Detecting prominence in conversational speech: pitch accent, givenness and focus

Vivek Kumar Rangarajan Sridhar[1], Ani Nenkova[2], Shrikanth Narayanan[1], Dan Jurafsky[3]

[1] University of Southern California
[2] University of Pennsylvania
[3] Stanford University

vrangara@usc.edu, nenkova@seas.upenn.edu, shri@sipi.usc.edu, jurafsky@stanford.edu

Abstract

The variability and reduction that are characteristic of talking in natural interaction make it very difficult to detect *prominence* in conversational speech. In this paper, we present analytic studies and automatic detection results for *pitch accent*, as well as on the realization of information structure phenomena like *givenness* and *focus*. For pitch accent, our conditional random field model combining acoustic and textual features has an accuracy of 78%, substantially better than chance performance of 58%. For givenness and focus, our analysis demonstrates that even in conversational speech there are measurable differences in acoustic properties and that an automatic detector for these categories can perform significantly above chance.

1. Introduction

In natural conversation, speakers make some words and phrases more prominent than others. These *pitch accented* words [1] are perceptually more salient to the listener and are presumably employed at least in part to draw the listener's attention to informationally *salient* words. There are many attempts to characterize what it means to be *salient* based on different aspects of *information structure*. For example words bearing pitch accent are contextually unexpected compared to non-prominent words [2, 3]. The degree of *givenness* of a referent seems also to be predictive of its prominence and speakers introduce new items by accenting them, while deaccenting familiar or old items [4, 5]. Finally, the *focus* of an utterance, semantically its most salient part, is also predicted to be the most prominent [6].

We report results on the automatic detection of prominence which explore the relationship between these three measures of information structure and the acoustic correlates of prominence in a richly annotated corpus of conversational speech. Conversational speech presents more complexities than read speech (massive reductions, disfluencies, pauses) and consequently, presents a greater challenge for automatic detection of prosodic structure.

Our first study employs Conditional Random Fields (CRFs) to combine text-based and acoustic features to detect *pitch accents* in conversational speech. This is a well-studied area, and our system achieves good performance. We then present two analytical studies on a much less well-studied task: the automatic detection of information structure, specifically *givenness* and *focus*.

2. Data and features used

Our experiments use a subset of the Switchboard corpus that had been hand-labeled with pitch accent markers [7]. 12 conversations from [7], amounting to 14,555 word tokens, had been manually annotated with additional tags such as givenness [8] and focus distinctions [9], features that the linguistic literature suggests are predictive of prominence [4]. The pitch accent labels are binary (accented or unaccented). A 10-fold cross validation was performed for testing, in all the experiments.

In our work we use acoustic, lexical and part of speech features to automatically detect prominence as associated with pitch accent, givenness or focus. The lexical features are word identity and accent ratio [10]. Accent ratio is an estimate of the proportion of times a given word was accented in a training corpus:

$$Accent\ ratio(w) = \begin{cases} \frac{k}{n} & \text{if } B(k,n,0.5) \leq 0.05 \\ 0.5 & \text{otherwise} \end{cases} \quad (1)$$

where k is the number of times word w appeared accented in the corpus, n is the total number of times the word w appeared, $B(k,n,0.5)$ is the probability (binomial distribution) that k successes occur out of n trials. Accent ratio was computed over 60 Switchboard conversations annotated for pitch accent [7] to compute k and n for each word.

A variety of acoustic feature were used. We extracted the pitch and energy contour over 10 msec intervals for each word and computed a set of representative statistics of the raw and speaker normalized pitch contour, duration and energy such as mean, standard deviation, slope, etc. The set of 26 features used in the experiments are summarized in Table 1.

3. Pitch accent detection

There is a vast literature on predicting and detecting pitch accents. To briefly summarize the most relevant studies, [11] recently showed that a maximum entropy classifier using local features (lexical, syntactic and acoustic) achieved good results on detecting accents in read speech, [12] showed that conditional random fields (CRFs) offer a good way to capture contextual influences for accent detection on Switchboard, and [10] showed that *accent ratio* was a powerful lexical feature.

We combined these three ideas to investigate the combination of a wide variety of acoustic and textual features in a conditional random field (CRF), a graphical model that conditions on an observation sequence [13]. While the maximum entropy model makes a decision for each state independently of other states, CRFs optimize over an entire sequence. Our work thus

Features used	Description
ling_dur	duration of word
f0_mean	f0 mean of word
f0_mean_ratio	ratio of f0 mean in word and conversation side
f0_mean_zcv	f0 mean in word normalized by mean and std dev of f0 values in conversation side
f0_std	std dev of f0 values in word
f0_std_ratio	log ratio of std dev of f0 in word and convside
f0_max	maximum f0 value in word
f0_min	minimum value of f0 in word
f0_mean_first	f0 mean in first half of word
f0_mean_second	f0 mean in second half of word
f0_slope	linear regression slope over all points of word
f0_slope_first	linear regression slope over first half of word
f0_slope_second	linear regression slope over second half of word
abs_f0_diff	difference in f0 mean in the second and first half of word
rel_f0_diff	ratio of f0 mean in second and first half
norm_end_f0_mean	f0 mean in second half normalized by mean and std dev in conversation side
norm_pen_f0_mean	f0 mean in first half normalized by mean and std dev in conversation side
norm_f0_diff	difference in f0 mean of second and first half, normalized by mean and std dev of f0 in convside
e_mean	mean rms energy in word
e_std	std dev of rms energy in word
e_mean_first	rms energy in first half of word
e_mean_second	rms energy in second half of word
abs_nrg_diff	difference between rms energy of second and first half
end_nrg_mean	mean rms energy in the second half
norm_nrg_diff	difference in normalized mean rms energy of first and second half
rel_nrg_diff	ratio of mean rms energy in second and first half

Table 1: Prosodic features used in the experiments organized by duration, pitch and energy categories.

		CRF		
Features used	Decision tree	current token	$\pm$ 1 tokens	$\pm$ 2 tokens
Words	67.94	68.13	75.66	75.11
POS tags	69.68	70.19	72.69	73.24
Accent ratio	75.51	75.31	75.24	75.14
Words+POS+Accent ratio	68.06	75.37	76.04	75.85
Acoustic only	75.84	73.12	73.99	73.93
All features	77.46	77.38	78.31	78.22

Table 2: Pitch accent detection accuracies (in %) using a decision tree and CRF for different features.

draws on the previous use of CRFs in pitch accent detection [12] while adding a much wide variety of rich lexical features (including word identity and accent ratio), and acoustic features ([12] only used duration, speaking rate and pause).

We chose the sentence as the basic unit for sequence labeling. Other choices such as turn, intonation phrases or pause-delimited fragment could be explored in future work.

In order to clearly demonstrate the benefit of the CRF model, we first use the features described in Section 2 in a simple decision tree classifier that uses cues from the current token alone. The results are presented in Table 2. For the CRF setting, in addition to using the lexico-syntactic features described in Section 2, we used only the most informative prosodic features in the decision tree classifier. The features were selected using an information gain metric. The most informative prosodic features that we used in our CRF model are ling_dur, f0_std, e_std, norm_f0_diff, rel_f0_diff and f0_slope. Interestingly, the raw acoustic features rather than the speaker normalized ones were the ones more helpful for the pitch accent prediction task.

Because the CRF++ toolkit that we used for our training does not support real-valued features, we had to discretize the acoustic features by taking their logarithm and performing uniform quantization. The duration feature was quantized into 5 bins and the other features were quantized into 10 equally spaced bins. The results are again presented in Table 2. ± 1 denotes the use of cues from the preceding and succeeding words and ± 2 denotes a window of 2 preceding and succeeding words, respectively.

The first implication from our results is the improvement (1.1% relative) in moving from a standalone per-word decision tree classifier to a sequence model. The ability of the CRF to optimize over a sequence is clearly important for the pitch detection task. Second, our results suggest that this contextual information is not particularly long-distance. A window of one preceding and succeeding word results in the highest classification accuracy of 78.31%; adding more context degrades performance. One possible explanation for these results is the known dispreferrence of speakers for *accent clash*, i.e. having two accented words right next to each other. The immediate context is sufficient to provide information about potential clash.

Acoustic features alone lead to good performance, indicating that despite the variability of conversational speech, prominence is marked acoustically in a systematic way. As in previous studies, the lexicalized accent ratio feature leads to very good performance, which is only slightly improved by adding more textual features (75.24% accuracy to 76.04%). The addition of acoustic features on the other hand improves accuracy much more, reaching 78.31%.

4. Givenness realization and detection

The givenness of a referent plays an important role in prominence decisions [14] and in the choice of referring expression [15]: new information is more likely to be acoustically prominent and realized using full noun phrase, while given information is reduced and often pronominalized. The annotation we relied on is based on the givenness hierarchy of Prince [16]: first mentions of entities were marked as *new* and subsequent mentions as *old*. Entities that are not previously mentioned, but that are generally known or semantically related to other entities in the preceding context are marked as *med*iated. Givenness annotation applies only to referring expressions, i.e. noun phrases whose referent is a discourse entity. Complete details of the annotation can be found in [8].

Is it possible to automatically distinguish between *new* and *old* items occurring in natural conversational speech, based on their acoustic properties? In order to answer this question, we first perform binary classification, combining the *med*iated class with *new* (*new+med* versus *old* distinction). We use a decision

tree classifier for this task. CRFs are not directly applicable here because the givenness tags are defined only for certain words. Since *Old* entities are systematically referred to by using a pronoun, we use part of speech distinctions (noun or pronoun) as a competitive baseline which achieves 88.29% accuracy. The classifier based on acoustic features is less accurate (79.09%) then the POS baseline, but still significantly outperforms the majority class (*old*) baseline, 53.98% (table 3).

Features used	Accuracy (%)
NN and PRO tags	88.29
Acoustic only	79.09
NN and PRO tags + acoustic	88.30

Table 3: Information status detection accuracies using decision tree for different features - *old* vs *med+new*

The results from the overall givenness classification, even based on acoustic features alone, are good, but they do not give a clear indication of how different the acoustic realizations of *nouns* from the three classes of givenness are. It is possible that most of the separability of the classes comes from differences between full nouns and pronouns. In order to examine the issue more closely, we further investigated the differences between nouns only. Table 4 gives the overall distribution of nouns in givenness classes and their realization as bearing a pitch accent or not.

NN	**pitch accent**	**no accent**
new	345	75 (18%)
old	186	69 (27%)
med	926	263 (22%)

Table 4: Distribution of nouns in givenness classes and presence or absence of pitch accent

As predicted from theories of givenness, the rate of *new* nouns that are prominent (bear pitch accent) is higher (and statistically significantly so) than that for the *old* and *med*iated categories: 82% vs 73% and 78% respectively. In order to establish differences between the givenness categories at a finer acoustic level, we performed analysis of variance with the acoustic features as dependent variables and the givenness classes as factors, followed by paired comparisons between each two classes,using Tukey's adjustment . The following acoustic *individual* features were significantly different:

new-med ling_dur, e_mean, e_mean_second, e_mean_first, end_nrg_mean, f0_mean_second, utt_f0_slope.

new-old ling_dur, e_mean, e_mean_second, e_mean_first, end_nrg_mean.

Our findings are consistent with previous work [14] in which, in a controlled experiment setting, it was found that the most salient difference between new and old nouns is in terms of duration. In that study amplitude was also found to be significantly different between the two classes, while the results for differences in fundamental frequency were weakest. Interestingly, no single acoustic feature was significantly different between the nouns in the *old* and *med*iated categories, suggesting that collapsing *old* and *med* classes may make more sense in binary classification tasks based on acoustic features.

Table 5 shows the detection accuracy of a decision tree classifier combining all acoustic features, downsampled classes with equal number of examples for each class. As the analysis

Classification	Accuracy (%)
new vs med	63.80
new vs old	57.84
old vs med	61.17

Table 5: Classification accuracy for noun givenness based on acoustic features. Chance level performance is 50%

of individual acoustic features indicated, there are biggest differences between the *new* and *med*iated classes, with classification accuracy 14% above chance level. Interestingly, while no individual acoustic feature was significantly different between the *med* and *old* classes, the *combination* of features achieved considerable improvement above chance performance (11%).

Our experiments with givenness distinctions show several important facts. First, even in conversational speech, there are measurable acoustic and prosodic differences between nouns with different givenness status. Second, somewhat unexpectedly, nouns from the *old* and *med*iated categories are very similar to each other prosodically and acoustically, with similar accenting rates and acoustic features that distinguish them from nouns in the *new* class. Finally, the combination of acoustic features can reliably distinguish among any pair of classes above chance levels.

5. Focus realization and detection

Some entities in an utterance are particularly salient because they are *focused*, i.e. contrasted with other semantically related entities [6]. Several classes of focus were marked in the the [9] corpus that we used: *adverbial* (when a focus-inducing particle such as "only" or "just" is used), *contrastive* (direct comparison of two lexical items), *subset* (two entities with common supertype are mentioned), and *other* (all other cases where the annotator perceived in item as being emphasized by the speaker but not falling in the previous categories). Entities that did not fall in any of the focus classes were annotated as *background*. Both transcripts and audio recordings were available to the annotators. A complete description of the annotation guidelines can be found in [9]. In a first examination of the data, [17] showed that sophisticated syntactic features as well as prosodic features were indeed correlated with focus. Our goal was to extend this preliminary work to understand what prosodic and acoustic differences exist between the focus classes and background items in natural conversational speech. Table 6 gives the distribution of three part of speech classes (nouns, adjectives and function words) in the respective focus and pitch accent classes. The table reveals that focus classes are indeed prosodically different from background items, with focus items being much more likely to bear a pitch accent. This tendency is consistent for all part of speech classes: nouns and adjectives tend to be accented even in the background case, but the rate of accenting increases when the item is focus. For function words, which typically do not bear pitch accent, the rate of accenting doubles when the words are marked as focus.

We again identified the acoustic features significantly different between pairs of classes using analysis of variance followed by Tukey's honest significant difference paired comparison tests. The different focus classes vary acoustically from background items in different ways:

background-adverbial f0_mean, f0_std, f0_sd_ratio, f0_mean_second , f0_mean_first, e_std

background-contrastive f0_std, f0_sd_ratio, f0_slope_second,

focus-pos	accent	none
adverbial-nn	32	5
contrastive-nn	276	55
other-nn	218	31
subset-nn	295	55
background	557	254
adverbial-adj	15	7
contrastive-adj	82	22
other-adj	87	16
subset-adj	89	17
background-adj	122	90
adverbial-fun	2	1
contrastive-fun	29	13
other-fun	29	8
subset-fun	27	14
background-fun	248	514

Table 6: Distribution of classes and accenting information.

	Accuracy (%)				
Features used	focus	adverbial	contrastive	other	subset
POS	72.95	67.21	71.10	68.97	78.24
acoustic	69.53	78.14	70.77	77.09	74.40
POS+acoustic	73.00	74.83	70.70	76.49	73.92

Table 7: Classification accuracy between background and focus classes. Chance level performance is 50%.

e_std

background-other f0_std, f0_sd_ratio, f0_slope_second, e_mean, e_std , e_mean_second, end_nrg_mean

background-subset f0_mean, f0_std, f0_sd_ratio ,f0_max, f0_min, f0_mean_second, e_std

Decision tree classifier accuracies for balanced classes of focus types versus background are shown in Table 7. As indicated by the individual feature analysis, the focus classes are not that homogeneous and different acoustic characteristics distinguish them form the background class. For *adverbial* and *other* focus classes which the ones mostly discussed in linguistic literature [6, 18] as associated with special prosodic realization, acoustic features perform much better than a baseline based on part of speech, with about 10% absolute improvement. This is not the case for the other focus classes and specifically for the general focus class (in which focus subtypes are merged into a single category), the POS baseline in fact performs better than the classifier based on acoustic features. These results indicate that while for all focus classes detection based on acoustic features is possible above baseline levels, the *adverbial* and *other* classes are acoustically most distinct from background elements and future studies could concentrate on only these two classes.

6. Conclusions

We have presented a study of prominence in conversational speech, as realized via pitch accents, givenness and focus. With a CRF based sequence model of pitch accent, we achieve a detection accuracy of 78.31%. This result is better in comparison with the non-sequence model (decision tree) results.

We also investigated how linguistic theories of prominence described through givenness and focus are correlated with pitch accents and presented preliminary results on their automatic detection. Our experiments suggest that there are measurable acoustic differences between *old* and *new* entities and such a distinction can be helpful in detecting prominence. Statistical significance tests with prosodic features demonstrated that collapsing *medi*ated class with *old* is more appropriate in binary classification of givenness. The prosodic features utilized in this paper also perform better than chance levels for focus classification, outperforming the part of speech baseline for given types of focus environments.

7. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proceedings of ICSLP*, 1992, pp. 867–870.

[2] D. L. Bolinger, "Accent is predictable (if you're a mind-reader)," *Language*, vol. 48, pp. 633–644, 1972.

[3] S. Pan and K. R. McKeown, "Word informativeness and automatic pitch accent modeling," in *In Proceedings of EMNLP/VLC*, 1999.

[4] W. Chafe, "Givenness, contrastiveness, definiteness, subjects, topics, and point of view," *Subject and Topic*, pp. 25–55, 1976.

[5] G. Brown, "Prosodic structure and the given/new distinction," *Prosody: Models and Measurements*, pp. 67–77, 1983.

[6] M. Rooth, "A theory of focus interpretation," *Natural Language Semantics*, vol. 1, no. 1, pp. 75–116, 1992.

[7] M. Ostendorf, I. Shafran, S. Shattuck-Hufnagel, L. Carmichael, and W. Byrne, "A prosodically labeled database of spontaneous speech," in *ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 119–121.

[8] M. Nissim, S. Dingare, J. Carletta, and M. Steedman, "An annotation scheme for information status in dialogue," in *Proceedings of LREC*, 2004.

[9] S. Calhoun, M. Nissim, M. Steedman, and J. M. Brenier, "A framework for annotating information structure in discourse," in *Pie in the Sky: Proceedings of the ACL workshop*, 2005, pp. 45–52.

[10] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, B. Beaver, and D. Jurafsky, "To memorize or to predict: Prominence labeling in conversational speech," in *Proceedings of NAACL-HLT*, 2007.

[11] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework," in *Proceedings of NAACL-HLT*, 2007.

[12] M. Gregory and Y. Altun, "Using conditional random fields to predict pitch accent in conversational speech," in *Proceedings of ACL*, 2004.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random field: Probablistic models for segmenting and labeling sequence data," in *Proceedings of ICML*, 2001, pp. 282–289.

[14] C. A. Fowler and J. Housum, "Talkers' signaling of "New" and "Old" words in speech and listeners' perception and use of the distinction," *Journal of Memory and Language*, vol. 26, pp. 489–504, 1987.

[15] J. Gundel, N. Hedberg, and R. Zacharski, "Cognitive status and the form of referring expressions in discourse," *Language*, vol. 69, pp. 274–307, 1993.

[16] E. Prince, "The zpg letter: subject, definiteness, and information status," in *Discourse description: diverse analyses of a fund raising text*, S. Thompson and W. Mann, Eds. John Benjamins, 1992, pp. 295–325.

[17] S. Calhoun, "Predicting focus through prominence structure," in *Proceeedings of Eurospeech*, 2007.

[18] D. L. Bolinger, "Contrastive accent and contrastive stress," *Language*, vol. 37, no. 1, pp. 83–96, 1961.