Prosodic Segmentation of Recorded Speech

W. N. Campbell

ATR Interpreting Telephony Research Laboratories
Seika-cho, Kyoto 619-02, Japan.

1 Introduction

Prosodic events in a speech signal, such as phrase boundary marking or stress on syllables, frequently correlate with an increase in the duration of the segments affected by the event. Campbell 1991 showed that an automatic algorithm can spot many of these events in a speech waveform from smoothed normalised measures of segment length, but the subsequent problem of distinguishing those segments that have been lengthened by stress from those lengthened by proximity to a boundary remains to be solved. As a step in thast direction, this paper will show that the phones in these two classes of environment show different characteristics of lengthening if viewed within the context of the syllable. It will be shown that the segment-based vowel-consonant distinction is less important than the syllable-based onset-coda distinction for an interpretation of lengthening characteristics.

2 Normalising phone-specific differences

Because different articulatory gestures produce sounds with different durational characteristics, some normalisation is required before a comparison of phone lengths can be made. Previous work (Campbell & Isard 1990) has shown that z-scores derived from the mean and standard deviation of each phone type can be satisfactory for this purpose, but the differences observed in the histograms of the durations of the phones of English can be better modelled by a bivariate Gamma distribution.

The bivariate gamma probability distribution function with parameters for shape and scale of a distribution can be defined as

$$\Gamma(x \mid p, s) = \frac{s^{-p} x^{p-1} e^{-\frac{x}{s}}}{\Gamma(p)} \tag{1}$$

where $\Gamma(p)$ is the gamma function, p is the shape parameter, and s is the scale parameter. Details of the maximum likelihood estimation of these parameters can be found in the Appendix. Figure 1 illustrates the degree of fit for eight English vowels, showing histograms of the raw durations alongside a smoothed representation of the histogram (dotted line) and the fit produced by the two parameters.

To remove the phone-specific durational differences from the measurements in a corpus of readings of 200 phonetically balanced sentences, individual duration measurements were

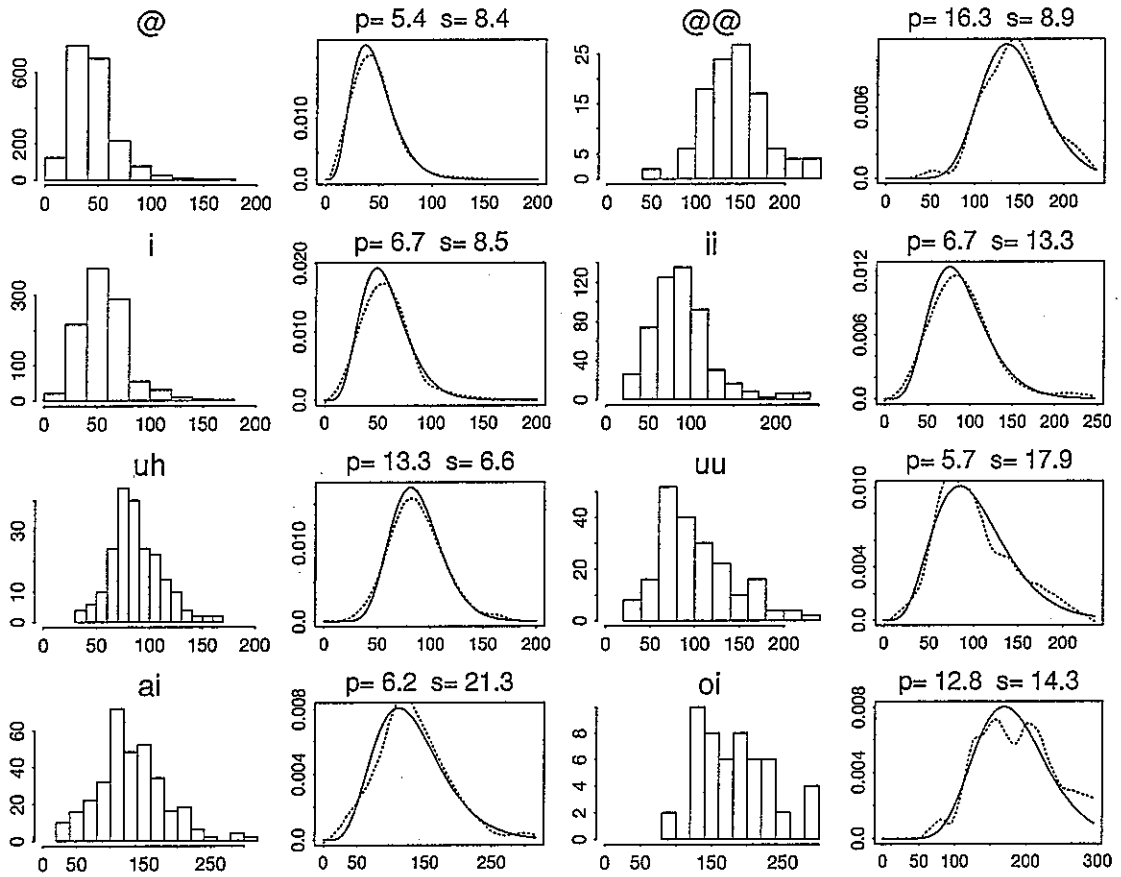

Figure 1: Gamma parameters derived for some representative English vowels

converted into quantiles (varying between 0 and 1) using a lookup table constructed using the above formula for the distributions determined for each phone. This transform allowed the length of each phone to be expressed as a real number, describing its position in the distribution of values for all phones of the same type in the corpus.

3 Lengthening within the syllable

The elasticity hypothesis (Campbell forthcoming) states that each segment in a syllable is lengthened equivalently, in terms of quantiles of its distribution, to accommodate to the duration determined for the syllable by its prosodic environment.

To test this theory of durational accommodation, known quantiles were compared with those predicted by the theory for a set of given syllable durations. First, syllable durations in the corpus were calculated, then an appropriate value (quantile) was determined for each so that the durations of the component phones, derived by lookup, would sum to the known syllable durations. A comparison was then performed between the quantiles predicted by this method and those computed from the actual segment durations. Systematic differences between the observed and predicted quantiles will show where other factors than accommodation to the syllable length are having an effect.

Since the greatest lengthening is found in phones in phrase-final or stressed syllables, the fit for these was examined first. A mean fit of 1.0 was observed, for observed/predicted, with 50% of the results between 0.81 and 1.15 (Figure 2). This confirms that most of the quantiles can be successfully predicted from the syllable durations, and a Chi-square of 0.378 (df 19) shows the distribution of error not to be significantly different from Normal.

Comparison of the fit for consonants and vowels shows no significant difference, but

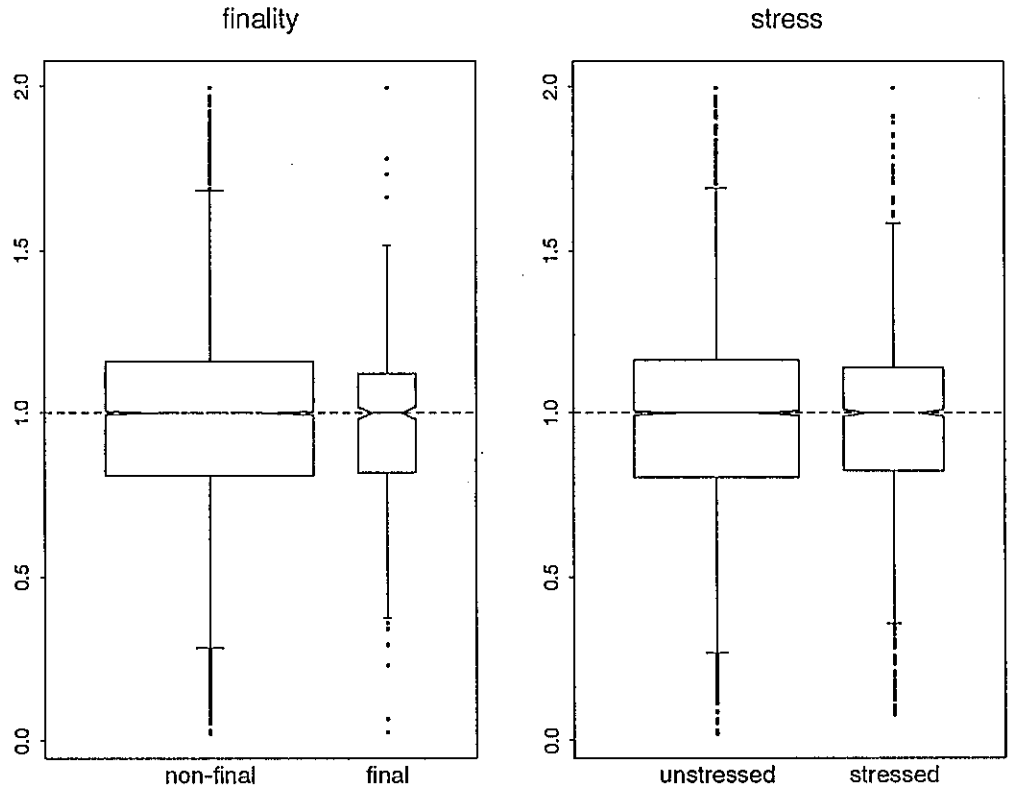

Figure 2: Comparing observed and predicted quantiles for stressed and phrase-final environments. (The figure shows observed/predicted values: The horizontal line shows a perfect fit; higher values indicate under-prediction of quantiles.

when we subcategorise the consonants into those that are in onset position and those that are in coda position in the syllable, a distinction becomes clear (Figure 3). Quantiles for onset consonants were typically greater than predicted, and those for coda ones less than predicted. This would imply that onset consonants are typically lengthened more than coda ones within the syllable. Consonants in ambisyllabic position were subcategorised separately, giving three classes in all.

Further analysis of phones subcategorised according to stressed and unstressed syllable contexts showed that although the unstressed group was similarly distributed about a fit of 1, the quartiles of the coda and medial consonants were being predicted less well in the stressed syllables (Figure 4¡). It appears that in a stressed syllable, the segments in the coda are not subjected to the same lengthening as those in the onset and peak.

In the comparable case of phones in phrase-final position, a much clearer separation could be seen between the onset and coda consonants (Figure 5). In this case though, the coda consonants were much longer than predicted by simple accommodation, and the onset and medial consonants were significantly shorter ('medial' consonants in final syllables are by definition in onset position). The lengthening undergone by phones in phrase-final syllables thus appears to be qualitatively different from that due to stress.

4 Discussion

These results are in accordance with those of Edwards & Beckman from articulatory data of jaw movements, and lead to the conclusion that an automatic algorithm for spotting prosodic events should be able to distinguish phones that are lengthened in phrase-final

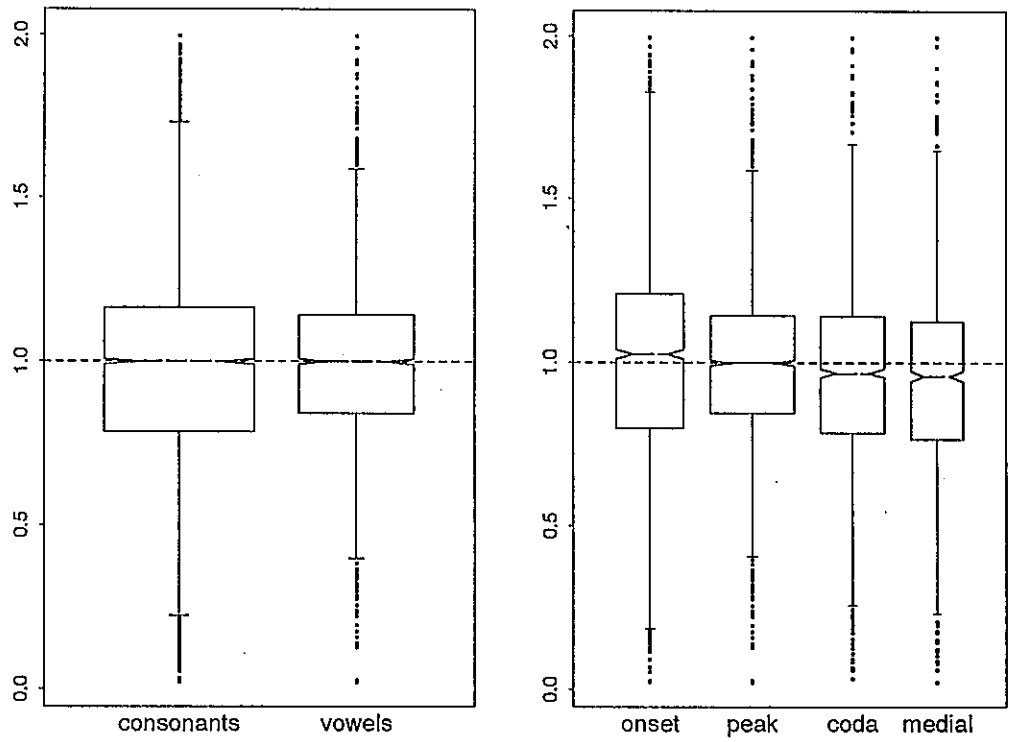

Figure 3: Fit for vowels and consonants showing the effect of position in the syllable

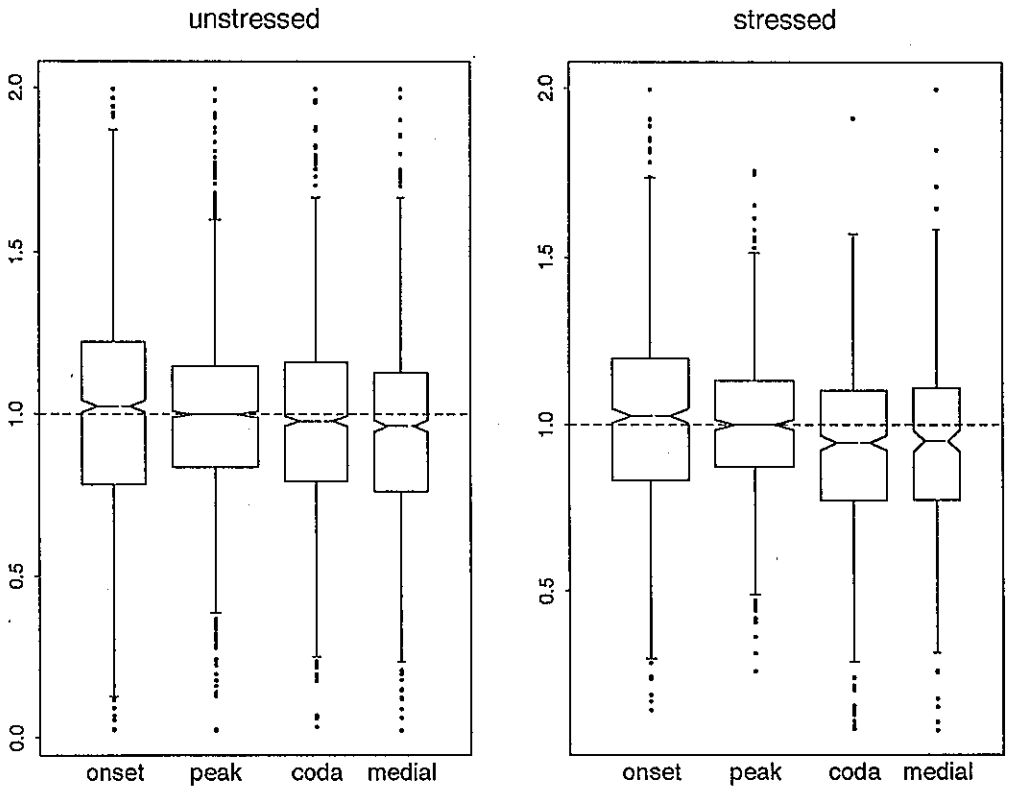

Figure 4: Fit for phones in stressed and unstressed environments

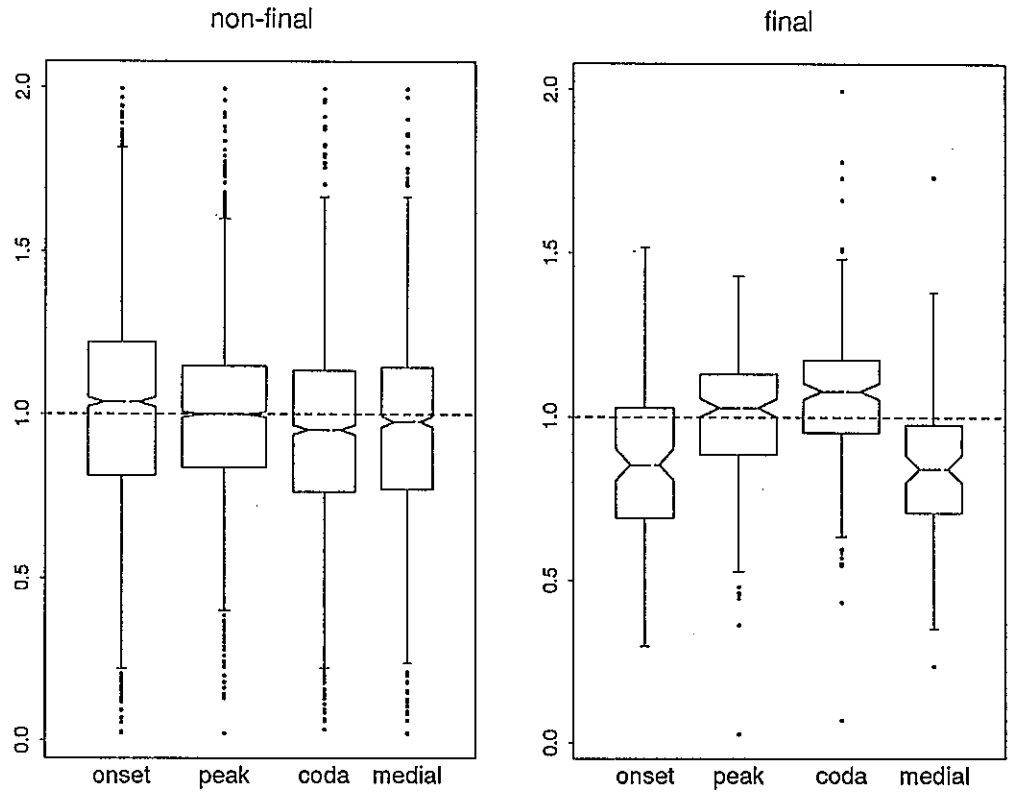

Figure 5: Fit for phones in neutral and phrase-final environments

position from those that are lengthened by stress from a consideration of the phone's position in the syllable. In this way, it should be possible to locate both phrase boundaries and their stressed syllables from duration measurements alone.

Studies are currently being carried out to determine the reliability of this type of prosodic segmentation. If successful, it will facilitate creation of the large amounts of training data needed for speech synthesis, and may provide a key to the use of prosodic information in speech recognition.

References

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

[2] Campbell, W. N. (1991) Phrase-level factors affecting timing in speech *Proc Eurospeech 91,* Genova, 629 - 632.

[3] Edwards J. & Beckman M. (1988) Articulatory timing and the prosodic interpretation of syllable duration, *Phonetica* **45**, 156 - 174.

Acknowledgement

I would like to express my thanks to the management at ATR for supporting this work, and to Shigeki Sagayama and Harald Singer for removing some of my ignorance of algebra.

Appendix

Maximum Likelihood Estimation of the Gamma Parameters:

$$\Gamma(x \mid p, s) = \frac{s^{-p} x^{p-1} e^{-\frac{x}{s}}}{\Gamma(p)} \tag{2}$$

Introduction of logarithms simplifies the calculations:

$$\log \Gamma(x \mid p, s) = -p \log s + (p-1) \log x - \frac{x}{s} - \log \Gamma(p) \tag{3}$$

To estimate the values of p and s for a given set of phone durations, we have to maximize the likelihood over all data points N.

$$\sum_{i=1}^{N} \log \Gamma(x_i \mid p, s) = -Np \log s + (p-1) \sum_{i=1}^{N} \log x_i - \frac{\sum_{i=1}^{N} x_i}{s} - N \log \Gamma(p) \tag{4}$$

Differentiating in respect to p and s leads to

$$\frac{\partial \left(\sum_{i=1}^{N} \log \Gamma(x_i \mid p, s) \right)}{\partial p} = -N \log s + \sum_{i=1}^{N} \log x_i - N \frac{\Gamma'(p)}{\Gamma(p)} \tag{5}$$

$$\frac{\partial \left(\sum_{i=1}^{N} \log \Gamma(x_i \mid p, s) \right)}{\partial s} = -\frac{Np}{s} + \frac{\sum_{i=1}^{N} x_i}{s^2} \tag{6}$$

By setting the derivatives to zero and dividing by N we get

$$-\log s + \frac{\sum_{i=1}^{N} \log x_i}{N} - \frac{\Gamma'(p)}{\Gamma(p)} = 0 \tag{7}$$

$$-ps + \frac{\sum_{i=1}^{N} x_i}{N} = 0 \tag{8}$$

or by taking the logarithm

$$\log p + \log s = \log \frac{\sum_{i=1}^{N} x_i}{N} \tag{9}$$

By substituting $\log s$ from (9) into (7) and rearranging, we get

$$\log p - \frac{\Gamma'(p)}{\Gamma(p)} = \log \frac{\sum_{i=1}^{N} x_i}{N} - \frac{\sum_{i=1}^{N} \log x_i}{N} \tag{10}$$

which can finally be rewritten as

$$\log p - \frac{\Gamma'(p)}{\Gamma(p)} = \log \frac{\sum_{i=1}^{N} x_i}{N} - \log \left(\prod_{i=1}^{N} x_i \right)^{1/N} \tag{11}$$

The right side of this equation consists in the difference of the logarithms of arithmetic mean and geometric mean; the left side can be tabulated for $0.1 < p < 40$. After calculating p, s can be determined by $s = \sum_{i=1}^{N} x_i / (Np)$.