Abstract for poster presented April 11, 2008 at *Experimental and Theoretical Advances in Prosody*, Cornell University. Poster available at http://www.novaspeech.com/documents/cornell2008.pdf

The Role of Prosody in Speech Parsing

Susan Hertz*†, Masayuki Gibson*†, Nina Glatthorn*, Patrick Hegde*, Harold Mills* and Isaac Spencer*

**NovaSpeech LLC, Ithaca, NY*
www.novaspeech.com
†*Cornell University, Ithaca, NY*

Just how listeners manage to make sense of the highly variable and complex speech signal has been a major puzzle facing speech researchers. How can a continuous speech signal be parsed into discrete linguistic units like phonemes and words even though the same acoustic sequence can often realize different linguistic structures depending on its context? This presentation addresses this question, demonstrating aspects of an iterative two-stage model of speech parsing in which listeners incrementally derive phonological structure from a speech signal by identifying certain *parsing anchors* (stage 1) and by interpreting the acoustic structure between successive pairs of anchors (stage 2). We focus on the role of prosodic units within this model, demonstrating how the speech signal is interpreted with reference to syllabic and phrasal context.

More specifically, we present evidence that, as the speech signal is received, the listener identifies certain robust context-independent acoustic events, generally periods of relatively high-amplitude vocalic resonances corresponding to stressed syllable nuclei, or long pauses corresponding to the ends of phrases. At a certain point—the *parsing anchor*—within each of these events, listeners parse the interval—the *parsing interval*—between that point and the preceding anchor, identifying phonological structure within the interval. This phonological structure might include, for example, the syllable-initial consonants of the current syllable, a preceding syllable boundary, and the syllable-final consonants and/or nucleus of the preceding syllable.

We show that in deriving phonological structure within a parsing interval (PI), listeners are sensitive to the phone-and-transition structure (in the sense of Hertz 1991) in the PI, as well to the perceptually important properties of *acoustically distinct intervals* (ADIs) in the PI. Examples of ADIs include (1) a stretch of high-intensity, high-frequency noise corresponding to one or more [s] phones or an alveolar stop burst, (2) a stretch of silence corresponding to one or more stop closures, and (3) a stretch of high-intensity periodicity corresponding to one or more syllable nuclei. We focus on how phonemes and syllable boundaries are inferred from the relationships of properties of both ADIs and phones and transitions. In the case of ADIs, relevant properties include intensity, duration, and certain spectral characteristics; in the case of phones and transitions, properties include duration and formant target values.

We present examples of how acoustically similar but phonologically different sequences are distinguished using prosodic information. One such example is shown in Figure 1 below, which contains spectrograms of the utterances *Say bead to me* and *Say beat* produced by the same speaker of American English at similar rates of speech. ADIs and parsing anchors relevant to the example are shown. Using prosodic information obtained at the parsing anchors—in particular, the fact that 1a is in a phrase-medial context and 2a in a phrase-final one—listeners perceive ADI 1a as a relatively long /i/ nucleus, while they perceive the acoustically similar ADI 2a as a relatively short /i/ nucleus (the perceptual difference helping to cue the distinction in voicing of the post-vocalic consonant in both cases). Similarly, ADIs 1b-d (the

long silence, burst, and aspiration, respectively) are interpreted as two phonemes /d/ and /t/ with an intervening syllable boundary, while ADIs 2b-d (the short silence, burst, and aspiration, respectively) are interpreted as a single syllable-final /t/. The influence of prosodic context on parsing becomes evident when utterance (1) is played up to the end of ADI 1d. In this case, listeners parse ADIs 1a-d in a phrase-final context, thus perceiving ADI 1a as a relatively short /i/ and ADIs 1b-d as acoustically appropriate for /t/. As a result, they hear the word *beat* rather than *bead*.

More generally, the model is supported by integrated results from more than thirty years of hands-on work in phonetics and phonology, including multi-language speech synthesis by rule (Hertz 1991, Hertz *et al*. 1999, Hertz 2006). It provides insights into both language-universal and language-specific timing patterns, phonological tendencies across languages, language-specific phonological and phonetic patterns, and the range of variation that can occur in the acoustic realization of a given linguistic structure in a given language.

[This work was supported by NIH NIDCD Grant 5R44DC006761 awarded to NovaSpeech LLC. The views are those of the authors and do not necessarily reflect the views of the NIDCD or NIH.]

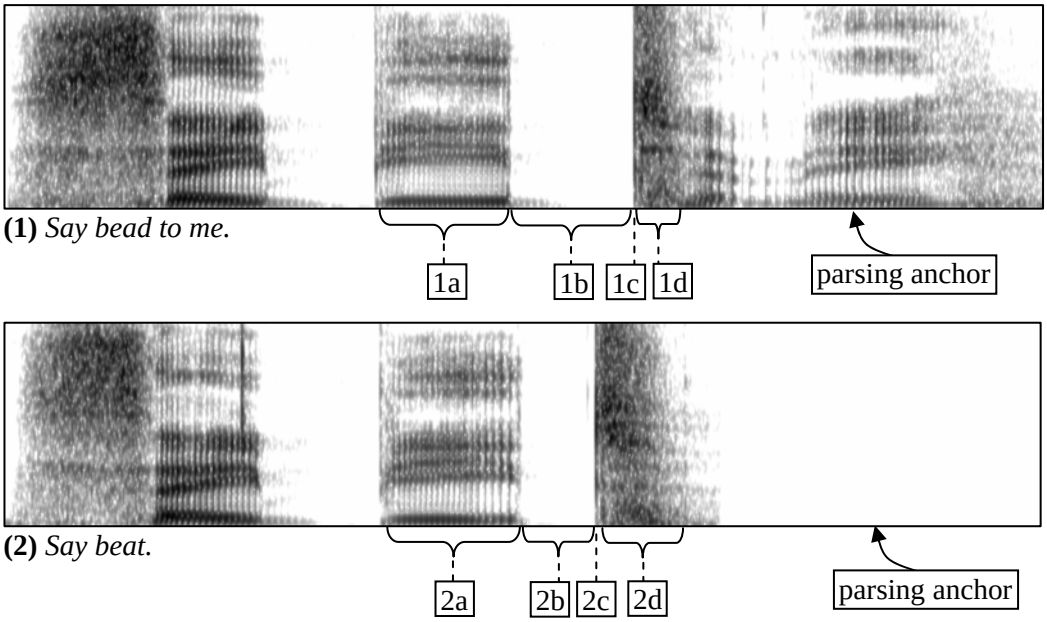

Figure 1. ADIs 1a and 2a (the nucleus [i]) above are both about 155 ms long. ADI 1b (the single closure for the consonants [dt]) is 140 ms long, while ADI 2b (the closure of the consonant [t]) is 80 ms long. Only the parsing anchors relevant to this example are shown; among others, there are parsing anchors early in ADIs 1a and 2a.

References

Hertz, S. R. (1991) Streams, phones, and transitions: toward a phonological and phonetic model of formant timing. *J. of Phonetics* 19, 91-109. *Special Issue on Speech Synthesis and Phonetics*, ed. R. Carlson.

Hertz, S. R., Younes, R. J., and Zinovieva, N. (1999) Language-universal and language-specific components in the multi-language ETI-Eloquence text-to-speech system. *Proc. 14th ICPhS*: 2283-2286.

Hertz, S. R. (2006) A model of the regularities underlying speaker variation: Evidence from hybrid synthesis, *Proc. Interspeech 2006*.