A Logistic Regression Model for Detecting Prominences*

Arman Maghbouleh

ATR Interpreting Telecommunications Research Labs
2-2 Hikaridai, Seika-cho,
Soraku-gun, Kyoto 619-02, Japan
and
Department of Linguistics
Stanford University
Stanford, CA 94305, U.S.A.

ABSTRACT

This paper describes the development of a model for identifying points of prominence in speech. This model can be used as a first step in intonational labeling of corpora that are used in some speech synthesis systems (Black and Taylor, 1995). The working definition of prominence is that starred ToBI accents (Silverman et al., 1992), that is, H*, L*, L*+H, L+H*, and H+!H*, are prominent. The prominence detection model developed here is based on the sums-of-products vowel duration model (van Santen, 1992).

The model was trained and tested on different portions of the Boston University Radio News corpus and achieves accuracy results of 86.3% correct identification with 12.5% false detection. The results are comparable to those of previous work (Wightman and Campbell, 1995): 85.9% correct identification with 10.7% false detection. The advantage of this model is that it can be trained quickly on as few as 600 data points, reducing the need for large corpora.

1. INTRODUCTION

Human speech is not monotonous. When people speak, some syllables, and in turn words that contain them, are perceived to be more prominent than others. Linguists refer to these prominent syllables as *accented* (Bolinger, 1986). Accented syllables tend to be longer, host local maxima or minima of pitch, and have a different voice quality than their non-accented counterparts (de Jong, 1995).

Points of prominence hold the key to identifying what a speaker believes to be new or given information in a conversation (Bolinger, 1989). Therefore, any program that attempts to converse with a person can benefit from recognizing prominences as well as words.

Furthermore, in speech synthesis systems, prominence detection can be used to achieve more natural speech. As mentioned before, accented segments tend to have a different voice quality from unaccented ones. Therefore, in a concatenative synthesis system (such as, Black and Taylor, 1995), one would do better by choosing an accented unit for concatenation when synthesizing a prominent syllable. An accent identification procedure, such as the one described here, can be used to identify accented units in the unit database.

Since duration is one of the primary acoustic correlates of accent, the guiding principle in developing this model has been to take advantage of existing research on duration models which include accent as a factor. Unfortunately, it is not possible to simply assign accent to longer syllables because duration is crucially affected by other factors as well. For example, given the same emphasis, the two words *bead* and *bat* will have very different durations because of their differing segmental makeup. The /i/ in *bead* can be shown to be inherently longer than the /ae/ in *bat* . That is, the /i/ is pronounced with greater duration than the /ae/ in the same context with same perceived emphasis. Furthermore, vowels before voiced stops such as /d/ are pronounced with greater length than vowels before unvoiced stops such as /t/. Therefore, the greater length in *bead* may be due to segmental factors and not a reflection of greater emphasis. In order to find prominences based on duration values, one needs to know precisely how the various factors interact.

A partial list of factors enumerated by various studies of durational effects includes inherent phone duration, features of following and preceding segment, location of segment within the syllable, location of the syllable within the word, speech tempo, location of the word with respect to the phrase boundary, and lexical stress (Campbell, 1992; Crystal, 1988; Klatt, 1976; van Santen, 1994). In principle, given a duration model, it should be possible to isolate the contribution of prominence to an observed duration and hence identify the presence or lack of accent.

This study used the sums-of-products (SoP) vowel duration model (van Santen, 1992, 1994) as a starting point because of reported high accuracy results and the low training requirements (Maghbouleh, 1996).

The remainder of this report describes the SoP model, how it was modified into a prominence detection model, the corpus used for training the new model, and the ensuing results.

2. SUMS-OF-PRODUCTS DURATION MODEL

The sums-of-products duration model rests on van Santen's observation that additions and multiplications are enough to

* I'm most indebted to Andrew Hunt who originally suggested modeling prominences directly and provided the code for reading the power data. I am also thankful to Nick Campbell, Alan Black, and Norio Higuchi at ATR, and Jan van Santen of Bell Laboratories for their help.

characterize the interaction of factors for large groups of phonemes. For example, he found that the sums and products in equation 1 sufficiently describe durational variation for vowels.

$$\begin{aligned}\log(\texttt{VowelDur}(\ldots)) = {} & \delta(\texttt{Vowel}) + \varepsilon(\texttt{Pre}) + \pi(\texttt{Post}) + \\ & \alpha(\texttt{Acc}) \times \beta(\texttt{Stress}) + \\ & \omega(\texttt{WordPre}) + \eta(\texttt{WordPost}) + \\ & \theta(\texttt{Post}) \times \kappa(\texttt{UttBoundary})\end{aligned}$$

Equation 1: Vowel duration model formulated by van Santen (1992)

The factors in equation 1 are: vowel identity (`Vowel`), identity of the preceding phoneme (`Pre`), identity of the following phoneme (`Post`), presence or lack of accent (`Acc`), presence or lack of lexical stress (`Stress`), position of the vowel with respect to the left edge of its enclosing word (`WordPre`), position of the vowel with respect to the right edge of the word (`WordPost`), and distance to the following phrase boundary (`UttBoundary`).

A term like $\delta(\texttt{Vowel})$ is a function that given the vowel identity returns a number representing that vowel's inherent duration. Similarly, the other terms are functions that represent the contribution of the other factors. If the amount of shortening or lengthening attributable to a given factor stayed constant and there were no interactions among the terms, then a simple linear model would perform as well as the sums of products.

Equation 1 only applies to vowels. Different factors and equations are needed for consonants. According to van Santen (1994), another 41 different sums-of-products equations are needed to model the consonants.

3. PROMINENCE DETECTION MODEL

Since equation 1 describes duration in terms of accent and other factors, it is possible to describe accent in terms of duration and other factors as shown in equation 2.

$$\begin{aligned}\alpha(\texttt{Acc}) \times \beta(\texttt{Stress}) = {} & \log(\texttt{VowelDur}) + \delta(\texttt{Vowel}) + \\ & \varepsilon(\texttt{Pre}) + \theta(\texttt{Post}) + \\ & \omega(\texttt{WordPre}) + \eta(\texttt{WordPost}) + \texttt{Energy}\end{aligned}$$

Equation 2: SoP model reformulated to have accent on the left hand side.

The utterance boundary term is missing from equation 2 because in most contexts where accent is unknown, the location of boundaries is unknown too. A new term, `Energy`, was added because it is one of the major correlates of prominence too. One way to properly include effects of energy, is to go through a process similar to what was done for duration: 1) identify relevant factors and interactions that affect energy. 2) isolate prominence from the other factors. 3) combine the information from duration and energy and the confounding factors for each in order to isolate prominence. As a first approximation, this study was based on equation 2 which assumes that the same factors affect energy and duration.

In order to use the tools available for working with linear equations, equation 2 was broken down to two separate linear equations, one for each level of stress. Furthermore, since accent is a categorical variable (accented vs. not), the regression is performed on a related continuous variable: the probability that a given set of factors will result in the perception of prominence. Finally, since the probabilities will be limited to the [0,1] range, the logistic function is used to expand that range to the real line. One of the two resulting equations is shown in equation 3.

$$\begin{aligned}\log\left(\frac{p(\texttt{acc})}{1 - p(\texttt{acc})}\right) = {} & \log(\texttt{VowelDur}) + \delta(\texttt{Vowel}) + \\ & \varepsilon(\texttt{Pre}) + \theta(\texttt{Post}) + \\ & \omega(\texttt{WordPre}) + \eta(\texttt{WordPost}) + \texttt{Energy}\end{aligned}$$

Equation 3: Equation for the two logistic regressions used in identifying accents.

The overview of parameter estimation and testing of the model is as follows. The available data, described in the next section, was split into two parts: one for training and one for testing. The training data was then further split in two parts: one for vowels in stressed syllables and one for the other vowels. Two sets of parameters to equation 3 were then estimated, one for each of the parts of the training data. Finally, the data points in the test data were fed through the resulting fully specified equations and if the resulting number was higher than a certain value, they were labeled prominent. The percentage of actually prominent vowels that were labeled so is reported as *hit rate*, and the percentage of actually non-prominent vowels which were labeled prominent is reported as *false alarm rate*.

4. CORPORA

This study used the Boston University Radio News Corpus which consists of recordings of broadcast radio news stories. Thirty-two of the stories (41,718 phones) that were read by a professional female announcer were used here. The corpus was collected at Boston University and is more fully described in (Ostendorf et. al., 1995). The recordings were made in the studio during broadcast. They were later digitized, transcribed, segmented, and hand-labeled with ToBI prosodic labels (Silverman et al., 1992).

The ToBI labels mark tones and break indices. The tones consist of pitch accents, phrase accents, and boundary tones. Pitch accents, the focus of this study, characterize prominent syllables with five labels for American English:

H*, L*, L*+H, L+H*, H+!H*. Each pitch accent type is related to a particular sort of pitch movement on or near the accented syllable. Phrase accents (L-, H-) and boundary tones (L%, H%), which characterize pitch movements at the edge of phrase boundaries, and break indices were ignored.

Reported inter-labeler agreement for presence versus absence of accent on a variety of texts is 86-88% (Silverman et al., , 1992). The agreement figures go up to 94% for fluent radio-style utterances (Wightman and Ostendorf, 1995).

5. RESULTS & DISCUSSION

Table 1 shows the results of prominence detection with training on one part of the f2b corpus and testing on another part for this study and some previously published studies with the same corpus.

	Hit Rate	False Alarm
Wightman & Ostendorf '94	0.836	0.127
Wightman & Campbell '95	0.859	0.107
Present work	0.863	0.125

Table 7: Results of prominence detection compared to other work.

The present model achieves accent identification results similar to the best previous work. The 86% accuracy results may seem impressively close to human performance but since 69% of the vowels are not accented, one can achieve 69% accuracy by simply always predicting no accent. With 69% as the baseline and 94% for human performance, we are only 68%, that is, (0.86-0.69)/(0.94-0.69), of the way through to achieving the best possible performance.

The advantage of this model compared to previous work is that it is computationally simpler to train. Here, training involves only regression instead of the more complicated hidden markov model grafted to a tree model used in Wightman and Campbell 1995. Furthermore, the model can be trained on as few as 600 data points because it has only 27 degrees of freedom. The model is lean because it takes advantage of existing research on duration to take into account only factors that have already been shown to be relevant.

Future work will try to improve performance results by properly including pitch and energy effects and extend work to identifying boundaries and accent types.

6. REFERENCES

. Black, A. and P. Taylor (1994). CHATR: a generic speech synthesis system. COLING-94, Kyoto, Japan.

. Bolinger, D. (1986) *Intonation and Its Parts: Melody in Spoken English* (Stanford, California, Stanford University Press).

. Bolinger, D. (1989) *Intonation and Its Uses: Melody in Grammar and Discourse* (Stanford, California, Stanford University Press).

. Campbell, N. (1992). Multi-level Timing in Speech. Ph.D. thesis. Sussex University.

. Crystal, T. H. and A. S. House (1988). "Segmental durations in connected speech signals: Syllabic stress." Journal of the Acoustical Society of America **83**: 1574-1585.

. Crystal, T. H. and A. S. House (1988). "Segmental durations in connected-speech signals: Current results." Journal of the Acoustical Society of America **83**: 1553-1573.

. de Jong, K. (1995). "The Supraglottal Articulation of Prominence in English: Linguistic Stress as Localized Hyperarticulation." Journal of Acoustical Society of America **97**(1): 491-504.

. Klatt, D. H. (1976). "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence." Journal of Acoustical Society of America **59**: 1208-1221.

. Maghbouleh, A. (1996). "An Empirical Comparison of Automatic Decision Tree and Hand-Configured Linear Models for Vowel Durations ." Computational Phonology in Speech Technology Workshop. University of California Santa Cruz.

. Ostendorf, M., P. J. Price, S. Shattuck-Hufnagel. (1995). The Boston University Radio News Corpus. Boston University Electrical Engineering.

. Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, (1992). ToBI: A Standard for Labeling English Prosody. Intl. Conference. on Spoken Language Processing,

. Van Santen, J. P. H. (1992). "Contextual Effects on Vowel Duration." Speech Communication 11: 513-546.

. Van Santen, J. P. H. (1994). "Assignment of segmental duration in text-to-speech synthesis." Computer Speech Language.

. Wightman, C. W. and M. Ostendorf (1994). "Automatic labeling of prosodic patterns." IEEE Trans. on Speech and Audio Processing.

. Wightman, C. W. and W. N. Campbell (1995). "Improved Labeling of Prosodic Structure." IEEE Trans. on Speech and Audio Processing.