THE ROLE OF PROSODY IN SPEECH PARSING

Presented April 11, 2008 at *Experimental and Theoretical Advances in Prosody*, Cornell University. Slightly revised.

Susan Hertz*, Masayuki Gibson*, Nina Glatthorn, Patrick Hegde, Harold Mills and Isaac Spencer **NovaSpeech LLC and *Cornell University** **www.novaspeech.com**

I. How do we parse continuous speech into discrete phonological units?

Answers have eluded us because, for example:

- The same phoneme may have different realizations in different phonological or phonetic contexts.
- The same acoustic information in different contexts may realize different phonological structures.

Despite the apparent complexity of speech, we show that speech is decoded in a principled fashion. We present a model of speech parsing, focusing on the central role that prosody plays within this model in accounting for a listener's ability to interpret the speech signal.

II. An iterative two-stage parsing model

OVERVIEW

Set *parsing anchor** A to start of utterance.

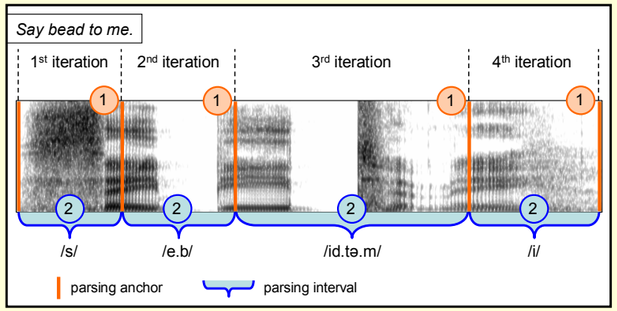
While parsing anchor A is not end of utterance:

1. **Locate parsing anchor B** in reference to next stressed syllable or phrase boundary.
2. **Identify phonological structure** (e.g., phonemes and syllable boundaries) by interpreting the *parsing interval** I_{AB} between parsing anchors A and B. More specifically, identify this structure by determining the *acoustically distinct intervals** and *transitions** in the parsing interval and interpreting their perceptually relevant attributes.

Set parsing anchor A to parsing anchor B.

* See explanations in Section III.

EXAMPLE: HOW PARSING PROCEEDS



III. Definitions and concepts

PARSING ANCHORS AND INTERVALS

A ***parsing anchor*** is a point at which phonological structure for an earlier stretch of the acoustic signal (a ***parsing interval***) can be unambiguously determined—i.e., the point after which the listener is unlikely to change his or her phonological interpretation of the prior portion of the utterance, except as a result of higher-level inferences (e.g., semantic inferences).

Parsing anchors are typically located:*

(1) shortly after formant transitions into the vowels of stressed syllables.†

(2) in pauses.

* Parsing anchors located at phrase boundaries are still under investigation, but F0 almost certainly plays a role in cuing them.

† We use the term "stressed syllable" loosely, as some parsing anchors may occur in sufficiently long non-reduced vowels lacking primary stress.

ACOUSTICALLY DISTINCT INTERVALS (ADIs)

Acoustically distinct intervals (ADIs) are abutted stretches of speech defined by collections of perceptually relevant attributes, such as:

- Type of sound source (e.g., noise and/or voicing)*
- Spectral characteristics (e.g., high vs. low center of gravity)
- Relative intensity

In a spectrogram, there are typically visible discontinuities at the boundaries between ADIs.

* Both a primary and secondary source may be associated with an ADI (e.g., voicing + aspiration for a breathy vowel, or silence + voicing for a closure with a voice bar, but this detail is ignored in the examples below.)

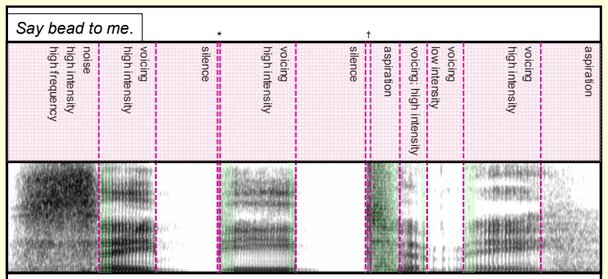
PHONES AND TRANSITIONS

Phones and ***transitions*** are define largely as in Hertz (1991). While both phones and transitions play a role in the model, the role of phones is outside the scope of this poster. ***Transitions*** have relevant perceptual properties, such as:

- Direction of formant movements (e.g., rising vs. falling)
- Length (i.e., short, medium, long)*

* Note that the resistance of the durations of transitions to contextual variation, described in Hertz (1991), is relied on by listeners for successful parsing.

EXAMPLE: ADIs AND TRANSITIONS



EXAMPLE: PHONEMES AND THEIR UNDERLYING PERCEPTUALLY RELEVANT ATTRIBUTES

Excerpt from *Say bead to me*: see also Example 2 in Section IV.

phoneme	…	i			d	t	ə		…
nucleus* length		long					short		
transitions		trans		trans			trans		
transition length		short		short			short		
ADI length					long	med			
source type		voicing			silence	noise	aspiration	voicing	
intensity		high				high		high	
center of gravity						high			

Note that the static nature of this and other diagrams is not meant to suggest a processing order. In fact, various attributes for ADIs may be arrived at in orders depending on prior and/or subsequent information within a parsing interval.

* See Hertz & Huffman (1992) for a definition of "nucleus" within the model.

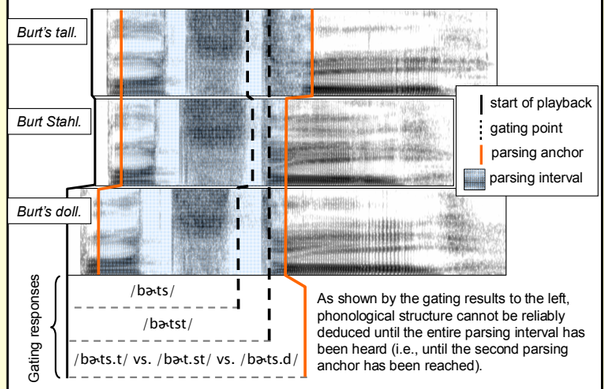
REFERENCES:

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. and Huffman, M. K. (1992) A nucleus-based timing model applied to multi-dialect speech synthesis by rule, *Proc. International Conference on Spoken Language Processing* 2, 1171-1174.

IV. Parsing examples

EXAMPLE 1. Identification of syllable boundaries and consonants in parsing intervals: /s.t/ vs. /.st/ vs. /s.d/



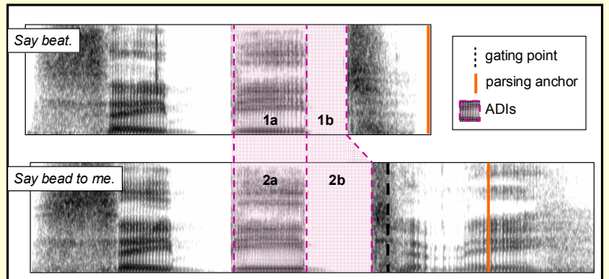
As shown by the gating results to the left, phonological structure cannot be reliably deduced until the entire parsing interval has been heard (i.e., until the second parsing anchor has been reached).

The table to the right, for example, shows some of the cues that are available at the parsing anchor for distinguishing the phonemes and syllable boundaries in the above utterances.

Perceptual attribute	Helps cue
Aspiration in transition immediately preceding parsing anchor	/.t/
Relative lengths of noise ([s]) and silence ([t] and/or [d]) intervals	/.st/ vs. /s.d/

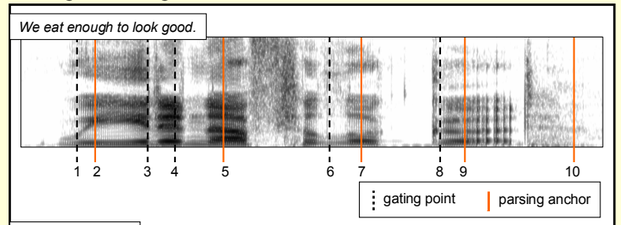
Other attributes used more generally to determine the place and manner of articulation of the consonants include the lengths of transitions, the directions of the formant movements in the transitions, and the frequency characteristics of noise intervals.

EXAMPLE 2. Interpretation of durations relative to phrasal context: phrase-final /it/ vs. phrase-medial /id.t/



- When the utterance *Say bead to me* is played up to the gating point shown, listeners hear *Say beat*. The model can explain this as follows:
  - ADIs 2a and 2b are interpreted as phrase-final (where lengthening is expected).
  - The nucleus (realized by ADI 2a) is interpreted as short given its phrase-final context (cf. ADI 1a in the utterance *Say beat*).
  - ADI 2b (the closure) and the following stop burst, coupled with the transition into the closure, have attributes that cue a /t/ in phrase-final position. Note that although ADI 2b is longer than ADI 1b in the naturally uttered *Say beat*, it is nonetheless interpreted as a single stop, a fact that can be explained by the model.
  - The "short" ADI 2a helps cue the perceived final consonant as voiceless.
- In contrast, when ADIs 2a and 2b are interpreted in their original, non-gated context, the listener recognizes their phrase-medial status and interprets ADI 2a as representing a long nucleus and, similarly, ADI 2b as relatively long and, hence, as representing two consonants.

EXAMPLE 3. Explanation of gating "confusions": *We eat enough to look good*



Gating responses

Gating point	Listener 1	Listener 2	Listener 3
1	[w]	[w]	[bʊ]
2	[wi]	[wi]	[wi]
3	[wi.ɨp]	[wi.ɨ]	[wi.ɨt]
4	[wi.rə]	[wi.ɨ.rə]	[wi.ɨ.rɨt]
5	[wi.i.rɨ.nə]	[wi.i.rɨ.nə]	[wi.i.rɨ.nə]
6	[wi.i.rɨ.nʌf.fəl]	[wi.i.rɨ.nʌ.fəl]	[wi.i.rɨ.nʌf.tə]
7	[wi.i.rɨ.nʌf.fə.lə]	[wi.i.rɨ.nʌ.fə.lə]	[wi.i.rɨ.nʌf.tə.lə]
8	[wi.i.rɨ.nʌf.tə.luk]	[wi.i.rɨ.nʌ.fə.luk]	[wi.i.rɨ.nuf.tə.luk]
9	[wi.i.rɨ.nʌf.tə.luk.gu]	[wi.i.rɨ.nʌ.fə.lʊ.gu]	[wi.i.rɨ.nuf.tə.luk.gu]
10	[wi.i.rɨ.nʌf.tə.luk.gud]	[wi.i.rɨ.nʌf.tə.luk.gud]	[wi.i.rɨ.nuf.tə.luk.gud]

Analysis of responses

Type of "confusion"	Sample "confusion"	Gating point (GP) and listener (L)	Selected reason for "confusions"
Number of syllables	/i.i/ → /i/	GP4 L1	The "vowel ADI" between GP1-GP3 is interpreted as part of a two-syllable phrase-final word, with the duration and other cues appropriate for a single /i/ in this context.
Manner of articulation	/w/ → /b/	GP1 L3	A major cue distinguishing /b/ and /w/ in this context is the length of the following transition, which has not yet been heard.
Stress	/i/ → /ɨ/	GP3 L1,2,3	The identity of a stressed syllable nucleus cannot be fully determined until the following parsing anchor, when its duration, F0 and intensity contours, etc. can be interpreted.
	/ʌ/ → /ə/	GP5 L1,2,3	
	/ʊ/ → /ə/	GP7 L1,2,3	
Syllable boundary	/ə.l/ → /əl./	GP6 L1,2	The [l] is considered phrase-final and, hence, part of the final syllable until the parsing anchor in the following vowel is encountered.
	/luk.g/ → /luk./	GP8 L1,2,3	The silence ADI is taken to be a single consonant due to its interpretation as phrase-final.
Place of articulation	/t/ → /p/	GP3 L1	Formant transitions from high front vowels into labials and alveolars are similar.
	/t/ → /f/	GP6,7 L1 GP6-9 L2	There are no ADIs unambiguously cuing the /t/. Listeners use semantic inferences at different points to infer the /t/.

In general, the model explains these sorts of "confusions" and how speech is organized to avoid them at parsing anchors—i.e., at stressed syllables and phrase boundaries.

V. Final remarks

While we have illustrated the model through gating examples, we stress that the model has evolved by integrating a wide range of data and results over the course of thirty-five years of hands-on research in multi-language synthesis by rule and related areas.

Given space, we would tell you about how the model can account for:

- Phonological and phonetic tendencies across languages
- Constraints on variation in the acoustic realization of phonological structure
- Timing patterns
- Patterns of language acquisition
- … and more

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 NIDCD or NIH.