1—6—13

Foot-level timing control for English speech systems

W. N. Campbell
ATR Interpreting Telephony Research Labs

1 Abstract

A two-level timing control system is described, showing that durations can be determined at the level of the syllable and accommodated at the segment level.

2 Introduction

Campbell & Isard (1991) discussed the notion of elasticity of speech segments and described a theory of accommodation into a durational framework determined at the level of the syllable. Campbell (1991) described a computer programme that uses a neural net to predict the syllable-level durations on the basis of input from a set of feature-descriptor units that describe the pragmatic and structural context of the syllable with only minimal awareness of its phonetic constituents. This paper will examine the extent to which segment durations predicted in this way differ from those observed in a reading of 100 phonetically balanced English sentences.

3 Syllable-level

The number of factors required to sufficiently describe a syllable's duration has yet to be determined, but using a three-layer neural net with input units sensitive to a) part-of-speech, b) stress, c) position in the phrase, d) position in the foot, e) number of phonemes in onset and rime, and f) the nature of the syllabic peak, we are able to account for 76% of the variance observed in the durations of the 1564 syllables in the first hundred of the SCRIBE phonetically balanced sentences with a RMS error of 42 ms[1]. The mean duration of these syllables read by speaker GSW is 191 ms (SD = 115 ms, median = 165 ms). The neural net was trained on 1000 syllable durations measured from a radio broadcast of a British (female) speaker reading a short story.

This is a conservative estimate of the power of the system as the input description was determined as a result of parses from earlier modules in the MITalk system (version 2, 1983) which serve as a front-end processor to the duration component currently being developed. Stress determination is only three-valued, largely lexically determined, and parses are restricted to noun-phrase and sentence-level discrimination. The pre-processing modules need to be improved or their output manually corrected for a full test of the algorithm.

The two main differences between the system in current use and MITalk's duration component are a) that syllable-level duration determination is insensitive to the phonemic content of the syllables, and b) that the model incorporates information concerning the syllable's position in the foot. These differences are motivated by a desire to separate the articulatory component of the speech signal from the cognitive component in order to test the theory that whereas phonemic differences must place constraints on the duration of an utterance, the time course is determined more as a result of factors operating at the higher level, such as rate of speech, and semantic and discoursal features such as stress and phrasing etc. This implies that segmental durations must be determinable from knowledge of the syllable duration and probability density distribution statistics of the individual phonemes.

4 Segment-level

In order to test the theory that segmental durations accommodate into a framework determined at the level of the syllable under a principle of elasticity, statistics were gathered and segmental durations predicted from knowledge of their distribution densities and the parent syllable durations alone.

The elasticity principle in its strongest form says that all segments in a given syllable fall at the same place in their respective distributions. Consider as an example the phonemes /a/ (μ_a = 119 ms, σ_a = 37ms), /t/ (μ_t = 41 ms, $\sigma_t = 21ms$), and /d/ (μ_d = 39 ms, $\sigma_d = 19ms$). If we are to combine these into the words *at* and *ad* then, by simply summing the means, the default durations for the words can be expected to be 160 ms and 158 ms respectively. Let us consider the case where the actual durations of the words is in fact longer, say 300 ms, and assume that onset and rime lengthen equivalently by a uniform amount in terms of their standard deviations. In this case a value can be found for k such that $\mu_a + \mu_t + k(\sigma_a + \sigma_t) = 300$, which yields /a/ = 208 ms and /t/ = 92 ms. A slightly different value of k gives $\mu_a + \mu_d + k(\sigma_a + \sigma_d) = 300$; where /a/ is 213 ms and /d/ is 87 ms. The overall word lengths are the same, but because the /d/ is shorter than the /t/, and has a smaller variance, the vowel seems to

[1]Unmodified MITalk durations predict 68% of this variance with an RMS error of 51 ms at the syllable level but the comparison may not be strictly fair as these durations are for American English and GSW is a British RP speaker.

英語音声合成システムのため階層的なタイミング制御
ウィルヘルム N. キャンベル

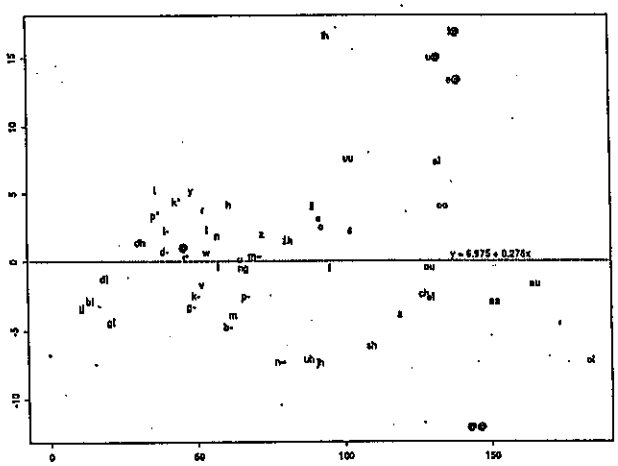

Figure 1: residuals of standard deviations against raw means

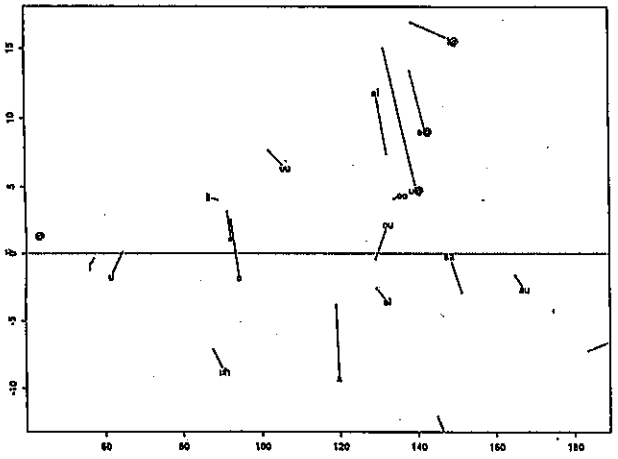

Figure 2: movement of means/sds through prediction - vowels

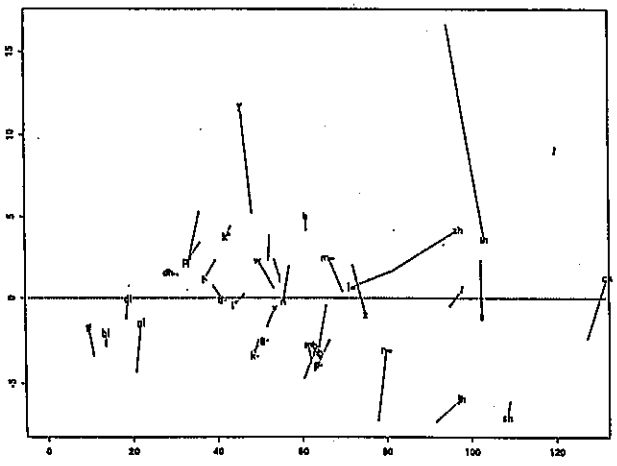

Figure 3: movement of means/sds through prediction - consonants

lengthen by 5 ms to accommodate. The lengthening applied to each segment in the syllable is the same for each phoneme, in terms of their elasticity, in each case. This appears to account quite simply, though not completely, for the lengthening that has been observed in vowels before voiced consonants.

Segment durations for each syllable in the first hundred SCRIBE sentences were calculated by feeding the observed syllable durations and the phoneme string to the *seg-dur* component of the duration algorithm. The resulting durations were compared with those observed in the readings of GSW and differences plotted graphically. As any reading by a human speaker is bound to show considerable variance in absolute durations, we will only consider trends here and not examine specific details.

Figure 1 shows residuals of standard deviations plotted against the means for the phonemes as segmented from the raw speech signal ($\sigma = 6.97 + 0.27\mu$). The notation used is the Edinburgh University Machine-Readable Phonetic Alphabet. Figure 2 shows the differences for the vowel durations calculated by the *seg-dur* component, and Figure 3, differences for consonant durations. Values in the plots are in milliseconds. Closer examination of Figure 1 shows related phonemes clustered together in the two-dimensional space. The lines in Figures 2 and 3 indicate change from the positions indicated in Figure 1. That the main orientation of these lines is in the vertical plane, indicates that very little change occurs in the mean durations as a result of the prediction process when vowels and consonants are lengthened equivalently. However, the extent of these lines in the vertical plane indicates a change in the variance of the phonemes. This variability is apparently balanced with respect to direction in the case of consonants, but represents a reduction in the case of most vowels. The implication of these changes is currently under investigation.

5 Summary

A two-level duration model has been described, and details of fit at both the syllable level and the segmental level have been presented. It has been shown that with equivalent lengthening, mean segmental durations are preserved through the accommodation process, but that some loss of variance is suffered, particularly in the case of vowels.

References

W. N. Campbell & S. D. Isard (1991) *Segment durations in a syllable frame*, Journal of Phonetics #19.

W. N. Campbell (1991) *Higher-level timing-control for an English language speech-synthesis system*, 信学技報, Vol 90, #423 SP-90-77.