PHRASE-LEVEL FACTORS AFFECTING TIMING IN SPEECH

W. N. Campbell

ATR Interpreting Telephony Research Laboratories, Kyoto, Japan,

ABSTRACT

This paper reports findings related to local changes in speech rate at the level of the phrase. It examines the tendency for local resets in speaking rate to align with boundaries delimiting groups of related words, and presents evidence for an overall slowing down of speaking rate throughout these phrase-sized chunks of speech.

The durations of all segments in a phonetically-balanced two-hundred sentence database were normalised to facilitate comparison of the lengthening or shortening undergone by each, regardless of any differences in actual duration that may be attributable to differences in manner or place of articulation.

The normalised data was fitted by a regression line having a positive slope that was steepest for shorter sentences. When longer sentences were divided into their component phrases, the slope of the regression lines increased. This indicates that segment durations tend to increase as a function of time within a phrase and are reset at phrase boundaries within the larger prosodic unit.

1 Introduction

Underlying the work presented here is the idea that timing in speech can best be described from the higher levels of the phrase, foot and syllable, and that it is only finally realised at the level of the segment as a result of an interaction with the effects at these levels. To reduce the effect of any phone-specific differences, all segment durations used in this study have been normalised by scaling. Conversion to z-scores yields a measure of the relative lengthening undergone by each segment which can then be more simply compared with events at the higher levels of analysis.

Earlier work [2][3] has shown that the syllable is a significant unit in the prediction of duration, and that segment durations can to a large extent be predicted by a process of accommodation into a syllable-level timing framework. However, it is not yet clear how many levels of the prosodic hierarchy must be taken into consideration to reduce the remaining variance in the residuals of such a prediction. Nor is it yet clear what levels of the factors already known to have an influence are significant. In particular, Ladd & Campbell [4] have shown that considering the phrasing of an utterance as being composed of two levels, as predicted by the strict-layer hypothesis (Selkirk [6]), is inadequate to account for the phrase-final lengthening, which appears to be correlated well with varying degrees of boundary strength.

It will be shown here that phrase-final lengthening may in fact be better described in terms of phrase-level resetting because of a gradual increase found in the typical lengthening of each segment throughout the phrase, and because of a resetting, or speeding up, of the local articulation rate at each phrase boundary.

2 Levels of timing control

By filtering out differences in duration that are attributable to individual phones, we are implicitly accepting that the level of the phone is significant in timing control, and that the mean and variance of each phone are sufficient descriptors of their contribution at this level. Syntagmatic relations between adjacent phones, such as the lengthening of a vowel before a voiced consonant in the same syllable, are also known to have an effect but are not taken into account in this study. They will account for some of the variance in the results.

After normalising, we are left with a measure of the rate at which each segment has been uttered, quantified in terms of the two parameters, μ_i and σ_i for each phone-type$_i$. Changes in this rate have been shown to correlate with syllable-level factors such as stress, with foot-level factors such as rhythmic grouping, and with phrase-level factors such as tone-group boundaries. Paragraph-level effects are also anticipated, but current databases do not yet provide sufficient data for an analysis at this level.

Ladd & Campbell have shown that a simple two-level description of tone-groups does not account very well for the lengthening found in syllable measurements from a twenty-minute recording of a radio broadcast of a short story and show that evidence can be found for at least two finer levels of distinction, while arguing that there may indeed be *no* principled limit to the depth of prosodic structure. Wightman et al [7], after a survey of recent theories of prosoodic constituency, consider seven levels of boundary to be sufficient to reflect the degree of separation between words that occur at the edges of the different levels of the constituents. These seven levels will be adopted for the determination of phrase boundaries for this study.

The two-hundred phonetically-balanced SCRIBE sentences recorded and segmented by the CSTR at Edinburgh University were used as data. They were read by three speakers of RP English, two male and one female, but only the results for one speaker are reported here[1]. The sentences were designed to include examples of all naturally occurring combinations of phonemes in English and were read in isolation in a studio environment. It is not expected that they will be representative

[1] Visual comparison of the data for all three speakers supports the findings reported here but the readings cannot easily be matched for statistical comparison as the pronunciations vary considerably from speaker to speaker.

to appear in Proc Eurospeech '91

index:	0	1	2	3	4	5	6
median:	-0.39	-0.21	0	.015	.15	.58	.34
mean:	-.253	-.122	.113	.173	.464	.356	.804
sd:	.861	.919	1.001	1.027	1.149	1.103	1.179
n:	1298	3317	2108	132	88	15	851

of the natural prosody of spontaneous speech or even the careful prosody of a performed broadcast reading, but there is clear phrasing evident in all the readings.

In order to label potential word separation in the readings, each word in the orthographic text of the 200 SCRIBE sentences was manually assigned a boundary value between 0 (clitic) and 6 (sentence-final) according to the criteria in [7]. These values are approximate and not based on the actual performance of any one reader, but rather represent an idealised way of phrasing the sentences. The tags thus produced were then inserted into the speech files. Since two parallel datasets of continuous and isolated-word readings are available, word boundary information was carried across into the continuous-speech readings by first aligning orthographic words with data from the isolated word readings, then aligning the isolated word data with those of the continuous readings. Duration measurements from segments not matched in the process (about 5%) were excluded from further analysis.

Using the break indices as a key to phrasing, chunking was performed to include all words whose index was higher than that of the preceeding word, terminating when the index of the following word was lower. This method was preferred to one using a fixed level of break index because it enabled chunking of short sentences that contained no major phrase breaks, with only low-valued indices. It is justified linguistically in that it groups related words and separates in accordance with the syntax, as in the following example (sentence 4). Double bars have been inserted into the sentence to show where a boundary is found.

> The 0 smell 2 // of 1 the 0 freshly 1 ground 1 coffee 3 //
> never 1 fails 2 // to 0 entice 1 me 2 // into 1 the 0 shop 6 //

The clitics are assigned a value of zero and can therefore never end a phrase; the adjective with a value of 1 binds closely to the following noun. Each preposition is included with the phrase it introduces because of its low index value. Phrase-final prepositions would be indexed accordingly.

Analysis of variance of the normalised durations (z-scores) showed the differences in the means of segments in the words thus labelled to be significant at $p < 0.001$ ($F_{6,7802} = 65.91$). Table 2 shows a tendency for lengthening of segments in the word to increase as the separation index value of the word increased. Median and mean measures of lengthening for each value are presented.

A regression line fitted through the z-scores for each level of break shows a positive slope of 0.119 with an offset of -0.27, r = 0.21 (this correlation increases to 0.38 for coda segments considered separately) and confirms the relation between lengthening of segments and the index of the word (its separation from the following word). A line fitted to the raw segment durations has a slope of 5.48 with an offset of 57 ms (r = 0.28 overall, 0.41 for coda segments). Table 2 shows individual means for onset, peak, coda and medial (ambisyllabic) segments, confirming that this lengthening is not just confined to the word-final or phrase-final segments. The lack of lengthening and smaller variance of ambisyllabic segments is interesting but will not be discussed further here.

	0	1	2	3	4	5	6
onset:							
mean	-.238	.139	.337	.175	.242	.395	1.195
sd	.836	.978	1.001	.968	.928	1.149	.856
n	434	780	463	25	25	4	18
peak:							
mean	-.228	-.170	.195	.587	.487	.657	.401
sd	.859	.884	1.012	1.161	1.229	1.111	1.191
n	601	1231	733	45	27	6	302
coda:							
mean	-.226	.005	-.339	1.102	-.009	.301	1.743
sd	.927	1.026	.923	1.148	1.061	1.168	1.415
n	240	845	516	28	24	3	192
medial:							
mean:	-.228	-.273	-.045	-.159	-.243	-.018	-.029
sd:	.594	.679	.854	.856	.797	.098	.529
n:	23	461	396	34	12	2	175

3 Lengthening throughout the phrase

The tendency for segment lengthening to increase throughout the phrase can be observed on visual examination of plots of the z-scores for many of the sentences. Figure 2 shows one sentence as an example. The strong line shows the z-scores for each segment, the dashed line the strength of the break (relative scaling), and the upsteppping dotted line indicates the phrase grouping as determined from word-break indices.

Closer examination of the z-scores in relation to the phrases shows some fuzziness in the domain of the lengthening. For example, the /w/ of *went* is lengthened as much as the /@/ (schwa) in *Clara*. Similarly, there is a marked high later that includes the unstressed schwa of *a* as well as the phrase-final word *phase*. This is followed by a reset on *when* and a gradual increase to *served*, which seems to include the /h/ of *Hungarian*. It is difficult to determine whether the lengthening on the /ng/ in this word is due to a mis-segmentation (including part of the /uh/) or whether it is sustained over-long in the articulation. Listening to the reading confirms the existence of a boundary after *served*, but the lengthening effect appears to be more global that the segment string would predict.

It is anyway the more global effects that are of interest here, and the reset-and-rise over *Hungarian goulash* and again over *followed by rhubarb* are clear. The speeding-up of the last word is less easy to explain, but this effect if found in many of the sentences of this reader and may be a result of reading the sentences as a list. The final lengthening

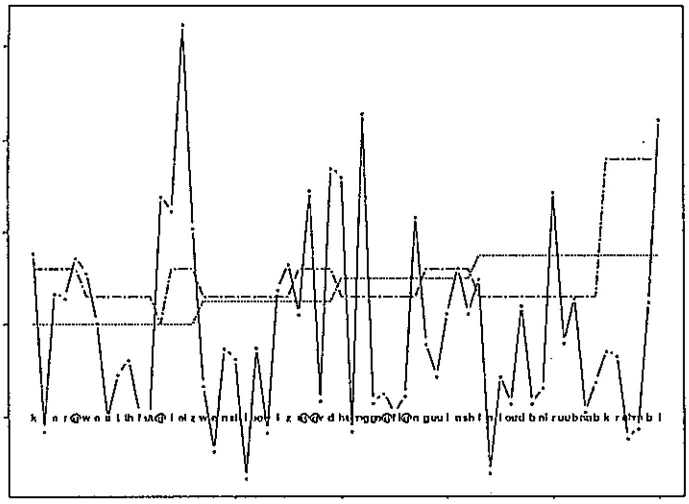

Figure 1: Sentence 63 showing z, breaks, and phrasing

of this word, however, is clear.

The positioning of the boundaries is reflected well in the lengthening characteristics, but some interesting uncertainties are also exemplified here. For example, *Clara* comes out as initial in the phrase *Clara went through a phase*, when it may function as a complete phrase in itself (an NP). The z-scores are similarly ambiguous, and as with the syntax, it is not clear whether to interpret the dip on *went through* as a reset or not.

4 Fitting a regression line

To obtain a more global perspective on these lengthening effects, robust regression lines were fitted to the data for the sentences and their component individual phrases. Robust regression [1] uses iteratively reweighted least squares to approximate the robust fit, with residuals from the current fit passed through a weighting function to give weights for the next iteration. It is less sensitive to the effect of outliers (such as the few phrase-final phones) and more likely to reflect the trend of the majority of tokens.

Sentence length varied from 22 to 70 phones, with the number of phrases determined per sentence between 2 and 7. Regression lines fitted to the normalised segment data for the 40 shortest sentences in the database (those segmented into less than 35 phones) showed a positive slope in 85% of cases (34 sentences out of 40). For the longer sentences (those segmented into more than 50 phones) a positive slope was found in 70% of cases (28 sentences out of 40). The average slope for the 40 shortest sentences was 0.021 representing an increase of half a standard deviation in the duration of the individual phones throughout a 25 phone sentence. The average slope for the 40 longest, however, was only 0.0019, giving an insignificant increase of less than a tenth of a standard deviation throughout a 50 phone sentence.

Since the average phrase count of the short sentences is 3.2 and of the long sentences 5.1, it may be more revealing to look instead at the slopes of lines fitted phrase by phrase. In this case we find an average slope of 0.061 for the 129 phrases in the shortest 40 sentences, which represents an increase of more than half a standard deviation throughout each phrase, of the same order as the sentence-level increase but within the smaller domain. In the longer sentences the slope also increases, to 0.03, representing an increase in durations of a third of a standard deviation during the span of a 10-phone phrase. It appears that the durations in all cases increase throughout the utterance, and the reason for lack of significance of the slope in the longer sentences is that a line is being fitted through several phrases, resulting in an average of them all rather than a fit to each.

Further evidence for this finding of lengthening throughout the phrase can be found in the difference in slope when comparing positive and negative slopes. Those in the positive direction in the shorter sentences have a mean of 0.13, more than twice that of those in the negative direction which have a slope of only -0.06. In the longer sentences similar differences are found, with slopes of 0.085 and -0.049.

A possible explanation for this difference that would lend strength to the interpretation of positive lengthening drift is that the phrasing determined by the text-based word breaks does not match the articulation of the utterance in some cases. Assuming a perfect match (however that may be defined), we can assume optimal regression lines if lengthening does increase throughout the phrase, but if there is a poor match then we can not expect the slope to be as steep. For randomly assigned phrase boundaries, we would not expect to see any slope at all for the same data, as exemplified by the results for the longer sentences as a whole.

Figure 4 shows the lengthening for sentence #19 *We have proof that the regime wields sufficient power in the North to exploit the entire population*. Here, we can see resets after *regime* (spreading across *wields*) and *in the North* (spreading across *to*). As before, the final word does not appear to be lengthened, but there is a tendency to rise across *exploit the entire*, and again over the final part of *population*[2]. The phrasing given by the word breaks places boundaries around *wields sufficient power* and *in the North*, but there appears to be a different grouping in the reading that would support a boundary after *sufficient* and group together *power in the North*. Such differences in phrasing are inevitable, but can only reduce the slope of the line fitted to the lengthening within the phrase. It can be assumed that optimal fits would produce steeper slopes.

5 Discussion

The ±1 standard deviation drift in the z-scores of the individual segment durations in the SCRIBE database appears not to be random but organised at the level of the phrase so that the minimum is at phrase-initial position and the maximum at phrase-final position. These positions do not coincide exactly with word boundaries, but appear to have a fuzzy match, the feature 'length' spreading over neighbouring phones as well.

Position of the segment in the phrase has been shown to be a significant factor in the timing of Japanese speech [5] and appears from the

[2] The speaker appears to have said /ih/ rather than /y/and/@/ in this syllable, but the unlikely reading 'popilation' appears to have been mistranscribed, resulting in unusually low z-scores for these two segments.

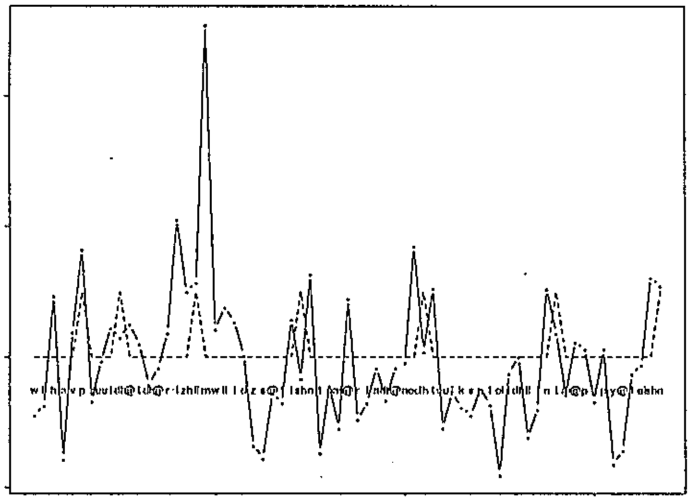

Figure 2: SCRIBE sentence # 19, showing peaks spotted by the algorithm

above results to be also significant in the timing of English. Determining phrase boundaries is a separate problem, but it has been shown that those derived from word indices do correlate well with the durational effects.

Many algorithms are being developed for optimising both stochastic and rule-based systems to predict the prosodic contours of speech, but these require large amounts of data to operate efficiently. Producing such data by hand is expensive and time-consuming work and does not allow easy re-transcription if an alternative labelling system is later found to be preferred. These results have an implication for the automatic segmentation of data that would also be of use in speech recognition systems. If we can utilise this resetting to locate the phrase boundaries of an utterance, we will be better able to produce 'islands of reliability' within which to focus lexical and semantic search.

Currently, an algorithm is being developed for this purpose that smooths the normalised measures of segment length (to reduce the effect of misaligned boundaries) and creates a boolean vector (true where phone$_{t+1}$ has a greater value than phone$_t$) to identify lengthening increases between successive pairs of values. This in turn is three-point smoothed to reduce the effect of local variation and produce a second vector that indicates rising vs falling sequences in sectors of the signal. Resets in this measure are taken as an indication of a prosodic event in the speech signal. Such an event, however, can indicate a phrase boundary or a stressed syllable. The dotted line in Figure 4 shows the points located by this algorithm for one sentence.

In all, 1300 such reset points are found in the SCRIBE sentences. The word-break index gives 794 phrase boundaries, of which 345 phrase-final words were found by the algorithm. This figure increases to 450 if the search space is expanded to include immediately neighbouring words (necessitated by the fuzzy domain of lengthening which does not always coincide with one single word). 640 stressed words were found out of 751 marked as stressed in the transcriptions.

Since there is in many cases a differential lengthening of onset and coda segments in the different cases of stressed and phrase-final lengthening [3], a modification to this algorithm is now being explored that will attempt to distinguish the two on the basis of lengthening differences within the syllables in the locality of the reset

6 Conclusion

Results have been presented which show that segment lengthening may increase as a function of position in the phrase, with resets at each phrase boundary. Seven levels of word separation were used in the determination of these boundaries, linking words whose separation value is low and positing a prosodic break where a sequence of high values is followed by a low value. These measures are derived by unrelated means, but show a correlation that supports their validity.

The data represent only one style of speech, and that a most artificial one, but the results are interesting enough to require further investigation with data from different speaking styles. If this phrase-level timing effect reflects a higher level of cognitive planning of the utterance, then the current models of duration in speech must be expanded to take account of more than just local segmental context.

Acknowledgement

Thanks are expressed to the labellers at CSTR who segmented the sentences on behalf of ATR, and to the management of ATR for enabling this research to be performed.

References

[1] Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988) The New S Language: A Programming Environment for Data Analysis and Graphics, AT&T Bell Laboratories, Wadsworth & Brooks/Cole Advanced books & Software, Pacific Grove California.

[2] Campbell, W. N. (1991) *Analog i/o nets for syllable timing,* in Speech Communication #9, Special Issue on Neural Nets and Speech, Elsevier Science Publishers B. V. (North Holland).

[3] Campbell, W. N.& Isard, S. D. (1991) *Segment durations in a syllable frame*, pp 37 - 47, Journal of Phonetics #19. 1990

[4] Ladd, D. R. & Campbell, W. N. (1991) *Theories of prosodic structure: evidence from syllable duration*, Proc XIIème Congres International des Sciences Phonetiques, Aix-en-Provence, France.

[5] Sagisaka, Y. & Tohkura, Y. (1984) *Phoneme duration control for speech synthesis by rule* (in Japanese), Trans IECE Vol 67-A No 7 pp.629-636.

[6] Selkirk, E. O., (1984) Phonology and Syntax: The relation between sound and structure. Cambridge, Mass. MIT Press.

[7] Wightman, C. W., Shattock-Hufnagel, S., Ostendorf, M. & Price, P. J. (1991) *Segmental Durations in the vicinity of prosodic phrase boundaries*, Submitted.