context sentence as well as the intended stimulus sentence together. The speaker was allowed to read the sentence pair ahead of the recording to make her aware of its context. In a second session (after a few days), the speaker was presented only the stimulus sentence to record.

For focus annotations, we recruited volunteers on Amazon Mechanical Turk (AMT) to mark the words they think are focused in a sentence. Given the different modalities in our corpus, we conducted two annotation tasks on AMT. For the text part of the corpus, annotators were presented with the sentence pair and asked to mark the words in the second one that they thought they would focus. The subjects had to choose at least one focus word, but could optionally select a second one.

For the isolated speech setting, subjects were presented with the stimuli recorded in isolation and were asked to annotate the word they perceive as being focussed. Finally for the with-context recordings, the presented stimulus contained both the context and the stimulus sentence. The subjects again were asked to identify the focussed word only in the second sentence. Speech annotations were only elicited for the primary focused word in the sentence. TestVox [16] is used to carry out annotations of the speech part.

While most subjects on AMT submit their responses genuinely, there are a sizable number of users who try to game the system to maximize their pay. We need to discount annotations from these participants. Also, the inherent ambiguity of which word should be focused makes this process particularly hard. There is often no single correct answer, so annotators tend to disagree. To handle these problems, we use the MACE algorithm [17] to combine the submissions from multiple annotators and obtain an optimal average annotation that is weighted by the estimated confidence in each annotator.

To determine inherent acceptable ambiguity in focus annotation, we need to evaluate the inter-annotator disagreement. The popular kappa metric [18, 19] is not appropriate in this scenario, since not all annotators are equally reliable. To measure agreement, we divided our annotators randomly into two groups, ran MACE on both, and computed the overlap between the predictions of the two groups. The results in Table 1 show that annotations vary considerably, an indicator of the task difficulty.

Setup	Agreement
Text	0.42
Speech w/ Context	0.21
Speech w/o Context	0.42

Table 1: Agreement of Focus Annotations

3. Focus in Text

In this section, we work with the annotations on the text portion of our corpus. We present an analysis of its properties, and describe a model that can predict focus annotations given a new sentence and its context. The reference labels for comparison are the annotations from AMT estimated with MACE.

3.1. Analysis of Focus Annotations

First we look at the annotator agreement on focus words. Considering only the primary focus, the averaged binary raw agreement among all annotators on the task was 0.27. Removing annotators that received a low competence estimate (< 0.5) by MACE, the agreement improves to 0.45. This agreement is respectable, given that these are per-sentence comparisons.

We further studied the focus annotations under two questions – (i) Are certain positions in a sentence more likely to be focused? and (ii) Are certain categories of words (e.g., parts of speech) more likely to be focussed? We select the most likely answer for each sentence as estimated by MACE and use them as reference annotations.

Figure 2 shows the distribution of how likely different positions in a sentence are to be focused. Since our sentences were of varying lengths, we used relative positions of the words in the sentence. For this dataset, we observe: (i) People can reliably discriminate between primary focus and secondary focus. (ii) Primary focus tends to placed early on in the sentence, whereas secondary focus happens towards the end. (iii) Primary focus distribution has two peaks, whereas secondary focus has one clear peak, suggesting that primary focus is more subtle and more subjective, presumably due to its dependence on the semantics and word order.

We also looked at the relative position of primary and secondary focus wrt one another. Annotators chose to annotate a secondary focus word in 330 instances of our data. The average distance between the primary and secondary focus word was about 4.7 words.

To find the relative importance of word categories in assuming focus, we clustered the focus words into their respective parts of speech (POS). Figure 3 shows the distribution of primary focus over the POS. We can see that nouns and verbs

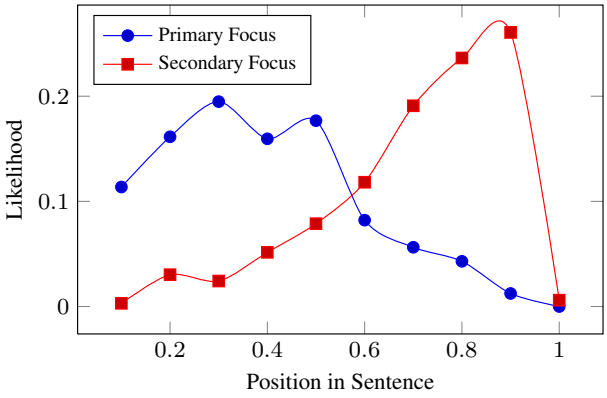

Figure 2: Relative word position of focus words in a sentence.

are the main groups that are focused. The same analysis for secondary focus produces similar patterns, with the important distinction that verbs were more likely to be (secondary) focussed than nouns. This suggests the two are complementarily distributed over nouns and verbs.

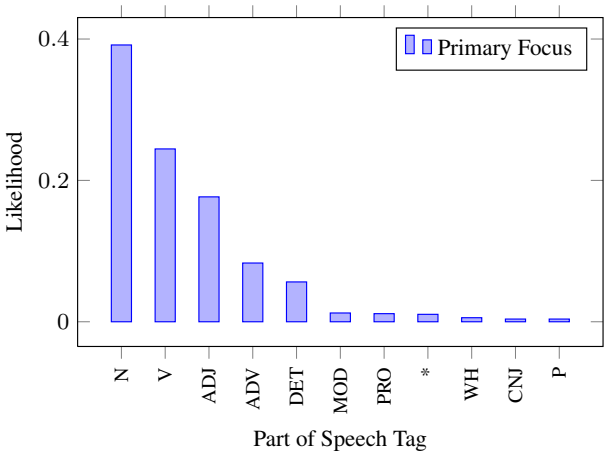

Figure 3: Distribution of parts of speech of focus words.

3.2. Focus Prediction from Text

Given the findings in the previous section, we attempt to predict the focussed word on an unseen sentence automatically. This has an immediate practical application in speech synthesis, where the knowledge about which word to focus improves the expressiveness of synthetic speech. In our experiments, we formulate focus prediction as a binary classification task: we classify each word in a sentence as being focused or not, and then select the word with the highest likelihood in the sentence as focus word.

We trained a support vector machine (SVM) classifier using SVMlight [20] to predict word focus. The training data comprised only the primary focus annotations. We split our entire text corpus (around 1000 sentence pairs) into 80%-10%-10% splits of training, development and test sets and extracted relevant features over all words, using the FANSE parser [21] and WordNet [22]. Table 2 lists the set of binary features used in the classifier (1353 dimensions). To find the contribution of context, we first predict the focus word by using features *only* from the stimulus sentence. We then build another model to predict the focus word that also uses information from the previous (context) sentence. Given the skewed distribution of the training data, with a vast majority of words without focus, we reweighed the positive (focussed) samples to a comparable level. This weight, as well as other SVM parameters (C, γ) were tuned against the development set.

Sentence Length	Word Length
Word POS	Word position
Positions of Content Words	Word Supersense
Word: first-noun?	Word: content-word?
Word: first-verb?	Word: function-word?
Word: in-list?	POS, Supersense, Semantic Role
Prev-Word: Lemma	

Table 2: Lexical features used in SVM training.

We compare the performance of the proposed classifier with respect to two baselines, based on the analysis we presented in the previous sections: (i) picking the first noun in the sentence, and (ii) picking the word in the relative position that was observed most frequently in training data (around the third word).

We present our comparison in terms of two metrics: Accuracy and Mean Reciprocal Rank (MRR). We calculate accuracy on a sentence level, i.e., the portion of sentences in our test corpus for which we accurately predict the focus word. MRR shows the average rank of the correct word in our prediction (this measure does not apply to the baselines).

Table 3 shows that our model performs significantly better than the baselines.

System	Accuracy	MRR
Baseline: first-noun	0.13	—
Baseline: most-likely-position	0.20	—
SVM	**0.44**	**0.62**

Table 3: Prediction accuracy per sentence for various systems

Counterintuitively, we find that contextual features do not improve accuracy. This suggests that mere co-occurrence is not sufficient, but that we have to take deeper semantic and syntactic links to the previous sentence into account.

4. Focus in Speech

In this section, we present the analysis of focus in the speech data, and build predictive models of the phenomenon. Note, however, that the speech data is much smaller (100 sentence pairs), making it prohibitive to run any large scale classification experiments.

4.1. Predictability of Focus in Speech

Prosody, particularly the fundamental frequency (F0), is widely held to be the primary correlate of perceived focus [1]. To verify this in the speech data, we ran a simple classification experiment, in line with the text-based focus prediction. However, given the small amount of data we resorted to using Classification and Regression Trees (CART) as a regression model on a simple set of features (Table 4) derived from the lexical and

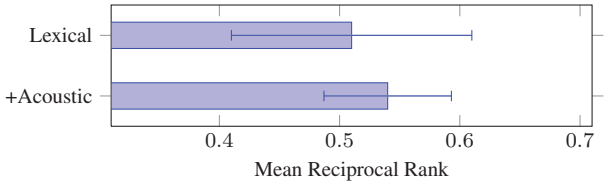

Figure 4: Mean Reciprocal Rank of focus prediction using lexical and acoustic features.

acoustic properties of each word in the stimulus and in the context. Output is a value between 0 and 1, where 1 is truly focussed.

lexical	acoustic
word : content ?	word duration
word POS	error in predicted dur.
#content words in left context	F0 TILT parameters
#content words in right context	F0 difference neighboring context
#words in left context	F0 mean in previous context
#words in right context	F0 max in context
#syllables in word	

Table 4: Text & speech features used in CART training.

The annotations obtained from AMT via MACE were used as reference values. We used k-fold cross validation over 10 different 80%-20% splits of the data. Two experiments are performed where the CART trees use i) only lexical features, and ii) lexical + acoustic features in the stimulus and context. For evaluation, the average MRR is computed across the 10 folds of cross-validation. Figure 4 compares the performance for both experimental settings.

The results indicate that using acoustic features improves the MRR considerably, and that the true focussed word is always within the top two model predictions. The lower standard deviation also indicates that the predictions are more reliable when acoustic features are used in conjunction with lexical features. We also found F0 related features as being the most informative to predict focus. To analyze any remarkable trends exhibited by the F0 contours in context and in isolation, we conducted the sentence-level analysis reported here.

In the case of our speech data, where we have the same sentence spoken both in isolation and in context, F0 is comparable. To study the explicit effect of context on speech, we measure the following global F0 related parameters: i) maximum value of F0, ii) mean value of F0, iii) mean F0 of first content word, iv) mean F0 of final content word, and v) dynamic range of F0.

Table 5 compares the Pearson's correlations among various recording conditions. 'isolated' and 'stimulus' respectively correspond to the same sentence spoken without and with context; 'context' denotes the context sentence provided. The mean and range (standard deviation) of the F0 for these conditions are shown.

Correlation	F0mean	F0range
isolated–stimulus	0.45	0.23
context–stimulus	0.23	0.22
context–isolated	0.13	0.13

Table 5: Correlations of F0 mean/range for various conditions

The numbers show that the F0 statistics are more correlated between the previous sentence and the stimulus recorded in context, as opposed to the previous sentence and the isolated recording. This implies the speaker employs systematic linear changes to the F0 statistics when speaking in context. For the statistics corresponding to starting F0, ending F0 and maximum value of F0, the averages of these values across all the utterances are illustrated in Figure 5.

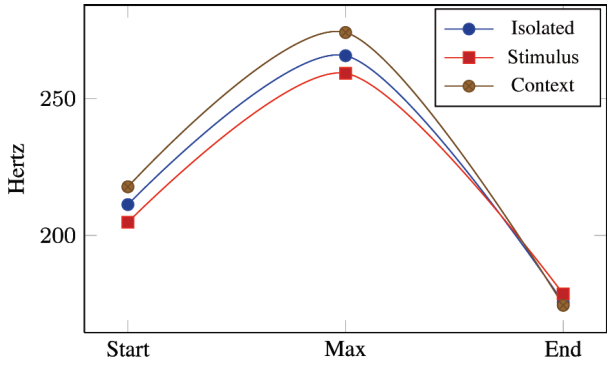

Figure 5: Analysis of mean starting/max/ending F0 across for different utterances

It is clear that these F0 statistics are consistently lower for sentences spoken in context than their counterparts spoken in isolation. These changes are likely in adjustment to the context of the previous sentence. While the data is not sufficient to run full-scale TTS experiments, we present here a simple experiment where the maximum value of F0 is predicted for a text stimulus, given the spoken context of its previous sentence.

4.2. F0 statistics prediction with context

To investigate if the observed trends in F0 analysis can be simulated, we created a regression model that takes into account the features corresponding to the stimuli (context and current sentence) and the F0 statistics of the context sentence to predict the maximum F0 value in the current stimulus. Over a 10-fold cross validation we observe that the F0 peak can be predicted with a lower error if the contextual features of the previous utterance's F0 are included as features in the prediction.

	without context	with context
F0 max RMSE	18.48	17.24

Table 6: Average F0 prediction error over 10 fold cross validation

5. Conclusion

We designed a corpus of sentence pairs to investigate the effect of context on a perceivable and model-able phenomenon we call "focus". We also recorded speech data to study the effect of context, presenting qualitative and quantitative aspects of the corpus with annotations collected on Amazon Mechanical Turk. Our preliminary results show the importance and inter-dependence of production and perception of "focus" in context. We intend to scale up the analyses and experiments presented here to audiobooks and towards complete prediction (as opposed to the illustrative maximum F0 prediction currently shown) of appropriate F0 contours that are sensitive to their synthesis context.

6. References

[1] D. Bolinger, *Intonation and its Uses*. Stanford University Press, 1989.

[2] Y. Mo, J. Cole, and E.-K. Lee, "Naive listeners prominence and boundary perception," *Proc. Speech Prosody, Campinas, Brazil*, pp. 735–738, 2008.

[3] M. E. Beckman, *Stress and non-stress accent*. Foris Publications USA, 1986, vol. 7.

[4] E. Fudge, *English word-stress*. Allen & Unwin London, 1984.

[5] D. v. Kuijk and L. Boves, "Prosodic stress revisited: Reassessing the role of fundamental frequency," in *Speech Communication*, vol. 27, 1999, pp. 95–111.

[6] M. A. Halliday, "Notes on transitivity and theme in english: Part 2," *Journal of linguistics*, vol. 3, no. 02, pp. 199–244, 1967.

[7] I. Lehiste, *Suprasegmentals*. MIT Press, 1970.

[8] P. I. Blok and K. Eberle, "What is the alternative? the computation of focus alternatives from lexical and sortal information," *Focus: Linguistic, Cognitive, and Computational Perspectives*, p. 105, 1998.

[9] R. Silipo and S. Greenberg, "Automatic transcription of prosodic prominence for spontaneous english discourse," in *Proc. XIVth Int. Cong. Phon. Sci.*, 1999, pp. 2351–2354.

[10] D. Wang and S. S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 690–701, Feb. 2007.

[11] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "A Statistical Phrase/Accent Model for Intonation Modeling," in *Interspeech 2011*, Florence, Italy, 2011.

[12] K. Prahallad and A. W. Black, "Segmentation of Monologues in Audio Books for Building Synthetic Voices," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 5, pp. 1444–1449, 2011.

[13] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Accent Group Modeling for Improved Prosody in Statistical Parametric Speech Synthesis," in *ICASSP 2013*, Vancouver, Canada, 2013.

[14] W. N. Francis and H. Kucera, "Brown corpus," *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Revised*, 1971.

[15] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media, Incorporated, 2009.

[16] A. Parlikar, https://bitbucket.org/happyalu/testvox/wiki/Home.

[17] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning Whom to trust with MACE," in *Proceedings of NAACL HLT*, 2013.

[18] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[19] J. Cohen *et al.*, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[20] T. Joachims, "A support vector method for multivariate performance measures," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 377–384.

[21] S. Tratz and E. Hovy, "A fast, accurate, non-projective, semantically-enriched parser," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1257–1268.

[22] C. Fellbaum, *WordNet: an electronic lexical database*. MIT Press USA, 1998.