TEXTUAL PREDICTION OF PROSODIC PROMINENCE IN SPONTANEOUS SPEECH WITH SEQUENCE CLASSIFIERS

Shiran Dudy and Kyle Gorman

Center for Spoken Language Understanding, Oregon Health & Science University
Portland, OR, USA
dudy@ohsu.edu

ABSTRACT

Speakers produce words with differing degrees of prosodic prominence, as do naturalistic text-to-speech systems. Prominence also marks contrasts in information structure. We describe models for predicting word prominence using textual features (such as part of speech). We employ sequence classification techniques so as to encode the tendency for prominent and non-prominent words to alternate. In experiments with the spontaneous speech from the Switchboard database, we show that these models produce a significant improvement in classification above the baseline. Our results suggest that word-level prominences can be accurately inferred from relatively shallow textual features.

1. INTRODUCTION

Prominence of a syntactic unit in a sentence intends to focus the listener's attention to a new concept presented by the speaker [1]. Cole et al. [3] showed that the emphasized words in a sentence shape the meaning of the sentence in an audio-listening perceptual test of non-expert, native speakers. It was also proven that incorporating prosodic cues to an utterance, in comparison to both arbitrary emphasis of words and monotonic speech, supports language acquisition [11]. Moreover, apart from conveying messages in an intelligible form, sentences lacking prominence were observed in children with autism by Shriberg et al. [17]. In her research, their speech was described monotonic and 'machine like'. In other words, prominence contributes to a natural and expressive manner of speech.

While humans naturally apply prominence, Text to Speech (TTS) systems still aim at producing synthesized speech that incorporates prominence [5]. There are two main reasons for which prominence can leverage performance in TTS systems. First, to convey an intelligible message that clarifies the meaning of a sentence. Second, to potentially make speech more natural through imitating humans' prominence assignments. The significance of prominence is illustrated in the following examples:

That's what she$^<$ said

That's what$^<$ she said

Monotonic speech, that contains no prominence, would sound not natural and make the hearer uncertain of the purpose of the utterance. However, when 'she' is emphasized, in the first example, it is assumed that the focus is on her and the information coming next would involve additional details to support that fact. The same goes with 'what' , given in the second example, only that this time the expected details would be on what is being said.

Where can we find evidence of prominence? Prominence of a word is determined by three parameters that are expressed in sound: a local maxima or minima of the fundamental frequency contour, longer duration of the emphasized unit, and a high energy region of speech [15]. More interestingly, in this work we hypothesize that linguistic features contribute to prominence. For example, since content words tend to introduce new information, they are likely to be accented more often than function words. Another example is the presence of contrast in a sentence. It is assumed that presenting a clause that undermines a former idea is expected to be emphasized in order to point out the contrasting information. According to Kimball and Cole [8] non experts native speakers are not capable of perceiving clashes resulted by adjacent prominent words and adjust their hearing to perceive a modified sentence that detaches and avoids adjacency. This finding might also suggest that sentence production involves scattering prominence to non-prominent environments by avoiding accenting two or more units in a row.

Therefore, our goal in this research is to develop a model that better predicts the prominence in a sentence by examining several different sequence classifiers that use contextual information. We use linguistic features of annotated spontaneous speech, which is known to be less predictive than read speech. We show that our model is significantly better than the baseline.

In the following sections we describe prior work in Section 2, methods to build our classifiers in Section 3, performance evaluation in Section 4, and conclusions in Section 5.

2. PRIOR WORK

The prior work in the field of predicting prominence in a sentence is divided to two schools of thought: the research that involves acoustics with or without linguistic features and the research that is focused only on linguistic features. We will review both and point out some interesting methods that were developed.

Rosenberg and Hirschberg [16] explored which domain of acoustic analysis predicts most accurately pitch accents and concluded that the word level reaches the highest accuracy of 84.2% on read speech. In another study [13] applying Maxi-

mum Entropy approach on read speech, using Boston University Radio News Corpus, scored 86.0% accuracy. Mehrabani et al. [10] used unsupervised learning to find several clusters to represent prominence level found in speech and then modeled the different clusters' relationship with linguistic features to predict prominence from a text. A German-language study that incorporated knowledge only from acoustics, used features like nucleus duration, spectral emphasis, pitch movements, and intensity to show correlation between perceptual tests to their predicted prominence [18].

In linguistic studies the researchers' focus is on utilizing the syntactic information. Windmann et al. [21] developed a rule-based voting system based on features such as part of speech tags, and dictionary based stress assignments, to determine the prominence assignment. Another study [2] suggested that using features such as the frequency of a word, the type of a noun, and accent ratio eventually results in a 77% in accuracy on spontaneous speech. Using features found in Switchboard spontaneous corpus such as contrast led to 76.58% in accuracy [12].

We would like to offer a better linguistic-based solution to predict prominence on spontaneous speech. Next, we'll present our methods and set of features we chose to use.

3. METHODS

3.1. Data

We used 'Switchboard in NXT' database [7] on annotated spontaneous speech. We used 40,647 tokens that constructed 6,425 sentences (or sequences). The database offers various features added from Penn Treebank and MS-State Transcript projects. The features we used were:

- **Terminals** containing the orthographic transcription along with its corresponding part of speech.
- **DialAct** containing the dialogue act description such as a statement and a question.
- **Kontrast** containing the level of kontrast such as word, phrase and whether its contrastive.
- **Disfluency** containing the places that the speaker hesitates. We used only the repair units.
- **Phonewords** containing similar information like terminals (both eventually were merged).
- **Phrases** containing groups of words that were assigned with a different prosody such as minor and major.
- **Accent** containing the information on level of prominence assigned to each token: 'full', 'weak', and 'none'.

3.2. Features

We extracted information from Switchboard and added additional features that may further tell about the accent which is the target prediction in our experiment.

In Figure 1, the blue represents the feature set referred to as X. Our feature set was: the word, its part of speech, its collapsed part of speech, whether it is a function and a negation word, the token's syllable's vowel-group by order, the token's nucleus, nucleus kind, kontrast level, kontrast type, whether it is a phrase, phrase kind, and the description of the discourse. In pink, there is the true target y describing the accent. In this case, only 'pretty' is assigned with 'full' type of accent and all the rest are assigned with 'none' accent.

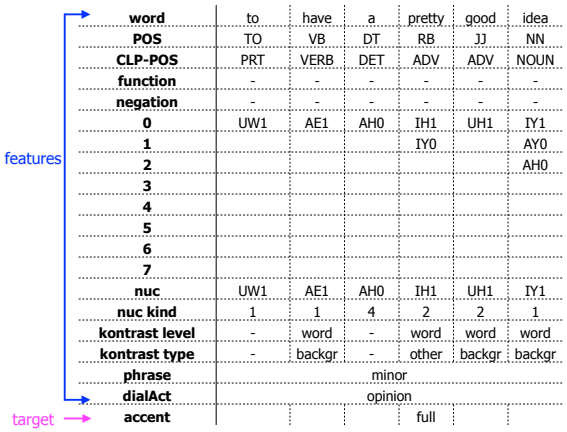

Fig. 1: sequence of feature set and its corresponding target accents

3.3. Models

We used sequence classifying models to exploit the internal relationships found in adjacent observations and consider the problem in its context. Models 1-3 were based on Average Perceptron approach [4] applying Hidden Markov Models assumptions to these models. For models 4, 6 we used a 'PocketCRF' toolkit' (https://github.com/kylebgorman?tab=repositories).

1. Local Search – we employed three models: zero order, first order, and second order. Zero orders's input contained feature set in time t: $X(t)$. First order's was $X(t)$ and $X(t-1)$. Second order's was $X(t)$, $X(t-1)$, and $X(t+1)$.
2. Greedy Search – we employed an input that contained feature set of time t together with last prediction for time $t-1$. The input was $X(t)$, $\hat{y}(t-1)$ and transition features.
3. Viterbi Search – we employed First order Viterbi algorithm [19].
4. Conditional Random Fields (CRF) Search – we employed an input that contained $X(t)$, $X(t-1)$, and $X(t+1)$[9].
5. Max Margin Markov Networks (M3N) Search – we employed an input that contained $X(t)$, $X(t-1)$, and $X(t+1)$ [14].
6. Baseline – our baseline was set by guessing the most probable target found in the database. In our experiment the guess was 'none' accent type.

4. RESULTS

We performed a 10 fold cross-validation test and averaged all accuracy results for token and sentence level tests.

4.1. Token Level Accuracy

The models were tested for their token accuracy. Accuracy was measured by counting the correct classifications over all classifications. Table 1 describes the accuracy of the classifications using the models. We noticed that all Local Searches performed the best, then, the Viterbi and the Greedy which were close to each other and lasts are the CRF as well as the M3N, yet still seemed to be different than the baseline.

Model	Accuracy[%]
Local Search 0 order	79.1
Local Search 1st order	79.2
Local Search 2nd order	79.0
Greedy Search	77.2
Viterbi Search	77.5
CRF	74.7
M3N	72.3
Baseline	69.9

Table 1: Classification Accuracy

All models performed better than the basline but we needed to understand how significant results were. Next, we applied Wilson Confidence Interval [20] to have a better idea of whether the models are different or that their confidence interval overlap and they were driven from the same population. Looking at Figure 2, we noticed that all models seem to differ than the basline since none overlap with it. We also saw that Local Searches demostrates overlap within group but does not overlap with Greedy and Viterbi. Greedy and Viterbi overlap but do not overlap with CRF. For the same reason, CRF and M3N do not seem to represent a similar population as well.

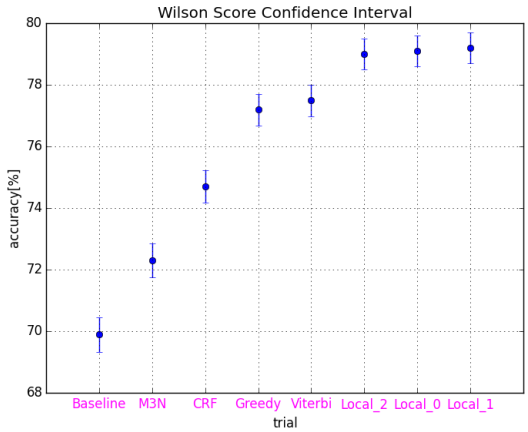

Fig. 2: Wilson Score

Finally, we performed a Mcneamar [6] test to determine the significance of our results. Mcneamar test compares two models in terms of how many true positive and true negtive were predicted in each model – this determines the differences in model populations.

Model A	Model B	P value
M3N	Baseline	1.82e-22
CRF	M3N	4.74e-197
Greedy Search	CRF	9.04e-137
Viterbi Search	Greedy Search	0.02
Local Search 2nd order	Viterbi Search	2.2e-38
Local Search 0 order	Local Search 2nd order	0.05
Local Search 1st order	Local Search 0 order	0.18

Table 2: McNeamar Significance Test

In Table 2, we sorted the groups in increasing accuracy order and showed a pairwise comparison between two adjacent models. P-values that were below 0.01 were demonstrated two significantly different models. We can see that M3N, CRF, Viterbi and Greedy together, and all Local Searches form four significantly different populations than the baseline. The Best model is the Local Search.

Our initial intuition of applying Averaged Perceptron approach was that it will perform better since during training the model is penalized proportionaly when the prediction is not like the true target. This regularization proved to be helpful. However, we expected that Viterbi and Greedy search would outperform Local Search since the former incorporate knowledge of the past prediction $\hat{y}(t-1)$ and it thought to be an important feature for deciding whether the current feature should be prominent. The possible explanation for that might derive from the transition features that were added to the input in Viterbi and Greedy and damaged the learning.

4.2. Sentence Level Accuracy

We know that it takes one misclassification of prominence to understand the sentence entirely differently as seen in the examples given in the Introduction Section. Therefore, we adopted a more rigid approach that measures how many fully correctly classified sentences the models predicted. This could help understand how many times we are confident that the message was conveyed the way the speaker intended. Table 3 describes the accuracy results for this experiment:

Model	Accuracy[%]
Local Search 0 order	48.4
Local Search 1st order	48.9
Local Search 2nd order	48.9
Greedy Search	48.9
Viterbi Search	48.9
CRF	34.8
M3N	32.8
Baseline	42.0

Table 3: Classification Accuracy

To finalize these results we used a Binomial Test to measure significance.We compared accuracies of baseline with all models that were above baseline. We conducted a Binomial Test that compares between two models' accuracy result and found that the Local Search 0 order is significantly different than the baseline with p-value of 5.19e-75. The p-value for

Viterbi, Greedy, Local Search 1 and 2 with the baseline is 2.22e-84. Our models, therefore, were able to send a more correct messages than using baseline predictions.

Table 3 shows that the Local, Greedy and Viterbi Search share similar pattern learning that is expressed in sentence accuracy. We can also infer that though the Local Search has a better accuracy result on a token level it is not expressed in terms of sentence accuracy.

5. CONCLUSION

In this research we showed that employing linguistic features extracted from spontaneous speech to create a prominence predicting model were significantly different than baseline guessing. This knowledge can be incorporated to a TTS system to improve intelligibly and to make it more natural.

6. REFERENCES

[1] Dwight Bolinger. *Intonation and its parts: Melody in spoken English*. Stanford University Press, 1986.

[2] Jason M Brenier, Ani Nenkova, Anubha Kothari, Laura Whitton, David Beaver, and Dan Jurafsky. The (non) utility of linguistic features for predicting prominence in spontaneous speech. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 54–57. IEEE, 2006.

[3] Jennifer Cole, Timothy Mahrt, and José I Hualde. Listening for sound, listening for meaning: Task effects on prosodic transcription. In *Proceedings of Speech Prosody*, 2014.

[4] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics, 2002.

[5] Martin Cooke, Catherine Mayo, Cassia Valentini-Botinhao, Yannis Stylianou, Bastian Sauert, and Yan Tang. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585, 2013.

[6] Morten W Fagerland, Stian Lydersen, and Petter Laake. The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC medical research methodology*, 13(1):91, 2013.

[7] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE, 1992.

[8] Amelia E Kimball and Jennifer Cole. Avoidance of stress clash in perception of conversational american english.

[9] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[10] Mahnoosh Mehrabani, Taniya Mishra, and Alistair Conkie. Unsupervised prominence prediction for speech synthesis. *Power*, 2(1.6):1–3, 2013.

[11] James L Morgan, Richard P Meier, and Elissa L Newport. Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive psychology*, 19(4):498–550, 1987.

[12] Ani Nenkova, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and Dan Jurafsky. To memorize or to predict: Prominence labeling in conversational speech. 2007.

[13] Vivek Rangarajan, Shrikanth Narayanan, and Srinivas Bangalore. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. In *Proceedings of NAACL HLT*, pages 1–8, 2007.

[14] Ben Taskar Carlos Guestrin Daphne Roller. Max-margin markov networks. *Advances in neural information processing systems*, 16:25, 2004.

[15] Andrew Rosenberg and Julia Hirschberg. On the correlation between energy and pitch accent in read english speech. In *INTERSPEECH*, 2006.

[16] Andrew Rosenberg and Julia Hirschberg. Detecting pitch accents at the word, syllable and vowel level. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 81–84. Association for Computational Linguistics, 2009.

[17] Lawrence D Shriberg, Rhea Paul, Jane L McSweeny, Ami Klin, Donald J Cohen, and Fred R Volkmar. Speech and prosody characteristics of adolescents and adults with high-functioning autism and asperger syndrome. *Journal of Speech, Language, and Hearing Research*, 44 (5):1097–1115, 2001.

[18] Fabio Tamburini and Petra Wagner. On automatic prominence detection for german. *Proceedings of Interspeech 2007*, 2007.

[19] Andrew J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.

[20] Sean Wallis. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, 20(3):178–208, 2013.

[21] Andreas Windmann, Igor Jauk, Fabio Tamburini, and Petra Wagner. Prominence-based prosody prediction for unit selection speech synthesis. *Proceedings of Interspeech 2011*, 2011.