ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 4pSCb: Production and Perception I: Beyond the Speech Segment (Poster Session)

4pSCb2. Exploring prosodic boundaries: Gradiency and categoricity of prosodic boundaries in articulation
Jelena Krivokapic*, Christine Mooshammer and Mark Tiede

*Corresponding author's address: Yale University, Haskins Laboratories, New Haven, Connecticut 06511, jelena.krivokapic@yale.edu

The prosodic hierarchy is a core concept of prosodic theory. Despite this, the number of categories in the hierarchy, and the structural relationships between them, is not clear. For English, a major and a minor category above the word level is usually assumed, and some studies have suggested additional categories (cf. Shattuck-Hufnagel & Turk 1996). Each of these category levels is assumed to be marked by categorically distinct boundaries, but experimental evidence for this view is sparse. In this work EMA is used to investigate this notion of the prosodic hierarchy (following a preliminary analysis in Krivokapic & Ananthakrishnan 2007). Seven subjects read six repetitions of 48 sentences, each containing one, two or three prosodic boundaries, for a total of 56 boundaries per speaker. The predicted boundary strength varied from a weak clitic boundary to a strong sentential boundary. The produced boundaries are evaluated using a Gaussian mixture model analysis. The results bear on the question whether prosodic boundaries behave categorically (i.e., prosodic boundary values cluster within a small number of categories), or in a gradient manner (i.e., they are more evenly spread), thus supporting an alternative view of the prosodic hierarchy. [Supported by NIH DC003172-16, DC008780, DC002717].
INTRODUCTION

A core question of prosodic theory is the number and type of phrasal categories in the prosodic hierarchy. Most theories of English assume a word category, and a minor and major category above the word level, but additional categories have been suggested as well (e.g., Beckman & Pierrehumbert 1986, Selkirk 1984, Nespor & Vogel 1986; see overview in Shattuck-Hufnagel & Turk 1996). Phonological arguments for these categories have been made, but not many experimental studies have addressed this question (see review in Ladd 1996, Krivokapić & Byrd 2012). Among the findings taken as evidence for the postulated number of categories is Wightman, Shattuck-Hufnagel, Ostendorf & Price (1992), who in a corpus study of read speech show that four prosodic categories can be distinguished by lengthening of the vowel of a phrase final syllable. On the other hand, the data shown for the lengthening of the syllable coda in Wightman et al. (1992) indicate that five categories could be distinguished. Ladd (1988) shows that depth of embedding of Intonation Phrase (IP) categories leads to boundaries of different strengths. In perception, a study by de Pijper and Sanderman (1994) observes that the perceived boundary strength values do not seem to cluster around a limited number of target values, as would be expected under a categorical model. Krivokapić & Ananthakrishnan (2007) show evidence that listeners perceive five categories. Together, these studies suggest the existence of more than the three commonly assumed categories and indicate that a more gradient view of the prosodic hierarchy might be more appropriate. In addition, Krivokapić & Byrd (2012) show that speakers can produce and listeners perceive IP boundaries of different strengths, thus motivating a recursive structure which could drive a more gradient production and perception of prosodic boundaries.

The goal of the present study is to examine whether prosodic categories are produced in a categorical or in a gradient manner. Specifically the question is whether the boundary production values cluster in a small number of clusters (indicating a categorical production), as would be expected under for example Beckman & Pierrehumbert’s (1986) model or Nespor & Vogel’s (1986) model, or alternatively whether they cluster in a more gradient manner (as suggested by Ladd’s 1988 analysis). In order to address this question, we conducted an electromagnetic articulometer (EMA) experiment and analyze the data using mixture modeling.

METHODS

Stimuli and Participants

Forty-eight sentences were constructed, each containing one, two or three prosodic boundaries, for a total of 56 boundaries. Each boundary was between the words “column” and “and”. To elicit variability in boundary strength, syntactic structure and phrase length before and after the boundary were manipulated. The predicted boundary strength varied from a weak clitic boundary to a strong sentential boundary. Seven native speakers of American English read six repetitions of these sentences for a total of 2352 boundaries. The sentences were pseudo-randomized in blocks of 48 sentences. A sample of the sentences is given in Table 1.

<table>
<thead>
<tr>
<th>TABLE 1. Sample stimuli (6 from 48 sentences). The boundary is marked by ‘#’. The bold words were read with emphasis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The magazine uses an interactive column-n-twitter design. Johnny likes to read a traditional column # and the weather report. At night, it’s the sports column, # and in the morning, it’s the obituaries. A sports column, # and a food column, # and a coffee, are all I need now. She wrote a long, unpleasant, and redundant column. # And then, she went to Paris. She was surprised Gary wanted to write a column. # And a sweet surprise it was.</td>
</tr>
</tbody>
</table>

Data Collection

Data were collected using two 3D EMA devices; the Carstens AG500 for four subjects, and the NDI WAVE system for three subjects. Five subjects were collected at Haskins Laboratories, and two at the University of Southern California. For the AG500 experiments sensors were placed on the upper and lower lip, tongue tip, tongue body, and tongue dorsum, and four reference sensors were placed on the nose, maxilla, and the mastoid processes behind the ears. Placement for WAVE experiments was similar, but lacked nose and tongue body sensors. For this study, only the lip sensors and the reference sensors were used.
The articulatory data were sampled at 200Hz (AG500) or 100Hz (WAVE), and acoustic data at 16kHz. Trajectory data were low-pass filtered at 20 Hz, corrected for head movement (using the reference sensors) and rotated to the occlusal plane.

Data Analysis

Data were analyzed using mview (Haskins Laboratories). We examine the pre-boundary Lip Aperture (LA) movement for /m/ in “column”. Lip Aperture is computed as the Euclidian distance between the position signals for the upper and lower lip, and its corresponding velocity is obtained through central differencing of the displacement signal.

Three time points were identified from the LA velocity signal for this study: the onset of the closing movement for /m/, the onset of the opening movement, and the end of the opening movement (see Figure 1). These time points were identified from the lip aperture zero-crossings on the velocity trajectory. From these data we derive the dependent variables, namely 1) the duration of the complete gesture (closing and opening movement) and 2) the duration of the release part (opening movement) of the gesture only. The pre-boundary opening movement is the movement closest to the boundary, and as such is expected to show the strongest effects of the boundary (based on the prediction of the π-gesture model of Byrd & Saltzman 2003, and empirical evidence, e.g., Byrd, Krivokapić & Lee 2006, Katsika 2012).

![FIGURE 1](image-url)

**FIGURE 1.** Example of a labeled lip aperture gesture ([m] in “column”). The vertical lines mark the landmarks described in the text: GONS marks the onset of the closing movement of the gesture, NOFFS the onset of the opening movement, and GOFFS the end of the opening movement.

Statistical Analysis

For the two variables, we examine the histograms, as a first approximation of the data, and we conduct a mixture model analysis. While histograms are reasonable for exploratory analysis, they cannot tell us about the number of underlying distributions within the data. We use Gaussian Mixture Models to examine this question. Mixture Models are an unsupervised clustering technique, which critically does not rely on a number of clusters defined by the experimenter. Instead, data are fit for a number of clusters (one, two, three, and so on) and the model determines the best fit of the data. We used R to conduct the Mixture Model analysis (mclust Version 4 for R, see Fraley, Raftery, Murphy, & Scrucca 2012). To select the best model, the Bayesian Information Criterion (BIC) is used with a penalty term for an increased number of parameters. The BIC used is given below. The first part of the formula gives the maximized log-likelihood of the data according to the model, and the second part represents the penalty term, which depends on the number of distributions assumed for the data, i.e., it increases with the number of distributions assumed by the model.

\[
BIC = 2 \log \text{lik}_M(x, \theta^*_k) - (\# \text{ params})_M \log(n),
\]
Results

We present the preliminary results for one subject analyzed to date. Analysis of the remaining subjects is currently being completed. The results for the whole gesture duration are given in Figure 2 and for the opening movement duration in Figure 3. For the gesture duration, the histogram shows one clear peak, indicating the existence of one prosodic category. However, in histograms it is unclear how many distributions are driving the peaks. The mixture models analysis shows that the best fit of the data to the model is with three underlying distributions (determined by the highest BIC score) indicating that there are two categories driving the distribution of the data. The center of these clusters are at values 183ms and 243ms.

For the duration of the opening movement, the histogram again shows one peak, but the results of the mixture model analysis indicate that there are three categories driving the distribution of the data. The centers of the clusters are at 68ms, 106ms, and 160ms, and the third component has only seven items. A preliminary analysis of twenty boundaries surrounding each of these cluster centers indicates that one of these clusters mostly consists of boundaries of the type Intermediate Phrase, one of boundaries of the type Intonation Phrase, and the third cluster also consists predominantly of Intonation Phrase boundaries.

FIGURE 2. Gesture (closing and opening movement) duration. Histogram and results of the mixture model analysis.

FIGURE 3. Opening movement duration. Histogram and results of the mixture model analysis.
DISCUSSION

Data from one speaker were analyzed. The distribution of the duration values for the pre-boundary movement closest to the boundary (the opening movement) shows evidence for three different categories. The results from the whole gesture movement suggests only two categories, indicating that the effect of the boundary decreases further away from the boundary (as predicted under the π-gesture model of Byrd & Saltzman 2003). Regarding the question of gradience vs. categoricity of prosodic boundary production, the results of the analysis of one speaker provide support for a categorical distribution of prosodic boundaries (as predicted under, e.g., Beckman & Pierrehumbert’s 1986 model). Preliminary analysis suggests that these categories are ip, IP and another IP. While further analysis and more speakers are needed, this distribution of boundaries indicates that one of these categories (IP) might be recursive.

ACKNOWLEDGMENTS

We are grateful to Gregg Castellucci and Patrick Verdier for data labeling. The work was supported by NIH (DC003172-16, DC008780, DC002717)

REFERENCES