

Predicting Types of Pitch Accent and Boundary Tone Using Structural Information

Tae-Jin Yoon

Department of Linguistics
University of Illinois at Urbana-Champaign
tyoon@uiuc.edu

The relationship between prosodic structure and syntactic structure is an unresolved area of inquiry, partly due to the shortage of prosodically transcribed speech corpora, and partly due to the complexity involved in the analysis of both syntax and prosody. Chomsky & Halle (1968, p. 372) state that “although there is a substantial literature on intonational and prosodic features in English, it is largely restricted to citation of examples, and we cannot draw on it for any significant insight into processes of a general nature.” More recently, Ladd (1996, p. 334) adds that “in the standard theory, the correspondence between syntactic constituent types and prosodic ones is highly variable, since the make-up of the prosodic constituents is influenced by a variety of essentially linear factors.” Despite the *status quo*, different views on the mapping from syntax to prosody have been proposed: (1) Syntax alone determines most of prosodic structure (e.g., Cooper & Paccia-Cooper 1980); (2) Speakers use prosody to signal syntactic information only when ambiguity is involved (e.g., Snedeker & Truswell 2003); (3) Many linguistic and para-linguistic factors along with syntax determine prosodic structure (e.g., Bachenko & Fitzpatrick 1990). Besides these linguistic and psycholinguistic studies, machine learning approaches have been employed to improve the performance of TTS or ASR by incorporating prosodic features, either using hand-built rules or using stochastic classification algorithms (c.f., Taylor & Black 1998).

Much of the linguistic and psycholinguistic research on prosody is limited in that it has often relied on data from impressionistic prosody labeling

and/or small data sets of recorded speech. Machine learning of prosody based on syntactic features is also limited by the availability of complete and reliable syntactic parsing, a concern that is reflected in Taylor & Black (1998) who state that “[a]lthough we argued . . . against using syntactic parsers for phrase break assignment, our reason stems from the basic inaccuracy of these parsers, not because syntactic parsers themselves are unhelpful.” One approach to overcoming the data limitation problem that characterizes this prior work is to search for prosody-syntax relations that can be established on the basis of less-than-complete syntactic structures. The method pursued in the research described here uses a shallow parse structure to obtain syntactic information that goes beyond part of speech features, and which is more accurate on continuous speech than the output of a full syntactic parser. The shallow parser produces output that is similar to ‘flattened syntactic structure’ (Chomsky & Halle 1968).

This paper presents results from machine learning experiments on predicting prosodic features related to pitch accents and boundary tones based primarily on shallow syntactic structure and grammatical relations, together with part of speech, basic syllable information, constituent length, and position of word within a syntactic phrase and a sentence.¹ The working hypothesis is that even though there is no one-to-one correspondence between syntax and prosody, the two grammatical components are highly correlated, such that syntactic information is a good pre-

¹POS, syntactic phrase chunk, and grammatical relations are tagged using shallow syntactic parser available at ILK. (<<http://ilk.kub.nl>>)

dictor of prosodic structure.

A subset of Boston University Radio Speech Corpus (BRSC) is used for the experiment (Ostendorf et al., 1995). The BRSC is a speech corpus produced by professional FM Radio News announcers. The corpus is prosodically labeled using ToBI (Tones and Break Indices). The total number of words used for the experiments reported here is about 10,000 and the number of sentences is about 600. Note that since all of the speakers in this study produced the same scripts, the number of word types is quite limited (about 900 word types).

Two machine learning algorithms are used for the experiment: (1) CART using Wagon (Taylor et al., 1999) and (2) Memory Based Learning using TiMBL (Daelemans et al., 2003). Contextual information is encoded for the previous and following n words, with $n = 1$ or 2 . The dataset is divided into training data (90%) and test data (10%). Word information is excluded from the feature set due to the limited number of distinct word types, in order to render the results more robust to a new test data. The accuracy for the predictions for type of pitch accent (H*, !H*, L*, No Pitch Accent) and type of boundary tones (L-, H-, L-L%, L-H%, H-L%, H-H%, and No Boundary Tone) are reported below. In general, Memory based Learning results in a little bit higher accuracy than CART based Wagon for both pitch accent and boundary type prediction. The baseline measure for pitch accent calculated from the test data is 48.04 % (490/1020). Confusion matrices for both Wagon and TiMBL show that while the accuracy for predicting the presence or absence of pitch accent is quite high (85.2%), the best accuracy for predicting the type of pitch accent is reduced to 75.18%. The baseline prediction for boundary tone is 72.6% (741/1020). The best accuracy for predicting the type of boundary tone is 81.86%. As with pitch accent prediction, the accuracy of prediction for the presence or absence of boundary tone is high (90.2%), but prediction of the type of boundary tone is negatively affected by the frequency of each boundary type. The accuracy obtained from this experiment is favorable compared to previous studies. For example, the best score for predicting the presence or absence of phrase break using 6-gram POS tagging in Taylor and Black (1998) is 86.6%, as compared to 90.2%

in this experiment.

The paper concludes with the discussion of possible ways to improve prosody prediction in the current system. The prediction of the type of pitch accent or boundary tone will be more accurate if acoustic information is utilized. For example, my work on the F0 correlates of pitch accent shows significant differences in F0 contour between H* and !H* if contextual F0 information is considered in linear regression analysis. Further improvement is expected if semantic information is introduced to reduce the confusion among types of pitch accents and boundary tones.

References

- Joan Bachenko and Eileen Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170.
- Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York, NY.
- William E. Cooper and Jeanne Paccia-Cooper. 1980. *Syntax and Speech*. Harvard University Press, Cambridge, MA.
- Walter Daelemans, Jakub Zavrel, Kovan. van der Sloot, and Antal van den Bosch. 2003. TiMBL: Tilburg Memory Based Learner, version 5.0. *Reference Guide. ILK Technical Report 03-10*,
- D. Robert Ladd. 1996. *Intonational Phonology*. Cambridge University Press, Cambridge, UK.
- Mari Ostendorf, Pattie P. Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University Radio News Corpus. *Technical Report ECS-95-001*, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA.
- Jesse Snedeker and John Truswell. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, 48(1):103–130.
- Paul Taylor, Richard Carey, Alan W. Black, and Simon King. 1999. Edinburg Speech Tools. *System Documentation Edition 1.2*,
- Paul Taylor and Alan W. Black. 1998. Assigning phrase breaks from part-of-speech sequence. *Computer Speech and Language*, 12:99–117.