THEORIES OF PROSODIC STRUCTURE: EVIDENCE FROM SYLLABLE DURATION

†D. R. Ladd & ‡W. N. Campbell

†Department of Linguistics and Centre for Speech Technology Research, Edinburgh University, Edinburgh, Scotland.
‡ATR Interpreting Telephony Research Laboratories, Kyoto, Japan, and CSTR, Edinburgh University, Edinburgh, Scotland.

ABSTRACT

A recent theoretical proposal to enrich the traditional fixed hierarchy of prosodic domain types (foot, phrase, etc.) by allowing the possibility of "compound phrases", has been tested with a model of syllable duration for English. By marking the input text to identify both subordinate and superordinate major and minor tone-group boundaries, a finer specification of the durations of phrase-final syllables can be achieved. The new description explains significantly more of the error in the predictions for these syllables in the duration model.

1. INTRODUCTION

Rules for segment and syllable duration remain one of the least satisfactory aspects of most speech synthesis-by-rule systems. Empirical studies [3] have established many of the factors that affect duration, including both segmental differences (manner and place of articulation, vowel height, etc.) and prosodic factors such as degree of stress and position in phrase. However, current models still fall well short of accurately reproducing the timing of natural speech.

There is reason to believe that part of the difficulty in modelling duration stems from theoretical shortcomings in the identification of the prosodic factors involved. A number of current issues in phonological theory concern the nature of prosodic structure and the relationship among different prosodic features. Obviously, if the definition of e.g. 'phrase' is open to debate, then this will affect the way 'phrase boundaries' are marked in any given corpus or text sample, which in turn will affect any empirical findings about the influence of phrase boundaries on syllable and segment duration.

The study reported here is an attempt to assess whether such theoretical issues are of any practical significance for empirical models, and more specifically to evaluate Ladd's theoretical claim (Ladd [4][5]) that there is no principled limit to the depth of prosodic structure. We do this by comparing the durational effects of phrase boundaries *within* and *between* what Ladd calls 'compound' phrases, i.e. phrases that are themselves composed of two or more phrases. We report two kinds of results: first, whether there are significant differences on this comparison, and second, whether inclusion of the distinction makes possible a significant improvement in the amount of variance accounted for. If the answer to both questions is yes, it will illustrate the potential relevance of these issues in phonological theory for practical applications in phonetics and speech technology.

2.MATERIALS & PROCEDURES

2.1 Modelling syllable duration

Our starting point is the model of syllable duration reported by Campbell [1]. and its application to a sample text. 3959 syllable durations were measured from a twenty-minute passage of speech

recorded (with permission) from a BBC Radio broadcast of a short-story. The passage was prosodically annotated by two British-trained phoneticians to indicate stress, accent-type, and both major (maj-tg) and minor (min-tg) tone-group boundary locations[1]. Each syllable was then tagged with a number of identifiers (e.g. stressed syllable in one-syllable foot, final syllable in maj-tg, etc.) and a neural network was trained to predict the durations from the annotated input.

In this study we concentrate on the model's predictions of the effects of position in the phrase. Five categories of syllable are defined with regard to two types of prosodic phrase: initial in maj-tg (1), initial in min-tg (2), medial (3), final in min-tg (4), and final in maj-tg (5).

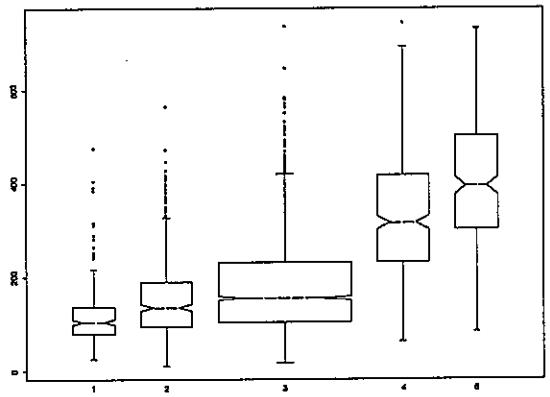


Figure 1: Durations of syllables factored by position in phrase.

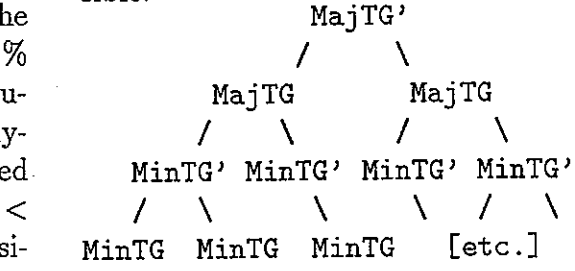
The boxes in Figure 1 are drawn with horizontal lines indicating the 25th, 50th and 75th percentiles of the durations of these syllables in milliseconds; the notches indicate significance at the 5% level in the difference of the distributions if they show no overlap. Analysis of variance of the durations factored in this way yields $F_{4,3954} = 529.4 (p < 0.001)$, showing that the effect of position is highly significant.

[1]Maj-tg corresponds roughly to Pierrehumbert and Beckman's 'intonational phrase' and min-tg to their 'intermediate phrase'.

However, the model's predictions, as suggested in the introduction, are only as good as the transcription on which they are based. The transcription is a traditional 'British school' analysis in which utterances are composed of major tone groups, and major tone groups are composed of minor tone groups. This categorisation of prosodic phrases conforms to the 'Strict Layer Hypothesis' (Selkirk [8]), according to which the prosodic structure of any utterance consists of a hierarchical arrangement of a fixed number of prosodic domain types. The Strict Layer Hypothesis is what is challenged in Ladd's work: specifically, Ladd has argued for the existence of 'superdomains' or 'compound prosodic domains', in which two or more adjacent domains of a given type are gathered together in a larger prosodic constituent *which is itself of that type.* Evidence for this proposal includes studies of acoustic cues to 'boundary strength' (e.g Cooper and Paccia-Cooper [2], Ladd [6]), in which considerable depth of structure is reflected in segmental duration and F0 properties in the vicinity of boundaries.

2.2. Refining the model with an enriched prosodic structure

With Ladd's proposal in mind, one of us (DRL) retranscribed the phrase boundaries in the corpus to allow for both compound maj-tgs and compound min-tgs. That is, we assumed that at least the following depth of structure is possible:



We thus have four hierarchically arranged types of tone group boundary rather than, as in the original traditional transcription, two.

It is important to note that there is no regular mapping from the traditional transcription onto the new one: the new one is based on a richer categorisation of the data. For example, many boundaries that were marked as min-tg boundaries in the original transcription became subordinate maj-tg boundaries in the new transcription, but so also did many of the original maj-tg boundaries. On the other hand, other boundaries marked as min-tg in the original transcrption were 'demoted' rather than 'promoted', becoming subordinate min-tg boundaries in the new transcription. In addition, there were many places at which no boundary was marked in the original transcription but where a subordinate min-tg boundary was marked in the retranscription.

Space does not permit any discussion of the kinds of phonetic cues that motivated the choice of boundary type in the retranscription; to some extent, as with the original transcription, choices were made on partly intuitive or impressionistic grounds. However, the point is not to argue in detail for one impressionistic transcription over another, but rather to show that, given a transcription that permits richer distinctions of boundary type, we can make more accurate predictions of syllable duration.

In the re-annotation, 174 new min-tg boundaries were inserted, 33 min-tg boundaries were demoted to subordinate, 190 min-tg boundaries were promoted to subordinate maj-tg boundaries, 37 original maj-tg boundaries were demoted to subordinate maj-tg boundaries, and 27 new subordinate maj-tg boundaries were inserted. 191 min-tg boundaries and 233 maj-tg boundaries remained unchanged.

3. RESULTS

Examination of the durations of the new tone-group-initial syllables (Figure 2) showed no significant difference between the sub-minor (4), minor (3), and sub-major (2) classes, although all three were significantly shorter than those in medial position (5), and major-initial syllables (1) were significantly shorter than any other group ($F_{4,3954} = 40.8$).

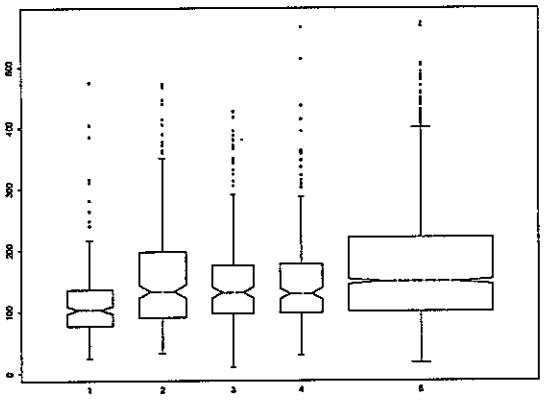


Figure 2: Durations of initial syllables.

Better separation is found in the lengthening of syllables in phrase-final position, i.e., those immediately preceeding a tone-group boundary, based on the re-annotation (Figure 3).

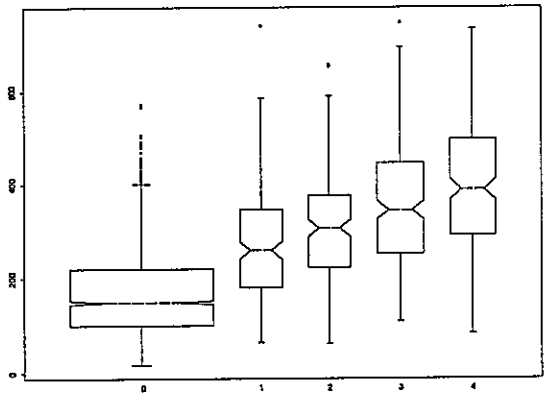


Figure 3: Durations of final syllables.

Clear and significant differences that correlate well with boundary strength can be found between those in subordinate min-tgs (1), superordinate min-tgs (2), subordinate maj-tgs (3), and superordinate maj-tgs (4). All are significantly longer than those in medial (0) position ($F_{4,3954} = 610.5$) and it can be concluded that the adoption of the finer classification provides better discrimination of the syllable durations.

3.1. Improving the prediction

In order to quantify the improvement that can be expected from incorporation of tone-group subordination in the synthesis model we can examine the residuals from a prediction. The best current prediction accounts for 86% of the variance in the durations, and by computing *predicted duration/observed duration* × 100, we have a percentage measure of the degree of fit for each syllable . Many factors contribute to the prediction error, and much of it may be randomly distributed. If a significant portion can be associated with any one factor, however, then retraining of the model with improved factorisation should account for that part of it.

The following table shows the percentage error that can be attributed to each class under both types of annotation.

Percentage error

error measure:	old	new
medial (unchanged)	3.7%	4.4%
minor -> sub minor	--	14.7%
medial -> sub minor	--	-5.3%
minor (unchanged)	5.17%	6.1%
medial -> sub major	--	-14.9%
minor -> sub major	--	2.7%
major (unchanged)	-3.4%	-2.4%
major -> sub major	--	-9.5%

For example, there was a 3.7% misprediction distributed among the syllables that had originally been classified as medial; those that are now classed as subordinate-major-tone-group-final syllabes form a small subgroup of that set which account for 14.9% of the error. By focussing the mispredictions in this way and retraining the network with data tagged according to the compound model, an improvement of up to 14% can be expected in the durations of that subgroup of syllables, which should significantly reduce the error in the group of medial syllables as a whole.

4. CONCLUSION

In this study we have compared two approaches to modelling position-in-phrase effects on syllable duration. The model previously employed defined such effects in terms of two levels of phrase, maj-tg and min-tg. This was replaced by a model distinguishing four levels of phrase (subordinate and superordinate groupings of phrases at both maj-tg and min-tg levels), to test the independently developed theoretical notion that there is actually *no* principled limit to the depth of prosodic structure. The second model gave a significantly better account of the distribution of syllable durations. This suggests that the notion of indefinite prosodic depth has merit and may be of practical empirical relevance.

Acknowledgements

We would like to thank CSTR in Edinburgh and ATR in Japan for their continued support for this research.

REFERENCES

[1] CAMPBELL, W. N.,(1991) *Analog i/o Nets for Syllable Timing*, in Speech Communication #9, Elsevier Science Publishers B. V. (North Holland).

[2] COOPER, W. & PACCIA-COOPER, J., (1980) Syntax and speech Harvard Univ. Press, Cambridge MA.

[3] KLATT, D. H., (1976) *Linguistic uses of segment duration in English* JASA #59 pp 1208 - 1221,

[4] LADD, D. R., (1986) *"Intonational phrasing: The case for recursive prosodic structure* Phonology Yearbook 3, 311 - 340.

[5] LADD, D R., (1988) *Declination "reset" and the hierarchical organization of utterances.* JASA #84: 530 - 544.

[6] LADD, D. R., (1990) *"Compound Prosodic Domains".* Occasional Paper, EULD; submitted to Language.

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