PREDICTING GRADIENT F0 VARIATION: PITCH RANGE AND ACCENT PROMINENCE

Ivan Bulyko and Mari Ostendorf

Electrical and Computer Engineering Department
Boston University, 8 Saint Mary's Street, Boston, MA 02215, USA

ABSTRACT

Many aspects of prosody prediction in speech synthesis could be improved, from placement of symbolic accent and phrase boundary markers to control of continuously varying parameters (e.g., duration, fundamental frequency). The goal of this work is to develop algorithms for predicting aspects of fundamental frequency typically said to have gradient variation: pitch range and prominence. In addition, the results of the automatic training methodology are used to investigate differences in prominence patterns associated with different genres of speech.

1. INTRODUCTION

Intonational patterns are characterized by sequences of pitch accents and boundary tones, which determine the shape of the F0 contour. Different syllables in an utterance can have different discrete levels of perceived prominence because of absence vs. presence of a pitch accent and/or word-level stress, but most theories also acknowledge that there is also gradient variation of prominence. Experiments on controlled stimuli show that perceived prominence is highly correlated with F0 peak height [12], and most studies focus on this cue.

Analyses of spontaneous speech [13] show that perceived relative prominence can vary for multiple accents within a prosodic phrase in ways that are not consistent across phrases with the same number of accents. However, the notion of completely free gradient variation of prominence is not very attractive from a theoretical perspective [6] or from the practical needs of speech synthesis prediction. One approach is to better understanding relative prominence is to look for constraints on the variation. An alternative, explored here, is to investigate prediction algorithms and determine how much of the variation can be explained by the local tone sequence vs. higher level factors. More specifically, this work looks at methods for predicting prominence (in the sense of F0 peak height for high accents) using corpus-based learning techniques.

As there is evidence in the literature that discourse structure affects pitch range (see [3] for a review), we decided to separate range from intra-phrase prominence by using a simple multiplicative model: $F_a = F_p K_{pr}$, where F_a is the target peak for a high pitch accent in a phrase with range F_p and prominence scaling coefficient K_{pr}. Note that the pitch range is constant over the intermediate prosodic phrase, and the prominence factor accounts for range compression and/or lowering. (Speaker-dependent overall range and baseline are omitted here because of a focus on speaker-dependent data.) Separating these two phenomena in designing predictors also allows us to investigate whether it is the case that different factors are important for predicting the two coefficients $F0_p$ and K_{pr}. For example, it may be that information status (e.g. new vs. given) is important for range but not for relative prominence within a phrase.

Within a prosodic phrase, a frequently observed phenomenon is that of progressive lowering of F0 peak height for successive accents, often referred to as downstep. Downstep cannot simply be accounted for by pitch range compression over the course of the phrase, since an accent can also be followed by another accent of the same or greater peak height. The sequence of accent tone labels is important for explaining peak variation, but there must be other factors since it does not predict raised peaks in non-initial pitch accents, which we and others have observed. This study aims to find such other factors.

In addition to investigating the importance of different factors for predicting inter vs. intra-phrasal peak variation, this work also considers the questions of which mathematical model works best and how these models translate across genres of speech. In particular, we assess different assumptions about dependence between prediction variables. In addition, we investigate whether different linguistic features are important for different genres. In other words, can we maintain the structure of a predictor across genres and simply retrain the key coefficients?

2. PREDICTION VARIABLES

We explored three main categories of prediction variables in this study, including tone labels and accent position in the phrase, syntactic structure, and features related to discourse structure, as described next.

Tone labels and relative phrasal position of accents are standard in rule-based systems of predicting peak height. The tone inventory used here, for both current accent and previous accent, is simply +/- downstep[1] Other accent-related features included the number of accents in the phrase and the number of downstepped accents so far. Both the tone labels and prosodic phrase structure are based on hand-marked labels.

The class of syntactic features include the part-of-speech label and syntactic constituent of the target word, both of which are annotated using automatic algorithms. Part-of-speech labels have been used both as a predictor of prominence and accent location. Syntactic parse features were investigated because of their usefulness in accent prediction and because of work suggesting an interaction between given/new status and subject/object position [4].

Discourse segmentation is well known to be correlated with pitch range. Here, we used a very simple representation based on sentence and paragraph structure, i.e. position (first, middle, last) of the intermediate prosodic phrase in a major prosodic phrase, in a sentence, in a paragraph and in a story. Finally, we included two features related to information status that attempted to capture "newness", including whether a content word was new to the paragraph and whether a word was part of a "named entity" (person, organization, location, date, money – hand-marked). The named entities are important because they typically specify the "who, what, where, when" of a story.

3. APPROACH

3.1. F0 Peak and Prominence Measurements

F0 values, used for computing accent and phrase peaks, were measured in Hz using the Entropic Waves pitch tracker with an F0 sample rate of 100Hz. A median filter (window size = 3) was applied to the F0 contour in order to remove spurious erroneous values and to a lesser degree compensate for segmental perturbations. Pitch doubling/halving errors were accounted for to some extent by ignoring regions with discontinuous jumps of high deviation within each syllable and large regions that were explicitly hand marked as having pitch errors. The regions with a high final rise were also excluded from finding phrase and accent peaks.

The pitch range (F_p) for a prosodic phrase was taken as the maximum F0 value for the phrase. Since it was difficult to obtain a reliable estimate of the local baseline, only the peak was used to characterize range. Then, for each syllable accented with a high tone, the highest F0 point was found (F_a). In some cases the peak F0 was not necessarily located within the syllable boundaries, therefore we expanded the interval for finding the peak into the adjacent syllables. The ratio $K_{pr} = F_a/F_p$ gave the value of a prominence. All peak measurements were automatic, so there was a small amount of measurement error. In all cases, the prosodic phrase position and pitch accent location is based on hand-labeled prosodic markers, so that errors due to automatic prediction would not be confounded with the prominence and range prediction errors.

3.2. Prediction Methods

Classification trees (or, decision trees) have been successfully used for predicting abstract prosodic labels [2, 11]. In this work, we use regression trees, which are another form of a decision tree appropriate for predicting continuous variables. The trees were grown using the minimum error criterion and then pruned using cross-validation. Advantages of decision trees in general are that they are well suited for categorical features and that they capture dependencies between different prediction variables. These dependencies become evident from analysis of how a given tree partitions the training data.

A significant drawback of decision trees is that at every step they split the training data and make only part of it available for further estimation. In contrast, a multiplicative model, used in [14], allows for training the coefficients for each factor on the entire corpus by assuming that the different factors are independent. Since multiplicative scaling corresponds to a sum in the log domain, the model can be implemented using multiple linear regression (MLR) of the term $\log K_{pr}$. Like decision trees, MLR training includes an automatic procedure for eliminating variables that do not improve performance.

We also investigated a compromise solution that separates the prediction variables into subsets, designs a regression tree based on each subset, and combines the output of the trees using MLR. Three subsets were used here. Features related to the type and location of accents belonged to $class_1$; features related to posi-

[1] Low accents are too infrequent in our data to be useful and bitonal accents were not labeled with sufficient reliability, e.g. labelers disagreed on L+H* vs. H*, so the labels were collapsed into a single category.

Table 1: *Experimental results on the radio news corpus: The baseline RMS error was 0.103; error reduction is relative to this value.*

Method	RMS error	Error reduction	Parameters
CART	0.069	31%	10
Multiplicative	0.073	29%	10
Mixed	0.073	29%	56

tion of the phrase in the discourse comprise $class_2$; and syntactic features together with named entity make up $class_3$.

4. EXPERIMENTS

4.1. Corpora

The main corpus was drawn from a collection of recorded FM public radio news broadcasts [8]. A training set that we used for these experiments contained approximately 50 minutes of speech representing the thirty-four radio news stories from a single female speaker. Four stories recorded later in a laboratory by the same speaker (12 minutes of speech) are used as independent test data. In addition, we used a 5-minute corpus of narrative stories by Robert MacNeil, which was all used for training since we were primarily interested in the structure of the predictor for assessing genre differences. The corpora were hand-labelled with the ToBI system of prosodic transcriptions [9], and automatically annotated with part-of-speech labels [7] and syntactic structure [1].

4.2. Prominence Prediction

In experiments on the radio corpus, the regression tree outperforms both the multiplicative and the mixed models, though the difference is not significant. Table 1 shows the percentage reduction in RMS error, compared to that obtained by predicting the mean training value for all prominences. Each of the models explains as much as 55% of variation in the training data.

The most important features in the prediction process for all of the methods that we applied were the tone sequence features. The decision tree used local (current and previous) tone labels, but not the number of downsteps up to the current point, though the MLR model did use that feature. The decision tree also used features we expected to be associated with range, i.e. the position of a phrase in a sentence and position of a sentence in a paragraph, though the impact on performance was small (31% vs. 30% RMS error reduction). Neither model was able to predict cases where the highest peak is not the first accent, though there were several instances of such in the training data. One limitation of the result is that it relies on accurate prediction of downstep, so we ran additional experiments using only accent location information (no tone labels) and found only a small loss in performance (29% vs. 31% error reduction).

Because of studies relating pitch range to perceived prominence [12, 6, 10], we tried additional experiments with this feature. If the true values for the phrase F0 range and the previous peak are available, then using these features in prediction reduces the RMS error by 41%, but so far predicted range values are too errorful to be of use.

4.3. Phrase Peak Prediction

Again using the radio data, models were trained for predicting peak F0 value within each phrase. The decision tree used information about minor phrase position in the major phrase and sentence, and sentence position in the paragraph, consistently giving higher values to the starting location at every level (major, sentence or paragraph). For example, phrases that start a paragraph have higher peaks than those that start a sentence or a major (273 vs. 255 vs. 226 Hz, respectively).

Predicting peak F0 values using a decision tree reduced the RMS error by 10% compared to that obtained by predicting the mean training value for all phrases. The MLR model was insignificantly different from using the training mean. In contrast to prominence prediction, the range prediction tree utilized information about the type of syntactic phrase and the number of named entities within the phrase, although these features were much less important than phrase position.

4.4. Genre Differences

Applying the F0 peak prediction methods to data representing a narrative speaking style exhibits similar results. Regression trees trained separately on the two corpora to predict the prominence scaling factor with the three most important (tone sequence) features have identical topologies (see Figure 1). Comparing the predicted values across styles one can see that narrative style prominences have lower values than corresponding radio news ones, which corresponds to bigger F0 drops when downstep occurs (since no raised peaks are predicted). While results need to be confirmed on a larger data set and more speaking styles, they offer hope that a style-independent model topology can be designed with retraining only a small number of parameters needed for capturing genre differences.

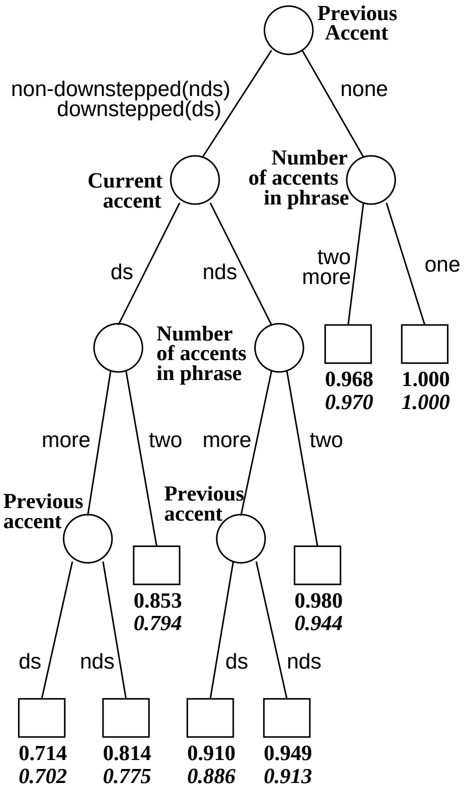

Figure 1: *Prominence prediction tree: numbers at the terminal nodes are normalized F0 peaks; regular font indicates radio style and italics is for narrative.*

5. CONCLUSIONS

In summary, this paper presents a model of gradient F0 peak height based on separate prediction of range and intra-phrase prominence. Experiments on radio news data show that, for relative differences in peak height within a prosodic phrase, over 50% of the variance of training data and 30% of RMS error in independent data can be accounted for by local tone sequence labels. Better results are obtained by incorporating pitch range as a feature, but automatic range prediction was not successful, probably because it requires more sophisticated discourse analysis than that used in this work. The features used for range vs. prominence prediction were, for the most part, different. However, the prominence models failed to predict cases where the highest accent is late in the phrase, and it may be that the appropriate cues to this event will also influence pitch range. Both regression trees and multiplicative models gave similar performance for prominence prediction with a small number of parameters, but multiplicative models were not effective for range prediction. Finally, regression trees trained on corpora reflecting different genres show identical structures, suggesting that the factors influencing relative prominence within a phrase may be style-independent though the specific size of factors like downstep will depend on the particular speaker and style.

Acknowledgements. This work was supported by the U.S. National Science Foundation, grant no. IRI-9528990.

6. REFERENCES

[1] M. Collins, "A new statistical parser based on bigram lexical dependencies," *Proc. ACL*, 1996.

[2] J. Hirschberg, "Pitch accent in context: Predicting prominence from text," *Artificial Intelligence*, 63:305–340, 1993.

[3] J. Hirschberg, "Studies of intonation and discourse," in *Proc. ESCA Workshop on Prosody,* ed. D. House and P. Touati, Working Papers 41, Lund University Dept. of Linguistics, 90–95, 1993.

[4] J. Terken and J. Hirschberg, unpublished study.

[5] D. R. Ladd, "Constraints on the gradient variability of pitch range, or , Pitch level 4 lives!" *Papers in Laboratory Phonology III: Phonological Structure and Phonological Form,* ed. P. Keating, 43–63, Cambridge University Press, 1994.

[6] D. R. Ladd, J. Verhoeven, and K. Jacobs, "Influence of Adjacent Pitch Accents on Each Other's Perceived Prominence: Two Contradictory Effects," *Journal of Phonetics*, 22:87–99, 1994.

[7] M. Meteer, R. Schwartz, and R. Weischedel, "POST: using probabilities in language modeling," *Proc. International Conf. Artificial Intelligence*, 960–965, 1991.

[8] M. Ostendorf, P. Price, and S. Shuttuck-Hufnagel, "The Boston University Radio News Corpus," BU Technical report ECS-95-001. 1995.

[9] J. Pitrelli, M. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," *Proc. ICSLP*, 1:123–126, 1994.

[10] T. Rietveld and C. Gussenhoven, "The Influence of Phrase Boundaries on Perceived Prominence in Two-Peak Intonation Contours," *Proc. EuroSpeech*, 2:859–862, 1997.

[11] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech and Language*, 10:155–185. 1996.

[12] J. Terken, "Fundamental Frequency and Perceived Prominence of Accented Syllables," *J. Acoust. Soc. Am.*, 89(4):1768–1776, 1991.

[13] J. Terken, "Variation of Accent Prominence within the Phrase: Models and Spontaneous Speech Data," in *Computing Prosody*, ed. Y. Sagisaka, N. Campbell and N. Higuchi, 95–111, 1996.

[14] J. van Santen, B. Möbius, J. Venditti and C. Shih, "Description of the Bell Labs Intonation System," *Proc. ESCA Speech Synthesis Workshop,* 293–298, 1998.