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

Prominent words in a sentence focusses 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. Furthermore, meaningful prosodic cues supports child language acquisition [11]. Monotonic speech is common in children with autism spectrum disorders Shriberg et al. [18]and has been described as 'machine-like'. All this suggests that appropriate use of word prominence is an important component of natural and expressive speech.

While humans naturally apply prominence, Text to Speech (TTS) systems aim to produce speech with appropriate prominences, for at least two reasons. [5]. First, prominence contributes to the broader meaning of an utterance, as the following example shows:

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. Secondly, appropriately placed prominences may serve to make synthesized speech more natural.

Where can we find evidence of prominence? At the word level, prominence is realized by at least three acoustic features, all defined relative to the larger context: a local maxima or minima of the fundamental frequency contour, increased duration of the word, and a higher overall amplitude [16]. 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.One key linguistic intuition in the model we propose is the idea that speakers tend to avoid "clashes"—sequences of adjacent prominent words—and long "lapses"–sequences of -adjacent non-prominent words. Kimball and Cole [8] find that native speakers have difficulty perceiving "clashes". We thus employ sequence classification models to model these local dependencies. 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.

Our goal is to evaluate textual models of word prominence for possible applications in text-to-speech systems. Unlike prior work, which uses read speech, we employ spontaneous speech data, which is likely to be more difficult. Our final model is far from perfect, but a significant improvement over a baseline model.

Section 2 describes prior work on predicting word prominence. Section 3 describes data and features used in our experiments. The results of our spontaneous speech experimental results are provided in Section 4. Section 5 concludes.

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 [17]pitch accent prediction in a variety of domains and found that word level lead to the highest accuracy predictions; their model achieved 84.2% prediction accuracy on read speech. Pitch accent is the observed change in pitch that is refereed to as prominence. Rangarajan et al. [14] applied maximum entropy classifiers and a mixture

of textual and acoustic features for prosody labeling and used Boston University Radio News Corpus that scored 86.0% accuracy. Mehrabani et al. [10] employed clustering in a study of word prominence. A study of German incorporated only acoustic features such as nucleus duration, spectral emphasis, pitch movements, and intensity to developed a model of prominence which was strongly correlated with human perceptual judgements [19].

In linguistic studies the researchers' focus is on utilizing the syntactic and semantic information. Windmann et al. [22] 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].

In what follows, we employ additional further linguistic features to improve prominence prediction for spontaneous speech. We proceed to describe the data, features, and machine learning techniques used in our experiments.

3. METHODS

In all our experiments, we used the Switchboard corpus of American English spontaneous speech[7] and the accompanying NXT annotation set, which consist of linguistic annotations added by other teams after the original Switchboard release. The target dataset consists of 40,647 word tokens from 6,425 utterances.

3.1. Features

We employed the following sets of features:

- Terminals: the orthographic wordform
- POS: the Penn Treebank part of speech tag
- CLP-POS: the POS tag mapped onto the Petrov et al. [13] universal part of speech tagset
- Function: is the terminal a function word? (in, on, of)
- Negation: is the terminal a negation word? (no, not)
- Vowels: indicators for each stressed vowel label, assuming left-aligned syllables
- Nucleus: the vowel label of the nucleus of the primary stressed syllable of the terminal
- Nucleus type: is the type of nucleus that receives _primary stress_
- Kontrast: kontrast level and kontrast type as defined in the the database
- Phrase type: grouping of words in the MS-State transcript into prosodic phrases
- DialAct: the dialogue act description (e.g., "question", "statement")

Finally, we used as the outcome variable the three prominence levels: 'full', 'weak', and 'none'.

Example feature and outcome vectors are provided in Figure 1. In the example, only the word 'pretty' was coded as having prominence; all other words are 'none'.

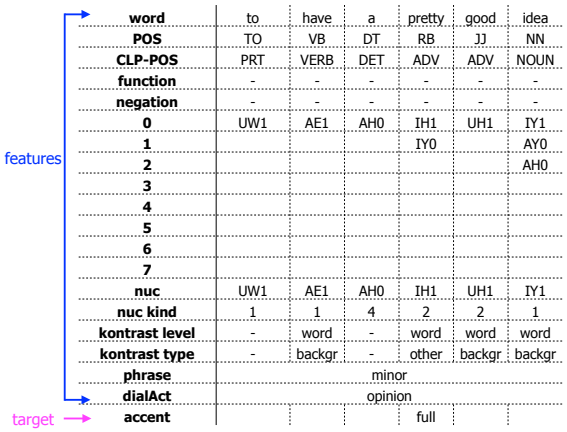

Fig. 1: sequence of feature set and its corresponding target accents

3.2. Models

We employed sequence classification models to predict prominence sequences. The baseline model simply guesses the most probable target over the dataset, which is 'none'. The next three models are variants of a hidden Markov model backed by a linear model (the averaged perceptron [4]). The final two models use the PocketCRF toolkit.[1]

1. Local Search (L): : the "zero order" model – L_0– mode employs only textual features of the current observation (X_t). The "first order" model – L_1– employs features of the current and preceding observation (X_t, X_{t-1}). The "second order" model – L_2– employs features of the current observation as well as the preceding and following observation (X_{t-1}, X_t, X_{t+1}).
2. Greedy search (G): The features consist of those extracted from the current observation as well as transition features generated from the best hypothesis for the preceding prominence label (X_t, $\hat{y}_{t-1}$).
3. Viterbi search (V): The same as the greedy model, but employing the Viterbi algorithm to consider all possible preceding labels $\hat{y}_{t-1}$.[20].
4. Conditional random fields (CRF): Global inference using a probabilistic classifier and the above "second order" features.[9].
5. Max margin Markov networks (M3N): Global inference using a maximum-margin sequence classifier and the above "second order" features.[15].
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.

[1] `http://pocket-crf-1.sourceforge.net/`

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 L 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[%]
Baseline	69.9
L_0	79.1
L_1	79.2
L_2	79.0
G	77.2
V	77.5
CRF	74.7
M3N	72.3

Table 1: Classification Accuracy

All models performed better than the baseline but we needed to understand how significant results were. Next, we applied 95% binomial confidence intervals based on the Wilson score [21] 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 baseline since none overlap with it. We also saw that L demonstrates 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.

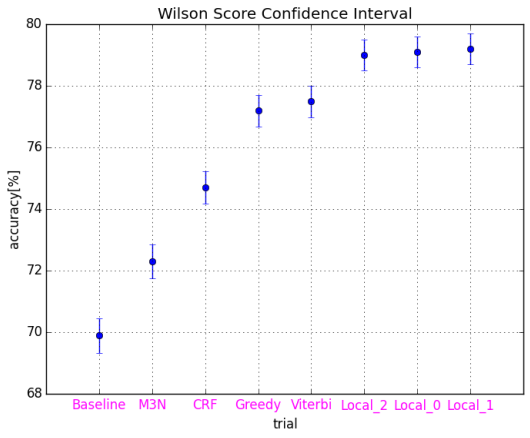

Fig. 2: Wilson Score

Since the baseline confidence intervals do not overlap those of any of the non-baseline models, we have reason to believe that these models are significantly better than the baseline. We formalized model comparisons by performing pairwise McNemar tests [6] . Table 2 shows the results of these comparisons.

Model A	Model B	p-value
M3N	Baseline	< 0.001
CRF	M3N	< 0.001
G	CRF	< 0.001
V	G	0.020
L_2	V	< 0.001
L_0	L_2	0.050
L_1	L_0	0.180

Table 2: McNemar 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 0.001 demonstrated two significantly different models. We can see that M3N, CRF, Viterbi and Greedy together, and all L form four significantly different populations than the baseline. The best model is the L.

We initially expected that the Greedy and Viterbi models would outperform local classification, as the expectations about the prior labels would prove to be predictive of the current label. However, these features appear to actually result in slightly less accurate models.

4.2. Sentence Level Accuracy

As pointed out earlier, even a single erroneous assignment of prominence may greatly affect the meaning of a sentence, and may also lead human listeners to judge the sentence as unnatural. Therefore, we adopted a more stringent measure of accuracy that required entire sentences to be correctly labeled. This is the one case where we can be confident that the prominence system is conveying a naturalistic message. Table 3 provides sentence-level accuracy results.

Model	Accuracy[%]
Baseline	42.0
L 0th	48.4
L 1st	48.9
L 2nd	48.9
G	48.9
V	48.9
CRF	34.8
M3N	32.8

Table 3: Classification Accuracy

In this case, the baseline is defined as by counting how many fully correct sentences are predicted by using a a model the guesses the common label in the corpus. According to the Binomial test, the local classifiers, greedy search, and Viterbi search are all significant improvements on the baseline. It is interesting to note that while the local classifiers were superior at the token level, their performance at the sentence level is matched by the greedy and Viterbi search.

5. CONCLUSION

We showed that simple textual features extracted from spontaneous speech can be used to predict human-like word promi-

nence sequences, and that these features, when exploited by modern sequence classification models, produce a prominence placement model that overperforms the baseline. This knowledge may thus result in better—more intelligible and more natural—text-to-speech.

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] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.

[14] 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.

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

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

[17] 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.

[18] 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.

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

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

[21] 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.

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