Visual Grouping and Prosodic Grouping: Effects of Spatial Information on Prosodic Boundary Strength

Edward Holsinger, David Cheng-Huan Li, Elsi Kaiser and Dani Byrd

Department of Linguistics, University of Southern California, Los Angeles, USA

{ holsinge, lidc, emkaiser, dbyrd } @ usc.edu

Abstract

We report two psycholinguistic experiments investigating whether grouping information presented in the visuo-spatial modality influences language production – in particular, whether different visual groupings influence the prosodic groupings that speakers produce. We used a picture-description task where three objects were grouped in different ways, and investigated whether spoken descriptions of objects that are spatially closer to each other are separated by weaker prosodic boundaries than descriptions of objects that are further apart. Our results suggest that prosodic boundary strength is influenced by the distance between objects, and that visual input influences linguistic production at the level of prosodic boundaries.

Index Terms: psycholinguistics, visual information, prosodic boundary, language perception, language production

1. Introduction

Prosody refers to the phrasal organization and accentual prominence in speech. Prosodic boundaries elicit systematic changes in the acoustic and articulatory properties of speech. Various acoustic studies have demonstrated that segments increase in duration at boundaries (Gaitenby 1965; Oller 1973; Klatt 1975; Shattuck-Hufnagel & Turk 1998) and lengthening of articulatory gestures around boundaries has likewise been shown (e.g., Byrd & Saltzman, 1998, 2003). Studies have found that listeners can systematically perceive which words are grouped into a processing unit based on these changes (Wightman et al. 1992; Lee & Cole 2006; Krivokapic 2007).

A number of factors influence prosodic boundary strength as indexed by lengthening, whether or not a pause will occur, and, if so, the duration of pausing. Relevant factors include constituent structure (Selkirk, 1981; Sanderman & Collier, 1995), speech rate (Fletcher, 1987; Trouvain & Grice, 1999), and discourse structure (Ayers, 1994; Venditti & Swerts, 1996). In addition, speaker-specific and task-specific effects are often observed.

1.1. Aims of this research

The research presented in this paper tests whether a non-linguistic factor, namely visuo-spatial grouping, can influence the strength of prosodic boundaries. In other words, we test whether visual grouping influences prosodic grouping. If a speaker sees two objects that are spatially closer to each other vs. two objects that are spatially further apart, will this influence the strength of the prosodic junctures that the speaker produces between noun phrases referring to those objects?

By investigating whether visual grouping influences prosodic grouping, this research aims to contribute to our understanding of the nature of the relation between visual input and language production. It is well-known that visual input shapes *what* people talk about. We are interested in a more indirect, more subtle connection: Does visual grouping information influence linguistic grouping, as reflected by prosodic boundary strength?

We investigate whether visuo-spatial distance correlates with prosodic boundary strength in two ways. First, we report a production study that tests whether pause duration in language production is influenced by visuo-spatial factors. Second, we test whether listeners are capable of perceiving differences in boundary strength that are the result of a visual manipulation during sentence production.

The reason for including both a production and a perception study is that this allows us to gain insights not only into the acoustic properties of the speech stream, but to test how humans perceive these acoustic properties. The perception study allows us to tap into people's perceptions of boundary strength, in essence providing us with a very holistic measure incorporating the myriad possible acoustic correlates of boundary strength.

2. Production study

This experiment has two main aims: First, it provides acoustic information regarding if and how speakers use linguistic prosody to encode spatial information. Second, it provides a set of naturalistic spoken stimuli for use in the perception experiment.

2.1. Method, procedure

Participants (n=7) produced scripted utterances based upon images displayed on the computer monitor. Utterances were recorded digitally. Participants' eye-movements were also recorded for future analysis.

Participants were instructed as follows: "In this experiment, you will see three objects on the screen. Your task is to construct a sentence describing the path that a little brown mouse uses to navigate around the objects." The experiment contained 48 trials. At the start of each trial, participants saw three grey boxes with an arrow above or below each box (see Fig 1a). Participants were told that the arrows represented the path taken by "a little brown mouse." The arrows were all oriented either left-to-right or right-to-left, and were in an Under-Over-Under or Over-Under-Over configuration.

After two seconds, this display disappeared and was replaced with three easily recognizable objects of different colors (Fig 1b), with no path arrows. Participants then produced a scripted sentence based upon the images and the previously displayed arrows, e.g. "The little brown mouse ran under the red helmet, over the yellow basket, under the green shorts and into the mouse hole." (The beginning and end of the sentence, "The little brown mouse ran" and "… and into the mousehole" were constant across trials.) Thus, each sentence contains three similar prepositional phrases which are

naturally spoken as prosodic intonational or intermediate phrases, in addition to the utterance-final prepositional phrase "and into the mousehole."

The number of Under-Over-Under and Over-Under-Over sentences produced by each participant was balanced, as was the number of trials on which the imagined mouse ran from left to right and from right to left. In this paper, we focus only on those trials where the mouse ran from left to right (24 trials per participant). This was done to keep the duration of the perception study (see Section 4) to a manageable length for participants, while simultaneously keeping directionality constant for this paper's analysis.

Participants were familiarized with the names of the pictured objects in a training session before the start of the experiment, to ensure consistency in lexical choice.

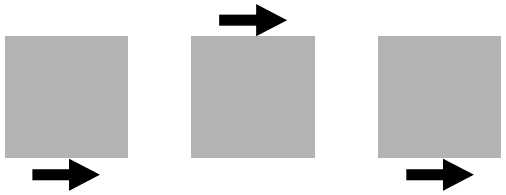

Figure 1a: *First display*

Figure 1b: *Second display*

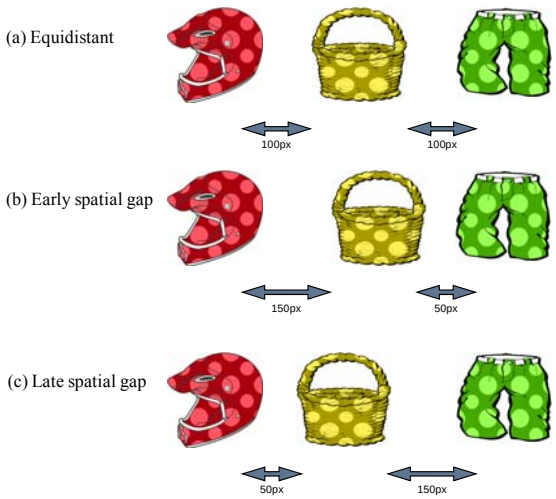

Figure 2: *The three spatial configurations*

2.2. Spatial manipulation

To investigate whether visuo-spatial information has an effect on prosodic boundary strength, we manipulated the distance between the three objects (see Figure 2). In the *Equidistant* condition, all three objects were equally spaced. There were 100 pixels between the first and the second object, and 100 pixels between the second and the third object. In the *Early Spatial Gap* condition, the first object and the second object were separated by a gap of 150 pixels, whereas the distance between the second and the third object was only 50 pixels. In other words, the second and the third object were grouped together. In the *Late Spatial Gap* condition, the first and second object were grouped together (separated by 50 pixels), whereas the second and the third object were separated by 150 pixels.

These three conditions can also be conceptualized in terms of *gap size*: In the Equidistant condition, we have two medium gaps, and in the Early Spatial Gap condition and the Late Spatial Gap condition we have one big gap and one small gap.

2.3. Predictions

We predicted that if visual gap size influences prosodic boundary strength, bigger spatial gaps should result in stronger boundaries. In particular, if we focus on pauses as an index of prosodic juncture strength, the prediction is that there will be more pauses and longer pauses when the spatial gap is bigger as compared to when the spatial gap is smaller.

By including three objects, we are able to test whether the first break (between the first and the second phrase) and the second break (between the second and the third phrase) pattern alike or not.

3. Results of Production Study

3.1. Analysis

Pause durations were extracted from the recordings. We focus here on the duration of the pause between the first and second prepositional phrases (Break 1) and the pause between the second and third prepositional phrases (Break 2), as indicated in this example: *The little brown mouse ran under the red helmet {Break 1} over the yellow basket {Break 2} under the green shorts and into the mouse hole*. Pause duration was measured from the offset of the preceding noun (e.g. 'basket') to the onset of the preposition (e.g. 'under').

Some of the utterances produced by participants contained disfluencies. All utterances where the disfluency affected one of the critical regions (Break 1, Break 2) were excluded from further analysis. We also excluded utterances where participants used an incorrect word (e.g. *sofa* instead of *couch*). These exclusions affected 10.2% of the total data. Breaks without any acoustic pause were also excluded; this affected 3.1% of the total data.

3.2. Pause duration

Figure 3 (on the next page) shows the average pause durations as a function of break order and spatial gap size (small, medium, big gap). The pause duration data was analyzed using an ANOVA with two factors: break location (Break 1 vs. Break 2) and spatial layout (equidistant, early gap, late gap). We found a marginal effect of break, with Break 2 being numerically longer than Break 1 ($F(1,6) = 3.7$, $p = .1$). However, there was no main effect of spatial layout and no break x layout interaction (p's > .2). Thus, although numerically spatially bigger gaps were associated with longer pauses, this effect did not reach significance.

4. Perception study

The analysis presented above for the Production Study focused on pause duration, commonly regarded as an indicator

of prosodic boundary strength. However, pause duration is only one of the numerous proposed indicators of boundary strength (see Section 1).

In order to tap into other potential indicators of prosodic boundary strength (to see whether they are sensitive to visuo-spatial distance), we decided to use humans as our 'measurement tool.' In other words, we tested whether *listeners' perceptions* of boundary strength in the absence of visual scene information are sensitive to the visual layout that the *speaker* saw when producing the utterance.

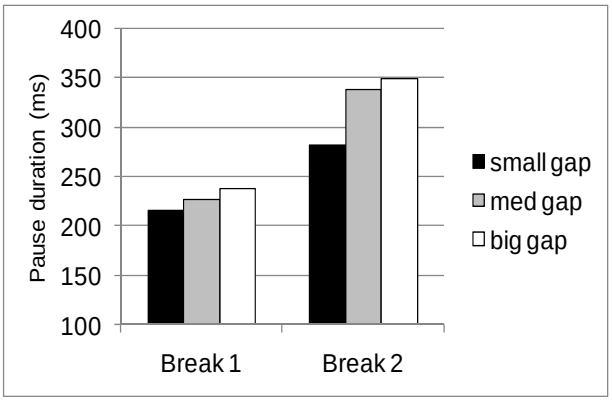

Figure 3: *Pause duration as a function of spatial gap size*

4.1. Method, procedure

In this study, the stimuli were the utterances obtained in the Production Study. Participants (n=28) listened to the utterances and provided ratings of prosodic boundary strength based on what they heard. A trial was structured as follows: First, participants saw the test word on the computer screen (e.g. 'helmet'). Then, a sentence was played over headphones (e.g. *The little brown mouse ran under the red helmet over the yellow basket under the green shorts and into the mouse hole*).

After hearing the target sentence (168 targets per participant), participants were instructed to indicate how strongly connected the test word ('helmet') is to the word that follows it ('over'). To do this, participants used the computer mouse to move a 'slider' bar along a linear scale which ranged from "weakest connection" to "strongest connection." Thus, they were able to freely choose which point on the scale best expressed the level of connectedness between the test word and the following word. This methodology is based on Krivokapić (2007), who showed that it is sensitive enough to obtain reliable data regarding prosodic boundary strength.

For a given sentence, a given participant was asked to rate either Break 1 or Break 2 (i.e., not both in the same sentence).

4.2. Predictions

The logic of the predictions is similar to that for the Production Study, except that now we are focusing on perceptions of connectedness rather than pause duration. If visuo-spatial information influences prosodic boundaries such that greater spatial distance results in a stronger boundary, we expect that participants' connectedness ratings will correlate with the spatial distance between the two objects that the speaker mentions. In particular, we expect that participants will give higher connectedness ratings to smaller gap sizes and lower connectedness ratings to bigger gap sizes.

5. Results of Perception Study

5.1. Analysis

For data analysis, the scale was divided into 100 equivalent proportions with 0 at one end of the scale, representing the *weakest connection*, and 100 at the other end of the scale, representing the *strongest connection*. Statistical analyses were conducted on these raw numbers, as well as on z-scores. The normalization procedure did not affect the results; the same data patterns were obtained with the raw scores and the z-scores. As in the Production Study, recordings with disfluencies that affected one or both of the critical breaks were excluded from analysis.

5.2. Correlations between pause duration and ratings

To ensure that participants' connectedness ratings are providing meaningful information about the nature of the prosodic boundaries, we looked at how well the ratings correlate with the acoustic measures of pause duration obtained in the Production Study. While pause duration is not the only factor related to prosodic boundary strength, it stands to reason that it contributes to perceived boundary strength. Thus, if participants' connectedness ratings are tapping into or providing a measure of prosodic boundary strength, we should see some correlation between our acoustic measure of pause duration and the perceptual judgments of connectedness.

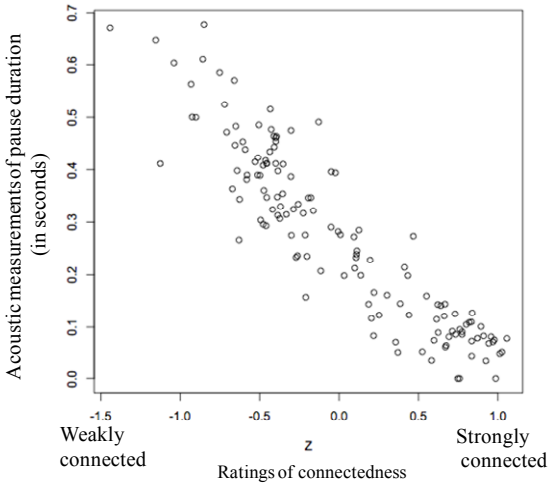

Figure 4: *Correlations between connectedness ratings and pause duration*

As shown in Figure 4, there is indeed a striking negative correlation ($r = -.68$) between pause duration and ratings of connectedness: the longer the pause, the less connected the two words are perceived to be. This confirms that participants' connectedness ratings are sensitive to subtle cues regarding prosodic boundaries.

5.3. Ratings and spatial layout

As shown in Figure 5 (next page), participants' connectedness ratings show a significant main effect of spatial gap size ($F(2,54) = 19.58$, $p < .001$). When Break 1 was produced on the basis of small spatial gap, it was judged to be significantly smaller (more connected) than when it was produced on the basis of a medium gap ($t(27) = -2.57$, $p < .05$) or a big gap

($t(27) = -2.31$, $p < .05$). Break 2 showed a similar significant effect of spatial gap size ($t(27) = -6.45$; $t(27) = -8.67$, p's < .001). In addition, as is expected on the basis of the Production Study, we found a significant effect of break order: Break 2 was judged to be overall bigger/less connected than Break 1 ($F(1,27) = 33.75$, $p < .001$).

In sum, listeners' ratings indicate that visual gap size has a significant effect on boundary strength. It is important to remember that listeners had no information about spatial gap size and did not know that speakers had produced the sentences on the basis of different visual displays. Thus, the sensitivity to spatial gap size in *listeners'* ratings must be attributed to acoustic differences in how *speakers* produced the sentences. Thus, we conclude that speakers' production of prosodic boundaries is sensitive to the spatial distance between the two objects being mentioned. Two objects that are spatially grouped together result in a smaller prosodic boundary than ungrouped objects.

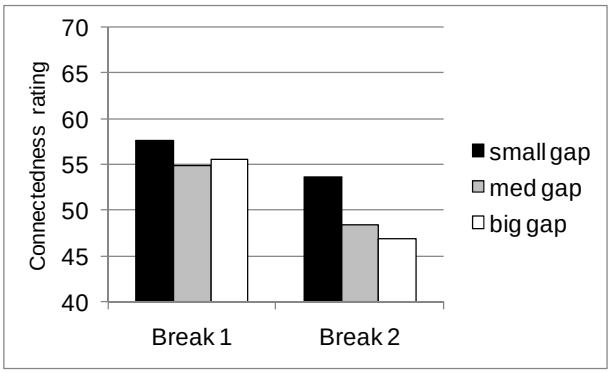

Figure 5: *Ratings of connectedness as a function of spatial gap size* (100 = 'strongest connection', 0 = 'weakest connection')

6. Conclusions

The research presented here suggests that in addition to linguistic factors such as prosodic structure and syntactic structure, prosodic boundary strength is also influenced by visual grouping or spatial gap size. We found effects of visual layout on a subtle property of language production: Listeners' rating data indicate that *visual scene layout* influences linguistic production at the level of prosodic boundaries.

These effects were clearest in the Perception Study – which provided a holistic measure of multiple acoustic correlates of boundary strength – and not as clearly present in the pause duration measurements, which fits with existing observations that pause duration is only one of many acoustic correlates of boundary strength. By using the fine-grained properties of the human perceptual system, we were able to detect effects of grouping in the visual domain on grouping in the linguistic domain – in particular, on the prosodic groupings that speakers produce.

7. Acknowledgements

This research was supported by NIH grant DC03172 awarded to Dani Byrd. We gratefully acknowledge useful feedback and comments from Louis Goldstein.

8. References

Ayers, G. M. (1994). Discourse functions of pitch range in spontaneous and read speech. Ohio State University Working Papers in Linguistics, 44, 1-49.

Byrd, D. & Saltzman, E. (1998) Intragestural dynamics of multiple phrasal boundaries. Journal of Phonetics, 26 ,173-199.

Byrd, D. & Saltzman, E. (2003) The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. Journal of Phonetics, 31, 2, 149-180.

Fletcher, J. (1987). Some micro and macro effects of tempo change on timing in French. Linguistics, 25, 951-967.

Gaitenby, J. H. (1965). The elastic word. Haskins Report SR-2, 3.1-3.12.

Klatt, D. (1975). Vowel lengthening is syntactically determined in connected discourse, Journal of Phonetics, 3, 129-140.

Krivokapic, J. 2007. The planning, production, and perception of prosodic structure. USC Dissertation.

Lee, E-K. & J. Cole (2006). Acoustic effects of prosodic boundary on vowels in American English. Proceedings of the Chicago Linguistic Society, Chicago, IL.

Oller, K. D. (1973). The effect of position in utterance on speech segment duration in English. Journal of the Acoustical Society of America, 54, 1235-1247.

Sanderman, A. A. & R. Collier (1995). Prosodic phrasing at the sentence level. In: Producing speech: Contemporary Issues. For Katherine Safford Harris. Edited by F. Bell-Berti & L. J. Raphael. New York: American Institute of Physics, pp. 321-332.

Selkirk, E. (1981). On prosodic structure and its relation to syntactic structure. In: Nordic Prosody II. Edited by T. Fretheim. Trondheim: Tapir, pp. 111-140.

Shattuck-Hufnagel, S. & A. Turk (1998). The domain of phrase-final lengthening in English. In: The Sound of the Future: A Global View of Acoustics in the 21st Century, Proceedings of the 16th International Congress on Acoustics and 135th Meeting Acoustical Society of America, 1235-1236.

Trouvain, J. & Grice, M. (1999). The effect of tempo on prosodic structure. In: Proceedings of the XIVth International Congress of Phonetic Sciences. San Francisco, CA, pp. 1067-1070.

Venditti, J. and Swerts, M. (1996). Prosodic cues to discourse structure in Japanese. In: Proceedings of the International Conference on Spoken Language Processing. Philadelphia, PA, pp. 725-728.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & P. J. Price (1992). Segmental durations in the vicinity of prosodic phrase boundaries. Journal of the Acoustical Society of America, 91, 1707-1717.