Variation in Final Lengthening as a Function of Topic Structure

Caroline L. Smith and Lisa A. Hogan

Department of Linguistics
University of New Mexico, Albuquerque
caroline@unm.edu

Abstract

This experiment shows that for an English speaker reading aloud, the topic structure of the text affects the amount of lengthening in sentence-final words. The speaker lengthened words less at the end of sentences that were followed by another sentence elaborating on the topic of the first, than at the end of sentences where the subsequent sentence added new information or switched topics. These results show that speech durations are affected by larger-scale linguistic organization, in addition to the well-known effects of local and phrasal structure. Modeling variation at the text or discourse level could improve the comprehensibility of longer passages of synthesized text.

1. Introduction

Present-day speech synthesis systems can model prosody at the level of the individual sentence remarkably well. Less attention has been given to modeling prosody over larger stretches of speech, even though research has identified a number of prosodic factors that human speakers modify as a function of higher-level discourse structure. These include the local and global structure of intonation contours ([1], [2], [3], [4], [5], [6]), speaking rate ([7], [8]), and the use of amplitude and pauses ([9], [10]). A role for topic structure in influencing the length of utterance units, as well as F0 patterns, was observed in Japanese [11]. Appropriate modeling of such factors would seem to be a fruitful tactic for increasing the naturalness and comprehensibility of synthesis [12], particularly in settings where the synthesizer speaks at some length.

The influence of discourse organization on intonation has been investigated more extensively than its influence on speech durations. In contrast, the durational effects of sentence-level prosodic phrasing are well-documented, and among the most robust in English is final lengthening. This has been described as added acoustic duration for speech segments ([13], [14]) and also as modification of the dynamic properties of speech articulations [15]. Lengthening was shown to mark 'paragraph' boundaries in speech ([9], [16]), but these studies were concerned only with the perceived degree of importance of a boundary, not the nature of the transition at that boundary. Since these two studies, there has been little, if any, investigation of how sentence-final lengthening is sensitive to the discourse structure of the material being spoken. The present study looks for a relation between the type of topic transition that occurs at a sentence boundary and the amount of lengthening at this boundary.

By taking a written text as the material to be analyzed, we could determine the topic structure independently of any speaker's rendition of that text when it was read aloud. In this study, we categorized the relations between the topics of consecutive sentences in the text as being indicative of the nature of the transition from one sentence to the next. Acoustic analysis of a speaker's reading of this text makes it possible to relate the durations of the sentence-final words to the nature of the topical transitions between sentences. Eventually, the patterns observed in the human speakers' text-reading will be developed into a durational model for synthesizing connected text.

2. Method

The goal was to determine how the nature of a topic boundary between sentences affects final lengthening. In order to determine the amount of lengthening in a particular production of a segment or word, the duration of that production must be compared to the duration of a non-lengthened production of the same material, in order to control for differences in the inherent duration of speech sounds. One technique to provide a baseline duration for estimating lengthening is to calculate the normalized duration of a speech sound over an entire large corpus and compare the mean normalized duration with its duration in sentence- (or phrase-) final position ([14], [17]). Even for a very large corpus, this method is sensitive to the particular contexts in which the segment occurs in that corpus [17]. The present study uses a different approach: we compared durations of words that occurred sentence-finally in a text with the same words produced in a controlled sentence-medial context.

2.1. Materials and recording technique

2.1.1. Text materials

The text chosen as the basis for this study is a passage from the manual for the computer drawing program Canvas [18]. This type of text was chosen because it has a clearly-defined sequence of topics, explaining how to use different features of the program. It was expected that this would simplify the task of labeling the topical organization of the text. The passage consisted of 60 sentences, with a total of 38 different words occurring sentence-finally. Lengthening was measured in these 60 sentence-final 'target' words. For each target word, a control sentence was constructed with the target word in sentence-medial position. All of the control sentences were of similar length, with 3 syllables before the target word and 8 syllables after. (Since the target words themselves varied between one and five syllables in length, the total duration of the control sentences varied accordingly.)

The control sentences were designed to favor the production of the target word with or without a pitch accent

depending on whether the presence or absence of accent seemed likely for that word in the text passage. In cases where the same word occurred sentence-finally in the text in both accented and unaccented contexts, two control sentences were constructed, one intended to favor production of an accent and one favoring an unaccented production. Two control sentences were also used for a few words where it seemed difficult to guess whether the speaker would accent the word. There were 11 words for which two control sentences were used, making a total of 49 control sentences in the experiment.

2.1.2. Recording technique

Five native speakers of American English were recorded reading these materials; data from one male speaker (E1) are reported here. Because this paper forms part of a larger study with the purpose of developing a model of naturally-occurring durational patterns for speech synthesis, the focus is on developing a detailed picture of the behavior of individuals, rather than averages over many speakers.

The recordings were made on a Sony Professional Walkman, using a Shure head-mounted microphone. Ten recordings were made of the same material, with intervals of 6 to 21 days (mean 10) between recording sessions for speaker E1. In addition to the materials discussed here, an additional text with matching control sentences was also recorded at the same time. At each recording session, speakers were presented with a different order of these four sections (two texts and two sets of control sentences) that alternated text and control sentences. The order of the control sentences was randomized across sessions, and two filler sentences were recorded at the beginning and end of each of the sets of control sentences. The ten recording sessions resulted in a total of 600 readings of sentences in the text reported on here, and 490 control sentences. Of these, 6 readings of text sentences and 3 of control sentences were omitted due to speaker error.

2.2. Analysis

The recordings were digitized on a Kay Elemetrics CSL system; all measurements were made with CSL. The acoustic duration of the target words was measured using the speech waveform and where necessary a spectrogram to identify acoustic landmarks. In addition, the duration of the coda of each final syllable was measured. In most cases even sentence-final voiceless stops could be measured, as the speaker tended to release these audibly. However, one reading of a text sentence and three control sentences were discarded due to difficulty in measuring them.

2.2.1. Pitch accents

All of the target words in the experiment except for “off” and “on” were lexical words, and thus potential sites for a pitch accent. (“Off” and “on” also receive pitch accent in some usages.) Words are generally longer when accented and have been shown to lengthen more phrase-finally when accented than unaccented [15]. In order to control for this factor, the target words in all readings of the text and the control sentences were categorized as accented or unaccented. The two experimenters listened to the recordings and decided, for each reading, whether the target word was produced with a pitch accent or not. Difficult cases were resolved by consensus, but for four readings of words in the text passage, no decision could be reached so they were excluded from further analysis.

The readings of the control sentences were divided into separate groups according to whether the target words were accented or unaccented. The mean durations of the words in control sentences, and their codas, were calculated separately for the two groups. These means, and their standard deviations, will be referred to as μ-s/acc, σ-s/acc and μ-s/unacc and σ-s/unacc for the accented and unaccented groups respectively. The readings of the text sentences were also sorted in each case as to whether the target word was accented or unaccented. Using a technique similar to that used by [14] and other studies, the duration of each target word in a text sentence was pseudo-normalized. If d-t is the duration of one reading of a target word in the text, then the pseudo-normalized duration was

$$\text{d-t/norm} = (\text{d-t} - \mu\text{-s}) / \sigma\text{–s} \quad (1)$$

where the mean and standard deviation used were those for the accented or unaccented productions, depending on whether the reading whose duration was being ‘normalized’ was accented or unaccented. In the rest of this paper, all references to duration refer to this pseudo-normalized duration. (Note that this is not the same as a *z*-score, because the mean and standard deviation being used are calculated over a sample distinct from the values being normalized.) For a total of 35 readings of words in text sentences, no readings of the same word with the same accent status were available in the control sentences, so these words were excluded from further analysis.

Preliminary examination of the data showed that the 18 readings of the word “two” (in two different sentences in the text) were quite anomalous. The grand mean of the durations of all target words in the text was 3.4, but the mean duration of the word “two” was 29.2; the word with the next highest mean duration was “three” at 12.9. After excluding all readings of the word “two” a total of 536 readings of target words in text sentences remained.

2.2.2. Topic boundaries

The written text was examined to determine the discourse relation between consecutive sentences. Adapting the topic labeling scheme used by [4] for spoken dialogue, each sentence in the text was labeled according to whether the following sentence constituted a Topic Shift (Shift), Topic Continuation (Cont), or Elaboration of the current topic (Elab). The categories were assigned with reference to the nature of the transition from one sentence to the next because it seemed more likely that the final word in the sentence would reflect the nature of the upcoming transition, rather than the topical transition from the preceding sentence. A fourth category, Text Marker (Mark), was assigned where the next “sentence” was an overt indicator of textual organization: these included the numbers in sequences of numbered instructions and the word “Note”. This meant that the sentence *before* “Note” was labeled Text Marker, whereas “Note” itself was labeled Topic Shift because the following sentence introduced new material. Table 1 shows the distribution of topic boundary types in the text.

Table 1: Count of tokens of each boundary type

Topic Boundary Type	Count
Topic Shift	69
Topic Continuation	251
Elaboration	149
Text Marker	57

2.2.3. Statistical analysis

Descriptive statistics were run in Statview with the pseudo-normalized durations of the target words and their final codas as dependent measures. Explanatory factors that were examined included measures of the phonological and discourse structure: topic boundary type, presence or absence of pitch accent on the target word, and presence or absence of stress on the final syllable. Additional factors were related to the physical layout of the text: did the target word occur at the end of a heading or at the end of a paragraph? did the sentence boundary coincide with a paragraph break or with a page break?

Topic Boundary Type was also analyzed as an independent factor in an ANOVA performed in SAS, although the cell n's are very unequal. A second independent factor was included in the ANOVA, the number of the recording session (1 - 10), to test if there were significant durational differences as the speaker became increasingly familiar with the text. The duration of the entire word and the duration of the coda were analyzed separately, although these are likely to be correlated.

3. Results

3.1. Lengthening of the entire word

3.1.1. Factors relating to linguistic structure

The type of topic boundary was found to be a significant factor in determining the (pseudo-normalized) duration of the sentence-final target words. In a two-way ANOVA, both Topic Boundary Type ($F(3,486)=11.97$, $p<.0001$) and Recording Session Number ($F(9,486)=2.16$, $p=.023$) were significant. The interaction between them was not significant ($F(27,486)=0.19$). Tukey's HSD post-hoc test showed that the duration of words followed by an Elaboration boundary were significantly shorter than those at any other type of boundary. This difference can be seen in Figure 1. Note that since these are pseudo-normalized durations, the vertical axis of the graph is dimensionless. A value of 2 corresponds to a duration that is 2 standard deviations longer than the mean duration of the matched words in the isolated sentences.

No consistent pattern of differences appeared in the durations at different recording sessions. Words recorded at the first session were significantly longer than those recorded at the 9th and 10th sessions, according to the post-hoc tests. No other differences were significant.

For the factors of Accent (words with or without pitch accent) and Final Syllable Stress (primary, secondary or reduced), the words were distributed so unequally among the categories that it was impossible to perform ANOVAS. Accented words tended to have slightly longer pseudo-normalized durations than unaccented (3.5 versus 2.9). Recall that since accented words were "normalized" relative to other accented words, and unaccented relative to unaccented, a difference here would mean that the accented words *lengthened* more, not that they were longer to start with.

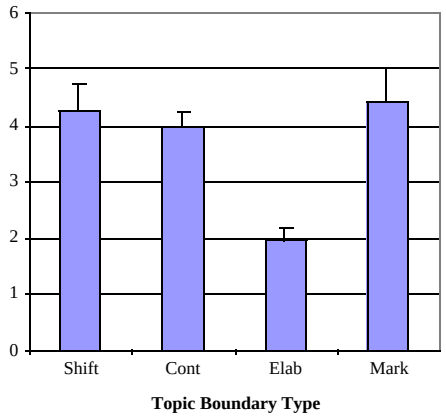

Figure 1. Pseudo-normalized duration of the words at different types of Topic Boundaries

3.1.2. Factors relating to physical layout of the text

In the labeling of topic boundary types, all headings in the text were classified as Topic Continuation, because the sentence after the heading provided information about the topic announced by the heading. Limiting the comparison of headings to non-headings that were also classified as Topic Continuation shows that the difference between them was not significant, with headings tending to be slightly more lengthened than non-headings. The position of a word at a paragraph end or immediately preceding a page break also showed no significant effect on the amount of lengthening. Overall, it seems that the physical layout of the text did not influence the amount of lengthening of the word as a whole.

3.2. Lengthening of the final coda

As for the entire word, lengthening of the coda was also significantly affected by Topic Boundary Type ($F(3,486)=19.49$ $p<.0001$). Recording Session was not significant ($F(9,486)=1.31$), nor was the interaction between these factors. As Figure 2 shows, the codas preceding an Elaboration type boundary were shorter than those preceding a Topic Shift or Continuation, the same as for the complete words. However, the codas preceding a Text Marker were also shorter than those preceding a Topic Shift or Continuation, contrary to what was found for the complete words. Post-hoc tests showed a significant difference between Elaboration and Topic Shift, and Elaboration and Topic Continuation, also between Text Marker and Topic Shift, and Text Marker and Topic Continuation, but no difference between Elaboration and Text Marker, or between Topic Shift and Continuation.

As was the case for the complete words, factors relating to the physical layout of the text appeared to have little effect on the lengthening of the codas. However, codas at the end of text headings showed a tendency to lengthen more than those at the end of ordinary sentences (headings: 5.1, non-headings: 3.5).

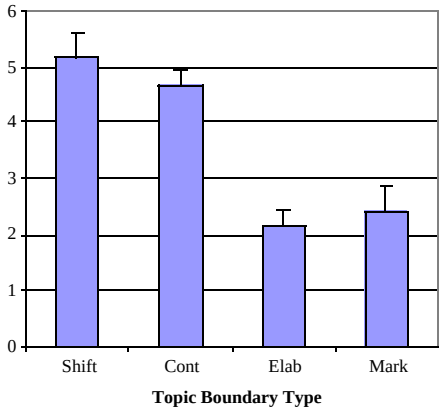

Figure 2. Pseudo-normalized duration of the final codas of words at different Topic Boundaries

4. Discussion

The type of Topic Boundary following a sentence-final word had a significant effect on the amount of final lengthening. When the following sentence provided more information about material already introduced in the current sentence, the speaker produced less final lengthening than when the following sentence continued a topic with new information, or switched to a new topic entirely.

The behavior of the words preceding the Text Markers is more difficult to interpret. For the overall duration of these words, there was substantial lengthening, suggestive of a more substantial boundary between the previous sentence and the Text Marker. But when the coda alone was measured, there was less lengthening before a Text Marker than before a topic shift or continuation. This difference between the word as a whole and the coda suggests that most lengthening occurred in earlier parts of the word. A possible explanation could be that the boundaries before these Text Markers were perceived by the speaker as substantial boundaries, leading to lengthening of the word as a whole, but the final syllable may have been more heavily reduced than in other contexts.

Previous studies [14] have suggested that most lengthening occurs in the coda of the word before a boundary. It is not clear that that was the case in this experiment. The pseudo-normalized durations are very similar for the codas and for the entire words, implying proportionately similar amounts of lengthening. This in turn would imply that in these words, other parts of the word lengthened about as much as the coda.

5. Conclusions

The results reported here show that at least for this speaker, the durational properties of his speech when reading a text aloud are affected by the organization of topics within the text. This study is the first phase of a larger study that will model the patterns, and the variability, of speakers' use of duration to signal topic structure when reading. Introducing this additional level of structure into the durational models of speech synthesis systems should improve their comprehensibility, by providing listeners with the same kind of information about the structure of the text that they are accustomed to receiving when listening to human readers.

6. References

[1] Brown, G., Currie, K. L. and Kenworthy, J., *Questions of Intonation*, University Park Press, Baltimore, 1980.

[2] Grosz, B. and Hirschberg, J., "Some intonational characteristics of discourse structure", *Proc. of the 2nd ICSLP*, Banff, pp. 429–432, 1992.

[3] Hirschberg, J. and Nakatani, C., "A prosodic analysis of discourse segments in direction-giving monologues", *Proc. of the 34th annual mtg. of the Assoc. for Comp. Ling.*, pp. 286–293, 1996.

[4] Nakajima, S. and Allen, J., "A study on prosody and discourse structure in cooperative dialogues", *Phonetica*, 50*:* 197–210, 1993.

[5] Swerts, M. and Geluykens, R., "Prosody as a marker of information flow in spoken discourse", *Language and Speech*, 37: 21–43, 1994.

[6] Wichmann, A. and House, J., "Discourse constraints on peak timing in English: experimental evidence", *Proc. of the XIVth ICPhS*, San Francisco, pp. 1765–1769, 1999.

[7] Brubaker, R., "Rate and pause characteristics of oral reading", *J. Psycholing. Res.*, 1: 141–147, 1972.

[8] Koopmans-van Beinum, F. and van Donzel, M., "Relationship between discourse structure and dynamic speech rate", *Proc. of ICSLP 96*, Philadelphia, pp. 1724–1727, 1996.

[9] Kreiman, J., "Perception of sentence and paragraph boundaries in natural conversation", *J. Phon.*, 10: 163–175, 1982.

[10] de Pijper, J. and Sanderman, A., "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues", *J. Acoust. Soc. Amer.*, 96: 2037–2047, 1994.

[11] Nakajima, S. and Tsukada, H., "Prosodic features of utterances in task-oriented dialogues", in Sagisaka, Y., Campbell, N. and and Higuchi, N. (eds.), *Computing prosody*, Springer-Verlag, New York, pp. 81–93, 1997.

[12] Sluijter, A. and Terken, J., "Beyond sentence prosody: paragraph intonation in Dutch", *Phonetica*, 50: 180–188, 1993.

[13] Klatt, D., "Vowel lengthening is syntactically determined in a connected discourse", *J. Phon.*, 3: 129–140, 1975.

[14] Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P., "Segmental durations in the vicinity of prosodic phrase boundaries", *J. Acoust. Soc. Amer.*, 92: 1707–1717, 1992.

[15] Edwards, J., Beckman, M. and Fletcher, J., "The articulatory kinematics of final lengthening", *J. Acoust. Soc. Amer.*, 89: 369–382, 1991.

[16] Lehiste, I., "Perception of sentence and paragraph boundaries", in Lindblom, B. and Öhman, S. (eds.), *Frontiers of speech communication research*, Academic Press, London, pp. 191–201, 1979.

[17] Campbell, N., "Syllable-based segmental duration", in Bailly, G., Benoit, C. and Sawallis, T. (eds.), *Talking machines: theories, models, and designs*, North-Holland, Amsterdam, pp. 211–224, 1990.

[18] Deneba Systems, Inc., *Canvas 5 User's Guide*, Deneba Software, Miami, 1997.

7. Acknowledgements

This work was supported by NSF grant BCS-9983106 to Caroline Smith. Many thanks to our speakers for their participation.