A Predictive Model of Prosody through Grammatical Interface: A Computational Approach

Tae-Jin Yoon
Ph.D. Department of Linguistics (2007)
University of Illinois at Urbana-Champaign

Advisor: Jennifer Cole
Committee members: Mark Hasegawa-Johnson, Richard Sproat, Chin-Woo Kim, Chilin Shih

Speech prosody is manifest in the acoustic signal through the modulation of pitch, loudness, duration, and source characteristics (voice quality), which combine to encode the prosodic structure of an utterance. Prosodic structure defines the location of prominent words and syllables, and the grouping of words into phonological phrases. Prosodic structure, in turn, relates the phonological form of an utterance to its morphological, syntactic, semantic, and pragmatic context. The listener’s task in comprehending speech includes decoding prosodic structure to aid in identifying the morphological, syntactic, semantic, and pragmatic contexts that comprise the meaning of the utterance.

The research reported in this dissertation focuses on acoustic and perceptual evidence for prosody in spoken language, and the relationship between prosodic structure and higher levels of linguistic organization. The study adopts a computational approach that employs natural language processing tools, machine learning algorithms, and speech and signal processing techniques to investigate prosody in speech corpus data. In this study, I show that prosodic features of an utterance can be reliably predicted from a set of features that encode the phonetic, phonological, syntactic and semantic properties of the local context. The study uncovers new evidence of the acoustic correlates of prosody, including prosodic phrase juncture and downstepped pitch-accent in American English, in features related to F0, duration, and intensity. The study also demonstrates in a series of machine learning experiments that these acoustic features and features from ‘higher’ levels of linguistic organization are highly correlated with each other, and that very accurate prediction of prosodic structure can be achieved on the basis of structural linguistic properties and that detection of prosodic structure can also be made with a high degree of accuracy on the basis of acoustic cues.

The research in this dissertation will make both theoretical and applied contributions to the study of speech prosody. On the theoretical level, my research will contribute to a better understanding of how different grammatical and/or acoustic features interact in forming prosodic prominence and phrasing. The proposal is expected to address the concern expressed by Ladd (1996), who states 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 (p. 334).” On the applied level, my research will inform the development of systems for the automatic prediction of prosodic categories, which in turn will enable the creation of Text-To-Speech (TTS) systems with enhanced intelligibility and naturalness. My research will also facilitate work on prosody detection for use in Automatic Speech Recognition (ASR) systems. While my research is not directly concerned with
improving ASR systems, it can be viewed as the first step towards the goal of automatically obtaining prosodically-labeled data as a means of bootstrapping prosodic analysis for ASR.

In what follows, I summarize remaining chapters in my dissertation:

In Chapter 2, I present the prosodic model that serves as the theoretical basis for my research, along with the speech corpus used for the experiments. I describe in detail the standard prosody annotation system, i.e., the ToBI system, for American English. Then, I present the Boston University Radio Speech corpus (Ostendorf et al. 1995), a large prosodically-transcribed database that is used throughout in this dissertation. While presenting the radio speech corpus, I review transcriber reliability studies reported for this corpus, and demonstrate the speaker variation (or consistency) observed in the corpus.

Chapter 3 presents an overview of a machine learning algorithm and summarizes earlier studies on the prosodic structure prediction. Probabilistic approaches are more suitable than deterministic approaches in describing and modeling prosodic structure, due to its variability. Machine learning approaches, as one of such probabilistic approaches, possess attractive characteristics in that a machine learning algorithm finds underlying generalization of the data. I review two such algorithms, the memory-based learning (MBL) algorithm and classification and regression tree (CART). The two algorithms have been successfully and widely used in many research areas including natural language processing as well as prosody modeling. I turn to the presentation of standard evaluation metrics such as baseline, precision, recall and accuracy that are typically employed to evaluate the performance of machine learning algorithms. I conclude the chapter by summarizing earlier studies of prosodic structure prediction.

Chapter 4 demonstrates the predictive models of prosodic structure through grammatical interface. I provide a probabilistic model of the mapping between prosody and phonological, syntactic, and semantic features. The model encodes phonological features, shallow syntactic constituent structure, argument structure, and the status of words as named entities. A machine learning experiment using these features to predict prosodic phrase boundaries achieves more than 92% accuracy in predicting prosodic boundary location. The experiment of predicting prosodic prominence location achieves over 87% accuracy. These accuracy rates are the highest rates reported for the tasks conducted on the same corpus. This study sheds light on the relationship between prosodic structure and other grammatical structures. But at the same time, the study reveals some aspects of prosodic structure that are not well understood and controversial. These aspects are further investigated in the following chapters.

Chapter 5 presents experimental results of predicting prosodic structure through the integrative set of acoustic and linguistics features derived from both the speech signals and the grammatical structures. In the previous chapter, I have demonstrated that linguistic features contribute much to the determination of the prosodic prominence location and the prosodic boundary location, as evaluated by the high accuracy rates. Prosodic structure can be approached from different perspectives: On one hand, the prosodic constituents are investigated based on the syntactic structures of an utterance (Selkirk 1984, Nespor & Vogel 1986, cf. Steedman 2000). The syntax-driven approach seeks to understand the mapping from syntactic structure to intonational phrasing. On the other hand, the Autosegmental-Metrical theory of intonational phonology
Summary of the dissertation

Tae-Jin Yoon

(Pierrehumbert 1980, Beckman & Pierrehumbert 1986), on which the ToBI system is based, investigates prosodic constituents on the basis of the perceived intonation pattern of an utterance. The phonology/phonetics-driven perspective seeks to understand the phonological structures that encode prosodic phrasing and accentuation, and how these structures relate to other aspects of phonological structure (e.g., syllable, metrical structure). It is also concerned with the acoustic correlates of intonational events, as a way of establishing the empirical basis of investigation. In this chapter, I show that experimental results obtained through the predictive model of prosodic structure, integrating features extracted from grammatical components and the acoustic signal show that the linguistic features and acoustic cues are highly correlated with each other. The results lead us to conclude that the prosodic structure can be predicted on the basis of structural linguistic properties, and detected on the basis of acoustic cues.

Chapter 6 investigates the acoustic correlates of aspects of prosodic structure, concentrating on the acoustic correlates of levels of prosodic phrasing (intermediate phrase (ip) vs. Intonational Phrase (IP)) on the one hand, and the acoustic correlates of downstepped pitch accent on the other hand. These two aspects of the prosodic structure are not well understood and are controversial, and the machine learning approaches in the previous chapters are limited in their ability to uncover new evidence. The study reported in this chapter uncovers new acoustic evidence for the distinction between two levels of prosodic phrase juncture and for the existence of downstepped pitch-accent. In the first part of the chapter, I present acoustic evidence from the radio speech corpus for a distinction between levels of prosodic boundaries. I investigate the phonetic encoding of prosodic structure through analysis of the acoustic correlates of prosodic boundary and the interaction with phrase stress (pitch accent) at three levels of prosodic structure: Word, ip, and IP. Evidence for acoustic effects of prosodic boundary is shown in measures of duration local to the domain-final rhyme. These findings provide strong evidence for prosodic theory, showing acoustic correlates of a 3-way distinction in boundary level. In the second part of the chapter, I present evidence from acoustic analysis and a machine learning experiment for a categorical distinction between non-downstepped and downstepped high-toned pitch accent (H* vs. !H*). The experimental findings from naturally occurring speech corpus provide evidence for !H* as a distinct prosodic category.

Chapter 7 concludes the dissertation. This research contributes to our understanding of the interaction between components of linguistic grammar, in demonstrating the dependencies between phonetics, phonology, syntax and semantics in the encoding of prosody. In addition, my work building on a stochastic model of prosody prediction has a direct application in the development of speech technologies that incorporate linguistic models of prosody, including text-to-speech and automatic speech recognition systems.