Syllable-based segmental duration

W. N. Campbell

ATR Interpreting Telephony Research Labs, Soraku-gun, Seika-cho, Kyoto 619-02, Japan.

Abstract

A two-layered model of timing in speech is described. In this model, syllable durations are first calculated at the higher level, to reflect the rhythmic and structural organisation of the utterance, and then segment durations are calculated at a secondary stage of the process, fitted to the higher-level framework by accommodation of the components of each syllable according to the probability density distributions observed for each phoneme in a phonetically balanced corpus.

A neural net was trained to predict the syllable timing based on exposure to a large corpus of feature – duration pairs from measurements of spontaneously occurring natural speech. In order to account for both the segment-specific characteristics at the phoneme level and the rhythmic and phrasal characteristics at the syllable level, two different corpora of speech were used in the training of the model. The first was taken from a twenty-minute radio broadcast of a short story, the second from readings of 200 phonetically balanced sentences.

To test whether a measure of lengthening can be found which applies to segments uniformly within the syllable, instead of differentially with respect to phonemic type or position in the syllable, the durations of each token in the segment database were converted to standard normal form by subtracting the means and expressing the residuals in terms of their standard deviations to yield a mean of zero with unit variance for each phoneme distribution. These normalised values were taken to represent the amount of lengthening or compression undergone by each segment relative to its elasticity.

To show that such a process of segmental accommodation is feasible, results are presented which confirm that segments in sentence-final position undergo greater lengthening in the peak and coda than in the onset, whereas segments that are lengthened sentence-internally, for stress and rhythmic reasons, are lengthened more uniformly throughout the syllable. The timing of segments in short syllables is similarly found to be reduced without differentiation with respect to position in the syllable or phonemic type. This greatly simplifies the process of accommodation to the higher-level framework.

1. INTRODUCTION

As much of the variance in the duration of syllables in a large corpus of read speech

can be accounted for by an algorithm determining durations from the level of the syllable as by an algorithm based on segment durations. The advantage of syllable-level duration determination is that it facilitates associations with even higher-level features such as rhythmicality at the level of the foot. It does, however, leave open the problem of how durations determined for each syllable are to be shared out amongst the components at the level of the segment.

Crystal and House (1988) in an analysis of a continuous-speech segment duration database showed that both consonants and vowels have similar positively skewed distribution densities. While they did not specifically make this point in their paper, it is clear from their data that the main difference between the distributions of the different phoneme classes lies in the maximum durations they could attain - in the positive end of the tail - not in the modal or most common durations, which for all segment types appeared to be around 50 ms. It would be extreme to infer from this that all segments can be assigned exactly the same duration in the majority of cases, regardless of type or context, but it may be an indication of smaller variability in the duration of segments in fluent speech than is found in the less typical modes such as readings of isolated words and citation form sentences, which make up many of the databases on which earlier studies of duration have been based. For a computer text-to-speech system, which is not yet likely to be required to express emotion, the need is more likely to be for the fluent variety of speech rather than the extremes.

Tests were performed to determine whether segments can be treated uniformly to fit a timing framework governed by the syllable. They show that, in the majority of cases, a single constant can be found that quantifies the expansion or compression undergone by all segments within a syllable.

1.1. Models of speech timing

Four different models of predicting speech timing for synthesis applications have been differentiated by van Santen (1990): 1) sequential rule systems, 2) equation systems, 3) table-lookup systems, and 4) binary trees. All these models predict the durations of phoneme-sized segments directly at the lowest level, either by the formulaic or rule-based modification of a set of default durations, or from direct specification of the typical durations of each (typically phoneme-sized) segment for each type of context.

The duration-assignment algorithm formulated by Klatt (1979) as part of MITalk and later incorporated in DECtalk and other synthesis systems shortens or lengthens the default duration for each phoneme according to an ordered set of about a dozen rules which make phoneme-specific adjustments to the inherent duration of each phone, either multiplicative or additive, according to phonetic and phrasal contexts. The concept of a minimum duration specified for each phoneme is required in such a system to ensure that shortening is never too extreme, although there is no explicit upper limit to the lengthening each may undergo. Rhythmic effects are assumed to be a by-product of segment-level modifications (Carlson, Granström & Klatt, 1978), such as the shortening of consonants in clusters, and no explicit control of these is offered per se.

A different type of timing algorithm, free of phonemic generalisations, is offered by models of gestural mechanics (Browman & Goldstein, 1989). By assuming the parameters of a task-dynamic model of articulatory gestures as the basics of phonetic

specification, Edwards & Beckman (1988) are able to distinguish the characteristics of lengthening applied as a result of phrase-final position from that applied as a result of stress phrase-internally. Unfortunately, the amount of data available for such movements is still very limited, and does not yet include continuous or spontaneous speech measurements. Until this is available, attempts will continue to be made to model the effects of these processes from measurements of acoustic rather than articulatory data.

The model described in this paper assumes timing to be a result of interactions between processes operating at different levels in the speech hierarchy (see for example, Beckman, 1991), and first calculates syllable durations to account for the higher-level prosodic processes, then accommodates segment durations into the timing framework thus determined for the utterance, according to known distribution densities. The two principal assumptions of this model are as follows

1. Acoustic measures of speech timing events are a function of interactions between
   (a) higher-domain phrasal and rhythmic processes operating at the level of the syllable and above, and
   (b) segmental characteristics, related to the manner, place, and context of articulation of each sound in the speech sequence.
2. Given the higher-level framework, segmental durations can be determined by accommodation according to known probability distribution functions for each phoneme.

In order to incorporate control of rhythmical features, it is assumed here that duration is a function of the syllable as a unit, and can be calculated accordingly, without prior consideration of fine phonetic constraints. The influence of the actual phoneme string at this stage is summarised by a count of the number of phonemes in each syllable, and a classification of the type of syllabic peak, sufficient to differentiate e.g., a schwa from a diphthong. Segment-specific timing effects are taken into account at a later stage of processing when a compromise is reached betwen the rhythmic and phrase-level framework and the articulatory features of the individual segments.

The distribution of syllable durations has been found to closely approximate a normal curve when transformed into the log domain if separated into classes that distinguish stressed syllables from unstressed ones. If the means and variances of these groups are known, and the distribution is approximately normal, then the problem of predicting syllable duration can be reduced to one of determining which factors contribute to any deviation from the mean, and to what extent. Since the interactions of the factors are complex to model, a neural net is employed to map syllable features onto durations.

One significant advantage of computing durations in the log domain, rather than with raw milliseconds directly, is the shift in granularity that results in syllable durations being more sensitively described at the lower extreme, where shortening frequently occurs and where, presumably, the most accurate adjustments have to be made if the small rhythmically motivated changes in duration are to keep vowel onset locations in a synchronous relationship (see for example Lehiste, 1977). The transform also reduces any statistical bias introduced by segments in the database that have undergone the extreme lengthening typical of phrase-final or sentence-final position.

Using a measure of elasticity of segments, it can be shown that differences in duration between contoid and vocoid segments in short syllables and in long syllables, although very different in absolute millisecond terms, appear uniform in terms of compression or expansion. Because of this, the problem of accommodating segment durations to an overall syllable duration is reduced to one of finding a suitable value which can be applied to modify the mean duration of each phoneme in the syllable, in terms of its standard deviation, such that the results sum to the desired duration for the syllable as a whole.

2. CORPORA

Two corpora of speech were segmented, and their durations analysed to provide data relevant to syllable and segment durations for the development of the model. The Spoken English Corpus (SEC) (Taylor & Knowles, 1988) was used for examples of fluent speech with natural prosody from a broadcast reading of a continuous text, and the Spoken Corpus Recordings in British English (SCRIBE) database of phonetically dense sentence readings was used for balanced segmental information.

2.1. The SEC Database

Section SECG01 of the Spoken English Corpus, a twenty-minute short story written by Doris Lessing and read on BBC Radio 4 by Elizabeth Bell was prosodically transcribed by two phoneticians. It was then measured for duration at the syllable level. Tests were performed to determine which features were most significant, and a testing/training database of 3959 syllables was constructed. This provides a good source of reliable and natural prosodic information at the syllable level, but lacks segmental detail.

2.2. The SCRIBE Database

The 200 sentences in the SCRIBE database were constructed to provide data on the permissible demi-syllables of English with almost all combinations of single consonants (in both initial and final position) and vowels as well as examples of consonant clusters up to length four. The sentences were read both in isolated word form and as complete sentences by an adult male speaker of RP English. They are all valid sentences of English, but the prosodic naturalness of such readings of artificially constructed texts has to be in doubt, as the speaker has no emotional involvement in their content and no hearer for whom their message is intended, other than a microphone and any future listeners to his recording. They are well annotated examples of the possible phonetic effects of combinations of segments in speech and, in conjunction with data from the SEC database, provide for a sound coverage of durational effects at both levels.

3. TRAINING A NEURAL NETWORK

Because of the difficulty of modelling by rule the interactions between the factors that affect syllable duration, a three-layer neural network (see Chapter 8 in Rumelhart, McLelland, & the PDP Research Group, 1988) is employed to predict the deviation of individual syllable durations from the mean. This mathematical mapping from feature

descriptors to syllable durations is similar to regression analysis but enables modelling of non-linear relationships as well. Its disadvantage is that there is no accountability, and that once trained, changes are difficult to make without lengthy retraining.

Training the network involves minimising an error term and is by no means a certain process. Experience has shown about a thousand epochs to be satisfactory with the current database; further training results in little improvement in the output, and severe over-training results in a worsening of performance, presumably as the net begins to learn the exceptions and over-generalises. Training time, however, is not a significant factor in the performance of a connectionist system, because once trained the operation is extremely fast and further training is only required for a change of speaker or improvements in the feature descriptors.

Data tables of the weights and biases learnt for each of the connections in the network and of the means and standard deviations of log-transformed segmental durations are stored in separate files that are compiled in with the program source code separately. In this way the timing characteristics for different speakers can be modelled by using different or interchangeable data tables, but at present only data for one speaking style and rate has been used and changes in the speed of speech are effected by modification to the segment accommodation.

Before training, all weights and thresholds in the network are set to a small random number. They are then incrementally modified during each training epoch by back-propagation of an error term to optimise the weights and thresholds on each of the connections between the input units, the hidden layer, and the output unit. No direct connections are employed between the input units and the output unit. In training, a $7 * n$ matrix (where n is the number of syllables in the training set) is constructed of feature-set – syllable-duration pairs and presented to the network in its training mode. Each pair consists of a set of feature settings for the syllable and the value in log-transformed milliseconds of the duration recorded for the syllable described. Each value in the input vector can be thought of as quantifying the activation of a lower-level network describing one of the features of a syllable as detailed in the folowing section. Input is categorial for the foot-level feature, and continuously varying in the range of 0 to 1 to describe the degree of activation of each of the other features. The analog value of the output unit, typically in the range of 0.3 to 0.7 is increased by a factor of ten and its exponential taken as the duration in milliseconds.

Operation of the network in prediction mode is extremely fast. Weights that have been learned during the training periood are applied to the connections from each input unit to each of five hidden units, and from each of the hidden units to a single output unit. Thresholds, similarly learned, filter the input to each of the hidden units and to the output unit. In this way a small number of very simple multiplications and additions converts the feature vector to a syllable duration.

4. SYLLABLE-LEVEL TIMING

The network is trained with syllables from the SEC database but used with unforeseen text. This text has to be described in a way that equates to the tagging of the training database. Each syllable is therefore described in terms of the following six

features:

1) The number of phonemes it includes. This is a simple measure of the complexity of the syllable in structural terms. Compare the monosyllabic word *"strengths"* with the indefinite article *"a"*. Both are single syllables, but the former is typically longer than the latter pronounced as a diphthong, and even more so in the usual case when it is pronounced as a short schwa.
2) The nature of the syllabic peak. The sonorant that forms the core of the syllable is categorised into reduced, lax, and tense vowels, syllabic consonants, and diphthongs. In this way, a syllable with a complex peak can be predicted to be longer than one with a tense one, which in turn can be predicted to be longer than one with a lax peak.
3) Position in tone-group. Syllables are distinguished according to initial, medial, or final position in major and minor intonational phrases. Final lengthening is a widely acknowledged feature of English speech (see Klatt, 1976, for a summary), and in addition, a degree of initial shortening has also been noticed in our corpus. Finer distinctions have been shown to be required (Ladd & Campbell, 1991) but the transcription process does not yet discriminate in enough detail for these to be implemented in an automatic system.
4) Type of foot. The foot is the unit of speech that contains one stressed syllable with a number (possibly zero) of unstressed syllables following it, the number of which has been shown to be significant to the length of both the stressed and the unstressed syllables (Campbell, 1988, 1990c). Phrase-initial unstressed syllables, occurring immediately after a prosodic boundary, and lacking a leading stressed syllable, are categorised differently from other unstressed syllables and typically undergo more extreme shortening. Each syllable is labelled with a code that describes the type of foot it occupies, which can then be considered in conjunction with its stress feature for a fuller categorisation.
5) Stress. Stress and accent are marked in the training database from a perceptual analysis of the original recordings. They have to be approximated for each syllable in the text to be synthesised by consideration of dictionary information for the parent word and position of that word in the intonational phrase. In synthesis, rhythm rules (see Monaghan, 1990) ensure that a suitable default intonation is achieved. Semantic and pragmatic criteria also have an effect on timing, for example in cueing contrastive stress and focal emphasis, but these are currently difficult to predict by rule and for labelling stress we simply consider whether a syllable has been transcribed with a full vowel (i.e. not reduced), whether it has been flagged as stressed, and, if so, whether it has also been flagged as accented, in which case it is further subcategorised according to whether the pitch movement is simple or complex.
6) Word-class. Syllables in content words, typically information-carriers such as nouns and adjectives, are distinguished from those in function words. Content words in a text may not always be marked as bearing stress but their timing characteristics are often found to be different from function words, and unstressed syllables in a content word may be articulated differently from those in a function word. The word, as a

prosodic unit, does not not appear to have a significant status in this database, but by marking all syllables according to the role of the parent word we ensure that any differences in articulation can be modelled.

A categorial factor analyis has produced partial correlations of $r = 0.36$ for number of phones, $r = 0.56$ for type of peak, $r = 0.40$ for position in the tone-group, $r = 0.15$ for type of foot, $r = 0.44$ for stress, and $r = 0.44$ for word class. The multiple correlation coefficient for the combined six factors is $r = 0.87$, which indicates an ability to account for 76% of the variance in the durations of the database. Since there is considerable variation in rate of articulation in the reading (to reflect different degrees of excitement in the story), it can be assumed that this combination of factors would be capable of predicting even more of the variance in a reading with a fixed rate of presentation. For comparison, the output of the MITalk duration algorithm, with 'speech-rate' optimised for a best fit to the data, accounted for 68% of the variance.

4.1. Determining segmental durations

Given the overall syllable duration, individual durations for each component segment can be calculated based on the distributions observed for each corresponding phoneme class in the sentence-level corpus under the assumption of an elasticity principle (Campbell & Isard, 1991). A single factor is computed that can be applied to each segment in the syllable in terms of standard deviations about its mean to produce an optimal fit to the overall syllable duration.

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 British English phonemes /æ/ (mean ($æ\mu$) = 119 ms, standard deviation ($æ\sigma$) = 37ms), /t/ (mean ($t\mu$) = 41 ms, standard deviation ($t\sigma$) = 21 ms), and /d/ (mean ($d\mu$) = 39 ms, standard deviation ($d\sigma$) = 19 ms). 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 total 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 peak and coda lengthen equivalently by a uniform amount in terms of their standard deviations. In this case a value can be found for k such that $æ\mu + t\mu + k(æ\sigma + t\sigma) = 300$, which yields /æ/ = 208 ms and /t/ = 92 ms. A slightly different value of k is needed for $æ\mu + d\mu + k(æ\sigma + d\sigma) = 300$; where /æ/ 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 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. This appears to account quite simply, though in fact not completely, for the lengthening that has been observed in vowels of English before voiced consonants.

Once the overall duration for the syllable has been determined by the network, a secondary function returns the durations for each of its phonemes. It takes as arguments values representing: 1) the phoneme string of the parent syllable, 2) the duration in milliseconds of the parent syllable, and 3) a value by which the duration of the syllable is to be modified to account for changes in speech rate.

The timings for each phoneme in the syllable are returned in milliseconds, computed by solving equation (1) for k:

$$= \sum_{i=1}^{n} \exp(\mu_i + k\sigma_i) \quad (1)$$

where is the duration determined for the whole syllable and n is the number of segments in that syllable. Segment(i) is assigned the duration $\exp(\mu(i) + k\sigma(i))$.

The factor k is initially assumed to be zero valued, and then grown in steps of 0.1 positive or negative until the resulting segment durations sum to fit the syllable duration. To reduce the coarseness of this fitting method, and to allow for influence from articulatory constraints on the production of the sound (presumed to be reflected in the typical durations for each segment), the final value of k is further reduced towards zero by a parameter (typically set to about 0.075, giving a granularity in the range of two to six milliseconds per segment) so that usually long segments forced into a short syllable duration will lengthen it, and usually short ones in a long syllable need not fully fill it.

Any further modification that is required, such as lengthening or shortening of segments uniformly or individually, can be performed by adjusting the factor after an optimal fit to the rhythmic framework has been found. Segments in phrase-final position are not lengthened in the same way as segments in a stressed phrase-internal syllable. This is currently implemented with the use of a decay variable that reduces the effect of any lengthening on segments earlier in the syllable, further away from the boundary, lengthening proportionally more towards the end of the final syllable. Thus in a final syllable, the *i*th segment is assigned a duration of $\exp(\mu_i + 0.75^{(n-i)}k\sigma_i)$.

4.2. Improvements to the model

The probability distribution functions have been determined for the durational characteristics of each phoneme class from the phonetically balanced SCRIBE recordings. The current version of the model employs log-transformations of durations in the segment database, which are close enough to a normal distribution to allow the means and standard deviations of the log-transformed raw-millisecond observations to be used in the algorithm, but current investigations show that a better fit may be achieved by the two-parameter gamma distribution, as described by Levinson (1986), and a slight modification to the algorithm may be necessary to incorporate this change and allow prediction without the iteration that is currently required to fit the exponential.

A good fit to either distribution can only be achieved, however, by a finer classification of the segments than simple broad-class phonemic labelling provides. For example, an /r/ in isolation shows quite different durational characteristics from an /r/ in a cluster such as /spr/, and the component parts of more complex segments such as plosives are better labelled separately, distinguishing closure from burst or aspiration.

5. CONTROLLING RATE OF ARTICULATION

A form of speech-rate adjustment is effected by post-modification of the k factor, but two further controls are included at the syllable level to allow both overall modification

of the length of syllables, and of the differences in contrast between longer and shorter syllables.

After the operation of the network there is both an absolute and a relative change that can be applied to modify the predicted syllable duration. An absolute increment, adding or subtracting a number of milliseconds to the duration of the syllable, will have stronger effect in the case of shorter syllables, reducing the contrast between long and short syllables. A relative duration change, multiplying the predicted duration by a given amount, will have the greatest effect on longer syllables, increasing the contrast between the extremes. These parameters are set to zero in the case of the absolute increment, and one for the relative. (See Campbell, 1990b, for a discussion of the relative reductions of variance in faster speech).

At the segment level, output modification is performed by adding or subtracting a fixed value to each constant (k) after it has been determined as optimal to fit the segment sequence to the syllable duration. Because this over-ride is applied after matching has been performed, it applies systematically to each segment in each syllable in terms of its own distribution while preserving the higher-level timing relationships between the syllables.

By separating the control of rate at the two levels, it is possible both to model the changes in the contrast that occur when speed of presentation is increased, and to control the length of each syllable, preserving the inter-syllable regularities, according to the distribution of the component segments. It must be emphasised here, though, that the types of changes described above are not sufficient in themselves to model the stylistic changes that typically characterise a change in speaking rate. They allow control over the articulation of individual segments but a separate control of elision, assimilation, pause insertion, and spectral characteristics is also required. Such segmental changes are beyond the scope of a timing alorithm.

6. ACCOMMODATING SEGMENTS TO SYLLABLE TIMINGS

The performance of the neural network and the syllable-level prediction has been reported previously (Campbell, 1990a), but the assumption that segments can be accommodated to a higher-level framework in a simple way needs to be confirmed. If expansion and compression are indeed uniform across the syllable, then application of the constant term k to determine the duration of the segments from the overall syllable duration can be considered valid in the model. An experiment was carried out to determine the extent to which the expansion or compression of each segment, measured in normalised terms, is uniform throughout the syllable.

6.1. Procedure

The SCRIBE sentences, read once in isolated word form and once as complete continuous sentences by an adult male speaker of RP English, were segmented to produce files of phoneme labels with associated durations and diacritics

Figure 1 shows the aligned machine-readable phonetic alphabet (MRPA) transcriptions of sentence #3, *'Amongst her friends she was considered beautiful.'*, in both isolated-word and continuous-speech form, to illustrate the procedures for determining

```
 @  M  UH  -   - NG  K  K  S  T  T  H  @   #   F  R  E  -  N  D  D   Z
 x  x  1n  -   -  x  c  b  x  c  b  x  1   x   x  x  1  -  x  c  b   x
66 51 120  -   - 61 44 12 96 51 81 54 279 607 90 46 146 - 84 56 12 152
 @  M  UH UH  UH NG  -  -  S  T  T  -  @   -   F  R  E  E  N  -  - ZH
 x  x  2n  x   n  x  -  -  x  c  a  -  x   -   x  x  1  n  x  -  -  x
32 41 15  48  17 56  -  - 54 21 29  - 66   - 129 31 72 41 133 -  - 35

..   #  SH  II  #  W   O   Z  #  K  K  -   N   S  I  D  D @@  D  D   #
..   x   x   1  x  x   1   x  x  b  a  -   s   x  1  c  b  x  b  c   x
.. 471 121 300 676 66 244 195 51 53 54  - 114 117 74 35 12 88 53 19 673
..   -  SH   I  -  W   @   Z  -  K  K  @   N   S  I  D  -  @  D  -   -
..   -   x   x  -  x   x   x  -  c  a  x   x   x  2  x  -  x  c  -   -
..   -  74  75  - 17  32  62  - 41 32 31  73 109 63 24  - 53 63  -   -

..   B  B  Y  UU  T  T  I   F  @   L  #
..   b  b  x   1  c  b  x   x  x   x  -
.. 49 25 49  67 11 30 62 100 30 102 650
..   B  B  Y  UU  T  T  I   F  @   L  -
..   c  b  x   1  c  a  x   x  x   x  -
.. 63 19 22 102 20 26 46  90 49  82  -
```

The diacritics represent: c - plosive closure, b - plosive burst, a - aspiration, n - a nasalised segment, s - a syllabic segment, 1 - primary stress, 2 - secondary stress, and x - a placeholder with no meaning. # is the symbol used for a word boundary).

Figure 1. MRPA transcriptions for two readings of SCRIBE sentence #3.

word and syllable boundaries in the continuous speech data. Each character is followed by a diacritic and a duration in milliseconds. The two datasets were matched phonemically where possible, allowing for insertions and deletions, and word boundaries (indicated by a '#' symbol in the figure) were inserted into the corresponding positions in the continuous speech data. The isolated-speech data was not used further in the experiment.

As the duration measurements are at the level of the segment and no syllable boundary information is marked in the database, it was necessary to group related segments into syllables. All vowels and syllabic consonants were tagged as *peak*. The majority of words in the sentences are monosyllabic, so for these all consonants preceding the first vowel or syllabic consonant after a word boundary were tagged as *onset*, and all following the vowel and before a word boundary were tagged as *coda*. Similar tagging was performed in the case of polysyllabic words, but with any internal consonants considered potentially ambisyllabic and tagged as *medial*. Syllabification of polysyllabic words was then performed such that a single medial consonant was assumed to be preceded by a syllable boundary, and functioning as *onset*, a pair of

Table 1
Mean lengthening for each group by position in the syllable

	long syllables			short syllables			final syllables		
	mean	sd	n	mean	sd	n	mean	sd	n
onset	1.56	0.93	102	-1.22	0.56	99	0.24	0.98	149
peak	1.47	1.16	187	-1.22	0.51	170	1.09	1.25	245
coda	1.03	1.08	87	-1.12	0.38	37	1.14	1.21	242
medial	1.48	0.92	63	-1.26	0.55	37	0.48	0.96	83

medial consonants were assumed to have a syllable boundary between them, the first being *coda* and the second *onset*, and a cluster of three or more medial consonants were assumed to have a syllable boundary between the second and the third.

An examination of the syllables produced in this way revealed few obvious cases of mis-syllabification, but in all cases, the *medial* tagging was retained when a consonant was assigned to either onset or coda position as a result of the application of the above rules, and syllables created in this way are thereby distinguishable from those derived from monosyllabic words.

Syllables were grouped into three classes of length by taking as a criterion the plus-minus one sd cutoff for the averaged z-score of component segments after the sentence-final syllables were removed. This gave 439 segments in long syllables, 343 in short ones. The means and standard deviations of the individual segment values in each group were then calculated for the subgroups of onset, peak, coda and medial segments.

Because of the tendency to maximise onset in the syllabification procedure, medial segments showed a clear bias in their distribution; of the 63 medial segments in the *long* group, only 15 are in coda position, and the remaining 48 in the onset. By definition, all medial segments in sentence-final syllables are in onset position.

6.2. Results

In millisecond terms there are considerable differences in the timing characteristics of vowels and consonants. If these are real differences, and not just artefacts of the different elasticities of the different phone types, then they should be preserved in a measure of the lengthening undergone by each. In this case, there will be significant differences between the results for segments in onset and coda position, and those in the peak.

The overall mean was 1.4 (sd 1.07, n = 439) for the *long* group, and −1.2 (sd 0.54, n = 343) for the *short* group. As can be seen in Table 1, for the individual onset, peak, medial, and coda segment categories we find little variation from that mean in long and short sentence-internal syllables, but considerably more in the sentence-final ones. Means for the components of the intermediate group of syllables were by definition close to zero, but the standard deviations were reduced from 1 to 0.8 as a result of the factorisation. Results of tests of significance of the differences of these means are shown in Table 2.

Table 2
Student's t tests for significance of difference in the means of Table 1

peak vs. coda				onset vs. peak				onset vs. coda			
	t	sig	df		t	sig	df		t	sig	df
short	1.13	n.s.	205	long	0.67	n.s.	287	long	3.62	<0.001	187
long	2.98	<0.01	272	final	7.08	<0.001	392	int.	3.09	<0.01	3501
final	0.45	n.s.	485								

In the *short* group, compression appears to be constant across all syllable parts and no significance was found in the small differences in the means of the peak and coda segments. This result indicates a uniform factor of compression for the shortening undergone by segments in these syllables.

The assumption of a factor of lengthening applying within the syllable and insensitive to any difference in absolute durations between vowels and consonants is also supported by the lack of significance in the difference between means for onset and peak segments in the *long* group. The differential shortening of coda segments in contrast to the onset and peak segments in this group is, however, found to be significant.

These results for coda segments may be due to an inherent difference between onset and coda classes in general, as a significant shortening of coda segments is also found in the reference group of *intermediate* syllables whose averaged values fall between +1 and −1.

That the lengthening of syllables in phrase-final position is different from that applied within a phrase is shown by the difference between means for onset and peak segments in these cases. In contrast to the phrase-internal case, no significance was found in the difference between peak and coda segments in syllables lengthened phrase-finally.

6.3. Discussion

The compression and expansion undergone by component segments of the longer and shorter non-final syllables shows that it is not the vowel taking most of the change, as raw duration measurements would indicate, but a factor operating in an apparently consistent way throughout the syllable. In the case of pre-pausal syllables, there is a significant difference between the lengthening found in onset segments and those occurring in the peak and coda of sentence-final syllables.

A further point of interest is the relative stability of consonants in the coda as compared with those in the onset. This may indicate the functioning of an intermediate-level construct, the ryme, but tests to determine this are still being carried out.

7. CONCLUSION

This paper has described a model of speech timing that is being used to predict durations in a text-to-speech synthesis system. The model is different from other duration algorithms in that it employs higher-level features to determine the timing framework into which segment durations can be accommodated. It has shown that

durations can be predicted at the level of the syllable by simple description of a few features that describe the main structural and informational aspects of the speech signal. It has also shown how the durations of the segments that make up these syllables can be determined from knowledge of the overall syllable duration and of the probability density distributions of the different phone durations in a given corpus of speech.

The best test of such a model is the actual performance of the system itself, but since it has been trained on data from two different speakers using very different reading styles, a direct numerical comparison of the resulting durations and original measurements is difficult. Informal perceptual tests show that it produces intelligible but generally fast timings. A formal comparison of the durations predicted by the algorithm with those produced by twelve human readers for a paragraph of text (see Campbell, 1990b) shows the variance to be within the range observed for the human readers, but the extent to which that variance is perceptually significant has yet to be quantified. Psychometric tests are currently being carried out at Edinburgh to measure any degradation in the performance of a task when natural speech timings used in synthetic speech interactions are replaced by those generated by this algorithm.

ACKNOWLEDGEMENTS

The model described in this paper was developed at the IBM UK Scientific Centre in England (syllable prediction), and at the University of Edinburgh Centre for Speech Technology Research in Scotland (segment accommodation). The author is also grateful to the management at ATR Interpreting Telephony Research Labs in Japan for enabling this paper to be presented.

REFERENCES

Beckman, M. J. (1991). "Metrical structure versus autosegmental content in phonetic interpretation." Proceedings of the XIIth Internatinal Congress of Phonetic Sciences, Aix-en-Provence, France.

Browman C. P., & Goldstein, L. (1990). "Gestural specification using dynamically defined articulatory structures." *Journal of Phonetics* 18, 299 - 320.

Campbell, W. N. (1988). "Foot-level shortening in the Spoken English Corpus." Proceedings of the First European Conference on Speech Technology, Edinburgh, 698 - 701.

Campbell, W. N. (1990a). " Analog i/o nets for syllable timing." *Speech Communication: Special Issue on Neural Nets and Speech*, 9, 57 - 61.

Campbell, W. N. (1990b). "Timing invariance in read speech." Proceedings of the ESCA Tutorial and Research Workshop on Speaker Characterisation, Edinburgh, 78 - 82.

Campbell, W. N. (1990c). "Shortening of feet in longer articulatory units." *Journal of the Acoustical Society of America*, Suppl.1, Vol 88, 6SP11.

Campbell, W. N. & Isard, S. D. (1991). "Segment durations in a syllable frame." *Journal of Phonetics: Special issue on Speech Synthesis*, 19, 37 - 47.

Carlson, R., Granström, B., & Klatt, D. H. (1987). "Some Notes on the Temporal Perception of Speech." in *Frontiers of Speech Communication Research*, Lindblom & Öhman, Academic Press.

Crystal, T. H. & House, A. (1988). "Segmental durations in connected speech signals: Current results." *Journal of the Acoustical Society of America* 83(4), 1553 - 1573.

Crystal, T. H. & House, A. (1990). "Articulation rate and the duration of syllables and stress groups in connected speech." *Journal of the Acoustical Society of America* 88, 101 - 112.

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

Klatt, D. H. (1976). "Linguistic uses of segment duration in English." *Journal of the Acoustical Society of America* 59, 1208 - 1221.

Klatt, D. H. (1979). "Synthesis by Rule of Segmental Durations in English Sentences." in *Frontiers of Speech Communication Research*, Lindblom & Öhman Academic Press, 287 - 300.

Ladd, D. R. & Campbell, W. N. (1991). "Theories of Prosodic Structure: Evidence from syllable duration." Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, France.

Lehiste, I. (1977). "Isochrony Reconsidered." *Journal of Phonetics* 5, 253 - 265.

Levinson, S. E. (1986). "Continuously variable duration Hidden Markov models for automatic speech recognition." *Computer Speech and Language* 1, 29 - 45.

Monaghan, A. I. C. (1990). "Rhythm & Stress Shift in Speech Synthesis." *Computer Speech and Language* 4, 71-78.

Rumelhart, D., McLelland J. and the PDP Research Group (1988). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Chapter 8). MIT Press.

Taylor L. J & Knowles, G. O. (1988). *Manual of Information to Accompany the SEC Corpus*, IBM UK Scientific Centre/Lancaster University.

van Santen, J. (1990). " Deriving text-to-speech durations from natural speech." Proceedings of the First ESCA Tutorial and Research Workshop on Speech Synthesis, Autrans, France, 157 - 160.