MODELING DURATIONAL VARIABILITY IN READING ALOUD A CONNECTED TEXT

Caroline L. Smith
Department of Linguistics
University of New Mexico, Albuquerque, NM 87131
caroline@unm.edu

ABSTRACT

One of the most striking features of speech produced by humans is its enormous variability, especially in the prosody. If speech synthesizers are ever to sound more natural, they must reproduce some of this variability. A portion of the variability can be attributed to known linguistic factors, but a substantial amount remains of unknown origin. Statistical techniques can be used to imitate this variability in a probabilistic way, but it may also be possible to reduce the proportion that is attributed to unknown factors. This study investigates durational variability in ten readings of an extended passage of text by an American English speaker. The focus is on how the structure of topics in the spoken material can explain some variability in the acoustic durations, and on how variability from this and other sources might be modeled in synthesized speech.

1. THE PROBLEM

A major deficiency in the prosody of synthesized speech has been the lack of variability, due principally to the use of deterministic algorithms for calculating intonation, pauses, and other prosodic characteristics [1]. The same prosody that sounds wonderful for one sentence can seem unnatural if it is repeated in all subsequent sentences with similar structure. Changes in prosody from sentence to sentence can also help the listener understand the structure of the spoken material [2, 3]. This paper presents a method for introducing variability in certain durational properties of a synthesized text. The properties studied here are sentence-final lengthening, number of syllables per second (a measure of speech rate) and pause duration.

Research on speech durations has typically focused on identifying the effects of factors relating to phonological context, speaker characteristics, etc. Correct implementation of the effects of these known factors is crucial to producing intelligible synthesis [4]. A major difficulty in introducing variability into synthesis is that most or all of these known trends must be implemented in order to preserve intelligibility.

This paper investigates a factor whose role in speech durations is relatively little studied, namely, the sequencing of topics in the material being spoken. Incorporation of this factor into synthesis, even in a deterministic way, may improve the intelligibility and naturalness, by more closely mimicking human speakers’ methods for marking the organization of the material being communicated. But in addition, it may be easier to synthesize variability in the effects of this factor than of many others, because the effects of topic sequencing operate over longer time spans than the local durational relations that must be correct in order to assure accurate segmental perception. These effects are investigated here in the reading aloud of several pages of text, a task which resembles what a synthesizer might be expected to do.

2. THE EXPERIMENT

The study investigated how speakers use durational patterns to mark the topic structure of a text being read aloud. Three durational properties were studied:

- Sentence-final lengthening, measured as the increase in duration of the same word in sentence-final compared to sentence-medial position.
- Speech rate, the number of syllables per second in the interpausal speech runs measured at the end of one sentence and at the beginning of the following sentence.
- Duration of the pause between sentences.

2.1. Procedure

Ten recordings were made of a male speaker of American English reading a collection of prepared material. The recordings were made at intervals ranging from 6 to 21 days, with a mean of 10 days. The present paper discusses the analysis of the readings of one text, which was a passage from the manual for the computer drawing program Canvas [5]. This type of text was chosen because its topic structure is relatively well-defined, and it is typical of what a synthesizer might be required to read in, for example, a Help system.

The structure of the text was labeled by categorizing the transitions from one sentence to the next as belonging to one of four types, adapting the labeling scheme used in [6]. These transition types were:

- Topic Shift. The following sentence introduces new material.
- Topic Continuation. The following sentence continues the topic, advancing the narrative.
- Elaboration. The following sentence provides more detail about the preceding sentence.
- Text Marker. The following sentence is an overt indicator of textual organization, similar to a discourse marker in speech. Sentences of this type were not analyzed separately but were included in the global analyses discussed in this paper.
Note that the label is associated with the sentence preceding the transition. The labeling was done by five linguistically-trained native speakers of English. Where the labelers disagreed, the choice of the majority prevailed.

In addition to the connected passage of text, the speaker also read a set of individual control sentences. Words which occurred sentence-finally in the text were placed in sentence-medial position in the control sentences. These were used as a baseline to measure the amount of sentence-final lengthening, which was measured as the increase in duration of the word in the text compared to its duration sentence-medially in a control sentence. This increase was calculated as the number of standard deviations above the mean in the sentence context, and is referred to as the “pseudo-normalized duration” of the sentence-final words (because it is analogous to the calculation of a z-score but with the mean and standard deviation taken from a different distribution).

Analyses of lengthening and speech rate excluded sentences with the final word “two”, as it was an extreme outlier. In the text it occurred twice, as a list number. Analyses of pause durations excluded pauses at page breaks.

2.2. Summary of major trends relating to topic structure

• Topic Shifts occur with significantly longer pauses than other types of transitions. Speech rate tends to be slower around a Topic Shift than at other transitions, but does not change at the transition from one sentence and the next.

• Sentence-final lengthening is similar in Topic Shifts and Continuations. However, at a Topic Continuation, speech rate increases significantly between the end of the first sentence and the beginning of the following sentence.

• Topic Elaborations have significantly less final lengthening than other transitions. Pauses are of similar duration for Elaborations and Continuations. Speech rate is faster in Elaborations, but does not change at the transition.

2.3. Distribution of durations

As a preliminary to modeling the speaker’s behavior for the four measures used in this experiment, the distributions of values were examined for each of the measures. When the data from all topic transition types were combined, the distributions for pause duration and for the pseudo-normalized durations of the final words were positively skewed, as is common for durational measures. For example, the distribution of pseudo-normalized durations plotted in Figure 1 is very similar to the histograms of durations in [7]. Application of a logarithmic transformation did not eliminate the positive skew of these data. In contrast, the distribution of measures of speech rate had a small negative skew, but these are not duration measures so they would not especially be expected to have positive skew.

However, in order to model both the differences among the topic transition types, the data must be divided into groups whose distributions will be modeled individually. When the data are divided in this way, the distributions are much further from normal. The smaller number of data points in each group is probably partly responsible for this, but it also appears that the labeling by topic transition type does not create groups which behave consistently with respect to these measures. For example, Figure 2 shows the distribution for the pseudo-normalized duration at Topic Shift transitions. While the mean in this context was 4.23, there were several negative values, which means that the word was shorter in the text than in the control sentence. Similarly irregular distributions were observed for pause durations and measures of speech rate.

The irregularity in the distributions could be due to other conditioning factors not controlled for in this analysis – differing sentence lengths, for example. Another possible explanation is that the topic labeling scheme is not fully appropriate for this task. Possibly the labelers did not apply the labels consistently, or the categories are too broad, so that subgroups which behave differently are being combined into one distribution, or the labeling scheme may not classify the transitions between sentences in a way that relates to their prosodic characteristics.

2.4. Category-internal variation: correlations among different measures within a topic transition type

In accounting for the variability in the durational prosody, one consideration is the relation among different measures of the same passage of speech. This study included multiple measures of closely-adjacent or overlapping spans of speech. In
particular, the measurement of rate at the end of a sentence included the final word of the sentence, whose lengthening was also measured separately. Examination of correlations among these measures may make it possible to identify dependencies among them, which would suggest a unified control of timing relations.

The relations among the four measures were tested using Spearman rank correlations. This nonparametric test was preferred to the more common Pearson correlation coefficient because of the non-normality of the data. The significance level of each of the Spearman correlations is listed in Table 1.

<table>
<thead>
<tr>
<th>measures</th>
<th>Shift</th>
<th>Cont</th>
<th>Elab</th>
</tr>
</thead>
<tbody>
<tr>
<td>length. final word and pause dur.</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>length. final word and speech rate end 1st sent.</td>
<td>ns</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>length. final word and speech rate beg. 2nd sent.</td>
<td>&lt;.05</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>pause dur. and speech rate end 1st sent.</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>pause dur. and speech rate beg. 2nd sent.</td>
<td>ns</td>
<td>&lt;.01</td>
<td>ns</td>
</tr>
<tr>
<td>speech rate end 1st sent. and speech rate beg. 2nd sent.</td>
<td>ns</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Table 1: Significance levels for Spearman correlations, data divided by topic transition type. ns = not significant.

The correlations between the amount of lengthening in the final word and the last speech run of the first sentence are what have been termed “part-whole correlations”. These would be expected to be significant unless there were a trade-off between the duration of the sentence-final word and the words preceding it in the final speech run. Thus the significant (positive) correlation found for Topic Continuations and Elaborations relates to the co-occurrence of a slower speaking rate over the last speech run with greater lengthening of the sentence-final word. It is more surprising that these two measures were not correlated at Topic Shifts. Those Topic Shifts which had more lengthening of the final word were not accompanied by a concomitant decrease (or increase) in overall speech rate.

On the other hand, at Topic Shifts there was a significant, but weak, tendency for greater amounts of lengthening of the sentence-final word to be followed by a more rapid speech rate at the beginning of the following sentence. There was no correlation between lengthening and speech rate in the following sentence for Topic Continuations and Elaborations. The only significant correlation involving pause duration was for Topic Continuations, where longer pauses tended to be followed by more rapid speech at the beginning of the following sentence.

Finally, speech rate at the end of the first sentence was correlated with the rate at the beginning of the following sentence for both Topic Continuations and Elaborations, but this correlation in fact reflects rather different patterns in the two types of transition. There is a significant increase in rate (p<.05 in a one-sample sign test) from the first sentence to the second at Topic Continuations, but no change at Topic Elaborations.

Although some of these correlations are statistically significant, none accounts for a very high proportion of the variation in the data. The highest correlation in this entire set is less than .5. The fairly large data set means that correlations can be statistically significant even if the trend is fairly weak. The overall conclusion from these tests is that the four measures are only weakly related, since even the part-whole correlations were not very strong. It would appear that each measure varies more or less independently.

### 2.5. Practice effects

As described in section 2.1, the ten recording sessions were separated by intervals averaging ten days. The purpose of this was to minimize the effects of repetition on the speaker’s behavior. Nonetheless, repeated readings inevitably result in greater familiarity with the material. Some effects of this can be identified in the data, but they do not suggest that the speaker modified his behavior greatly over the course of the experiment.

As might be expected, the measurements of speech rate show an increase in rate from the first reading to the last, although the increase is not completely consistent. Measured in the last speech run of the sentences, the rate increased from 5.05 syllables/second in the first reading to 5.87 syllables/second in the last reading, with a mean of 5.55 syllables/second. In the first speech run, the mean was 5.60 and the range was 5.10 to 5.78 syllables/second.

Interestingly, the pattern of increasing speech rate from the first repetition to the last (measured over speech runs between pauses) is not mirrored in the duration of pauses between sentences. One might have expected a trend towards shorter pauses as speech rate increases. Instead, as Figure 3 shows, there is generally a trend towards longer pauses in readings with faster speech rates, but the pause duration trends differently over different subsets of the readings.

Figure 3. Duration of pauses between sentences.

Unlike speech rate which trended upwards over the ten readings, pause duration increased from the first to the seventh reading, and then decreased between the seventh and tenth. This difference is a further piece of evidence that the different measures are independent of one another: the speaker did not speed up in all possible ways as he became more familiar with the text. Rather, he seems to have apportioned time differently.
As he came to know the text better, he read more rapidly but made longer breaks between sentences.

The amount of sentence-final lengthening showed a general, but inconsistent, trend towards less lengthening with increased familiarity. The pattern was somewhat similar to the pattern seen in the measurements of speech rate: the first repetition was the slowest and had the most lengthening, and the last reading was the fastest and had the least lengthening.

![Figure 4. Increase in duration in sentence-final position compared to sentence-medial position.](image)

3. **MODELING VARIABILITY**

The purpose of modeling a set of data is to reduce the number of degrees of freedom while still accurately representing the patterns in the data. In this study, the goal was a little different: to create a methodology that would make it possible for synthesized speech to mimic both the trends relating to the topic structure of a text and the variability observed.

From section 2 it is clear that the data were not normally distributed. In fact, tests suggested that they do not conform to any well-known distribution, such as the gamma distribution used to model speech segment durations in [7]. Therefore, the methodology that was chosen to model these data is bootstrapping, which makes no assumptions about the structure of the data [8]. Bootstrapping uses random sampling to calculate robust estimates of distribution parameters. This technique is often used for sampling the results of statistical tests. The drawback is that it does not derive a functional specification of the data, but that was not essential in this study.

The absence or weakness of the correlations among the different measures means that they cannot be modeled by a single distribution, so each of the four measures was treated as a separate data set. For each of these four data sets was further divided by topic transition type, resulting in a total of twelve distributions of data to be modeled. The number of values in each distribution varied from 56 to 274.

One hundred sets of ten values were randomly sampled (with replacement) from each distribution, using Matlab. Each of these sets of ten provides a set of values that can be used as a model of the variation in the larger distribution. The mean and standard deviation of the samples provide estimates of the mean and standard deviation without assuming that the data are normally distributed. By selecting as the model for each distribution the sample of ten whose mean is closest to the mean of all one hundred sample sets, the models for each topic transition type have means with the same relative magnitudes as the original distributions, thus preserving the central tendencies summarized in section 2.2 above.

Inspection of the distributions resulting from this procedure suggests that they do indeed preserve many of the characteristics of the original distributions, and are therefore suitable for modeling both the global patterns and the variability of the data.

4. **CONCLUSIONS**

The purpose of this paper was to examine some of the sources of variability that occur in the reading aloud of a connected text, focusing on how the sequencing of topics in the text contributes to the variation. The methodology chosen for modeling the data does not derive a function to model parameters of the data, which would perhaps be the most convenient representation. But by deriving a model without any assumptions as to the structure of the variation being modeled, we avoid the possibility of creating a model which alters the variation occurring in the data.

The next step is to use the modeled durations to create synthesized versions of the text which incorporate the same trends, and variability, that the speaker produced. The synthesized versions can then be presented to listeners for evaluation. Preferences emerging from this experimentation can guide efforts to introduce these kinds of variability into synthesized speech.

This research was supported by NSF grant BCS-9983106. Thanks to Lisa Hogan for making the acoustic measurements, and to the UNM Statistics Clinic for consultation.

5. **REFERENCES**