# Higher-level Timing Control for an English-language Speech Synthesis System

## 英語音声合成システムのため階層的なタイミング制御

*W. N. Campbell*
ウイルヘルム N. キャンベル

ATR Interpreting Telephony Research Laboratories

ABSTRACT: The timing component of many current Text-to-Speech systems is based on rule-governed modification of a set of inherent segment durations. In contrast, the system described here is two-layered; durations are first calculated at the syllable level, to reflect rhythmic and structural organisation of the utterance, and then accommodated amongst the components of the syllable at the segmental level. A neural net has been trained to predict syllable timing based on exposure to a corpus of syllable-feature – duration pairs from measurements of spontaneously occurring natural speech. When tested with three different thousand-syllable passages from the same corpus as the training set, it accounted for approximately 70% of the variance observed in each of the test cases. Segment duration is calculated as a second stage of the process, fitting the predicted syllable duration while staying within observed distribution probabilities for each segment.

まえがき: 従来、英語の規則合成システムにおける継続長制御は、音韻に固有な継続長を規則に基づいて変える方法によっていた。これに対し本システムにおいては、最初にリズムや構造を反映した音節単位に継続長を算出した後、音韻の継続長を算出する２つのステップを踏むことを特徴としている。音節の継続長は、継続長、音節の種類が付与されている自然発声の音声データベースを用いて学習したニューラルネットによって推定する。次に、推定された音節長内での音韻の継続長は、各々の音韻の平均と分散を用いて音節内のちらばりが一定になるように決定する。最後に、作成されたネットワークに、各々１００の音節を持つ、３つの未知データセットで評価した結果、それぞれ約７０％の分散で推定されることがわかった。また、音韻の継続長推定結果も示す。

## 1 Theoretical background

The two main assumptions of this method are

1. That speech timing is a function of higher-domain processes operating at the level of the syllable and above.
2. That segment duration can be determined by accommodation into a syllable frame according to known probability distribution functions for each phoneme.

The most widely used duration-assignment algorithm, proposed by Dennis Klatt [14] and incorporated into the MITalk system, operates at the level of the segment, shortening or lengthening the inherent duration of each according to an ordered set of rules which make phoneme-specific adjustments, either multiplicative or additive, according to phonetic and phrasal contexts.

The concept of a minimum duration [15] for each phoneme is required to ensure that shortening is never too extreme, although there is no explicit upper limit to the lengthening each may undergo.

It is assumed in the Klatt system that rhythmic effects are a by-product of segment-level modifications, and no explicit control is offered [10].

In order to have a better control over rhythmical features, it is assumed here that duration is a function of the syllable as a unit, and is calculated accordingly, without consideration of phonetic constraints[1] or segment-specific effects, which are then accommodated at a later stage of processing.

The distribution of syllable durations has been shown to fit a normal probability function when transformed into the log domain if separated into classes that distinguish stressed syllables from unstressed ones.[3]. If the mean and variance 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 working in the log domain rather than with raw milliseconds directly, is the shift in granularity that results in syllable durations being more sensitively predicted at the lower extreme, where shorten-

[1] the number of phonemes in the syllable, and a simple description of the nature of the peak of the syllable are the only aspects considered at this stage

ing frequently occurs and where, presumably, 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 [16]). The transform also reduces any statistical bias introduced by segments in the database that have undergone extreme lengthening in phrase-final position.

Campbell 1990 [6] has shown that onset and peak segments are lengthened and shortened by an equivalent amount within a syllable in terms of deviation about their mean durations[2]. Using a measure of elasticity of segments it can be shown that differences in duration between segments in short syllables and equivalent segments in long syllables, although very different in absolute millisecond terms, are uniform in terms of compression or expansion. The problem of accommodating segment durations to an overall syllable duration is thus 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 anaylsed to obtain data relevant to syllable and segment durations. The Spoken English Corpus (SEC) [12] was used for examples of fluent speech with natural prosody in a connected text, and the SCRIBE corpus [13] of phonetically dense sentences 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 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

These 200 sentences were constructed to provide almost total coverage of the permissible demi-syllables in 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 committment to 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.

[2]and coda segments similarly, but with less variation.

# 3 Syllables

Each syllable in the SEC database is described in terms of

- The number of phonemes it contains.

  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.

- The nature of the syllabic peak[3].

  This distinguishes lax and tense vowels, syllabic consonants and diphthongs. In this way, a syllable with a complex peak can be expected to be longer than one with a tense one, which in turn can be expected to be longer than a lax peak.

- 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 [14], and a degree of initial shortening has also been noticed in our corpus [1].

- Position in foot.

  Rhythmic constraints are incorporated in the determination of syllable timing. The foot is the unit of speech that contains one stressed syllable and a number of unstressed syllables following it, the number of which has been shown to be significant in determining its duration [2][8].

- Stress.

  Stress and accent are determined for each syllable according to the part-of-speech of the parent word and the position of that word in the intonational phrase. Semantic and pragmatic criteria also have an effect here, in cueing contrastive stress and emphasis, but these are currently difficult to predict by rule and the criteria for training were simply whether a syllable in the corpus had been transcribed with a full vowel (i.e. not reduced), whether it had been noted as stressed, and if so, whether it had also been transcribed as accented, in which case it was further subcategorised according to whether the pitch movement was simple or complex.

- Word-class.

  Syllables in content words, typically information-carriers such as nouns and adjectives, are distinguished from those in function words.

[3]*Peak*: the vowel, vowels or syllabic consonant that form the obligatory main part of the syllable

Syllable-level timing is determined by an interaction of the above factors.

# 4 Segments

Probability distribution functions have been determined for the durational characteristics of each phoneme class from a phonetically balanced two-hundred sentence single-speaker database of recordings of male RP speech.

Because the distribution functions that describe the durations of many segments are not significantly different from normal when plotted in the log domain[4], the means and standard deviations used in the programme are calculated from the log-transformed raw-millisecond observations.

# 5 A Computer Implementation

A module for a Text-To-Speech programme that implements duration is described here. It consists of two main functions *Syll-dur* and *Phon-dur*, of which the former calls a subsidiary sigmoid transfer function to calculatedetermine output of hidden units in a three-layer neural net.

Data tables of weights and biases for each of the connections in the net, and means and standard deviations of log-transformed segmental durations are stored in a separate header file and compiled in with the program source code before run time. Timing characteristics for different speakers can be changed by using different data tables, but at present only data for one speaker is used.

## 5.1 Controls

Whereas no attempt is made to model speech rate in this implementation, some control for speed of speech will be required in a practical system. Three controls are included for this purpose, and between them they allow both overall modification of the length of syllables, and of the differences in contrast between longer and shorter syllables.

Within *Syll-dur* there is both an absolute increment and a relative change that can be applied after the duration has been predicted. An absolute change in duration has the effect of adding or subtracting a fixed number of milliseconds to the duration of the syllable, which will have strongest effect in the case of shortest, thereby reducing the contrast between long and short. 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 variables default to zero in the case of the absolute increment, and one for the relative.

At the segment level, output modification is performed by a parameter to the *phon-dur* function. The effect of this is to add or subtract a fixed value to the constant ($k$) determined as optimal to fit the segment sequence to the syllable duration. This over-ride is applied after matching has been performed and applies systematically to each segment in the syllable.

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 and further away from the boundary.

## 5.2 Determining Syllable durations

As its name implies, *syll-dur* predicts the duration to be assigned to a syllable. When called with a numeric vector of six feature descriptors this function uses a three-layer neural net to return a duration in milliseconds for the syllable thus described. Each value in the vector can be thought of as quantifying activation of a lower-level net describing one of the features of the syllable described in Section 3.

## 5.3 Training the neural network

Training the net involves minimising an error term and is by no means a certain process. Experience has shown 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-generalise. 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.

A $7 * n$ matrix (where n is the number of syllables in the training set) is constructed of feature-value pairs and presented to the netwotk in its training mode. Each pair consists of the six feature settings for the syllable, tagged as described above, and the value in log-transformed milliseconds of the duration of that syllable.

Weights are on 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 are on each of the hidden units and on the output unit. No direct connections are employed between the input units and the output unit. Initially all weights and thresholds are set to a small random number prior to training.

The net is trained with standard back-propagation of error [17] to optimise the weights and thresholds on each of the connections between the input units, the hidden layer, and the output unit. Input is not binary, but continuously varying in the range of 0 to 1 to describe the degree of activation of each of the six 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.

The weights thus determined are included as a data table in the program at compile time and serve to predict the durations when the network is switched from training mode to run mode.

[4]It may be that the lack of normality in the distribution of the other segments can be accounted for by the dense construction and exceptional production environment of the sentences, in which internal pauses are kept to a minimum and phrase-final lengthening therefore artificially low.

The major drawback of a neural-network system is that once trained, there is no way of interpreting the information contained in the weights. If changes are to be made, it is essential to retrain a completely new set of weights rather than attempt to modify any that are in an existing set.

### 5.4 Determining Segment durations

Once the overall syllable duration has been determined, individual durations for each component segment can be calculated based on the distributions observed for each corresponding phoneme class in the sentence-level corpus.

Assuming an elasticity principle [9] under which all segments in a syllable fall at the same place in their respective distributions, 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. This factor 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[5].

Any further modification that is required, such as lengthening or shortening of segments uniformly or individually, can be performed by adjusting the factor as required after an optimal fit to the rhythmic framework has been found. In the case of phrase-final syllables, the factor is applied in a non-uniform way, lengthening proportionally more towards the end of the syllable. Speech-rate modification is effected by similar post-modification of the factor.

Given the overall duration for the syllable, the function *phon-dur* 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 of the parent syllable in milliseconds.
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 the following equation for $k$:

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

where: $n$ is the number of segments in the syllable,
and $\Delta$ is the duration determined for the whole syllable.
Segment$_i$ is assigned the duration $exp(\mu_i + k\sigma_i)$.

The means and variances of each phoneme class are held in a lookup table derived from analysis of the twelve-thousand segments of the SCRIBE database.

## 6 Testing the model

Since there is so much variability inherent in natural speech it is not an easy matter to judge the correctness of a single duration value determined by such a system. Durations are evaluated perceptually in an informal manner in everyday use of the synthesiser, but although gross mispredictions may be obvious in such an informal evaluation, minor errors within the difference limens for individual segments may pass unnoticed while perhaps compounding to degrade the overall performance in less easily discernable ways.

For a more objective quntification of the performance of the system, durations can be compared to those observed in real speech, but it is not yet established which numerical timing differences may be perceptually significant. Thus, even if a predicted duration is within a predetermined percentage or range of the timings of a naturally spoken model, it cannot necessarily be assumed that the original can be deemed typical or representative in itself. The same speaker may produce a quite different durational structuring of the same utterance on a different occasion. Nor is it certain that a mismatching duration wouldn't be correctly predicted by a different description of the syllable. Since it is rare for even two human transcribers to agree on so subjective an area as intonational or suprasegmental features, we have to accept a degree of uncertainty in any quantitative measure of the output.

The quantification of fit at the segment level should be easier. If syllable duration is simply the sum of the durations of the segments comprising that syllable, and since all durations are obtainable from the database, then we ought to be able to use the information in the database to test the performance of the procedure absolutely. Syllabification, however, is uncertain; many consonants in polysyllabic words are ambisyllabic, and as it is still unclear which syllabification procedure is optimal. Even with monosyllabic words, (the majority), a degree of resyllabification may take place across word boundaries and such efects can neither be discounted nor easily quantified except in a circuitous manner.

Results of a test of segment accommodation are reported in Campbell 1990b [6] and show that for the subgroups of syllables lengthened or shortened by an amount greater or less than one standard deviation when averaged across all segments in the syllable, the principal of a common factor describing segmental elasticity can be shown to apply at least to the onset and peak elements, and similarly though with less variation to coda segments.

In a follow-up study (Campbell 1990c [7]) three different types of lengthening were shown; stress-induced lengthening is accommodated differently within the syllable from phrase-final lengthening, and modification for phonetically-motivated lengthening (such as occurs in the case of vowels before a voiced plosive) were also found. The system described here incorporates only the first two; future work will test how much more can be accounted for by incorporation of the third by including sensitivity to context-specific segment timing characteristics.

### 6.1 Fitting one sentence

By way of illustration, details are reproduced here for the prediction of the first sentence from the SCRIBE corpus. Each line of output from the program is followed first by the durations observed for readings as continuous speech, then for the durations of the words read in isolated mode. Values shown are: syllable-descriptor array, syllable duration, sum-

[5] To reduce the coarseness of this fitting method, the final factor is further reduced by 0.05, giving a typical granularity in the range of two to six milliseconds.

of-means, factor, and individual segment durations. Transcription is according to EU MRPA conventions. The text is: *"The price range is smaller than any of us expected."*

```
011121   80   67   0.3 dh:32 @:48
cont:    40                28   12
isol:   246                30  216

244849  394  354   0.2 p:96 r:46 ai:131 s:105
cont:   390             119   35    117   119
isol:   534             105   42    165   222

224849  305  330  -0.2 r:44 ei:113 n:66 jh:82
cont:   240              35   108   50    47
isol:   546              68   195   91   191

233821  170  120   1.0 i:77 z:94
cont:   143              56   87
isol:   356             156  200

234535  268  276  -0.1 s:93 m:57 oo:121
cont:   300              87   63    150
isol:   354             166   37    151

241225  191  187   0.1 l:46 @@:142
cont:   199              33    166
isol:   207              53    154

021131  118  141  -0.3 dh:22 @:35 n:64
cont:   143              61   41   41
isol:   383              69  149  165

233711  129   85   1.2 e:131
cont:    97              97
isol:   118             118

231425  137  157  -0.3 n:64 ii:72
cont:    90              22    68
isol:   237              51   186

031421   90   88   0.1 @:42 v:48
cont:    92              31   61
isol:   234             190  144

031421   90  179  -2.1 uh:47 s:43
cont:   140              47   93
isol:   421             198  223

231421  120  135  -0.2 i:47 k:74
cont:    87              26   61
isol:   144              77   67

234641  289  355  -0.6 s:77 p:73 e:69 k:65
cont:   272              60   83   71   58
isol:   307              56   97  115   39

251331  441  154   2.1 t:117 i:98 d:119
cont:   250             111   66    73
isol:   238              92   73    73
```

## 6.2 Comments on the output

Word one, *the*, varies considerably in the two readings; from 246 ms when read in isolation down to 40 ms when read with and presumably cliticised to *price*. The prediction of 80 ms is closer to the latter but the distribution of the segment durations shows a reversal, with the vowel taking less of the syllable in the read speech and more in the computer-generated version. Twelve milliseconds is indeed very short for even a schwa, and the 32 ms 48 ms distribution accords better with the overall averages for those segments, but perhaps the timing algorithm has missed a generalisation in this case that requires the special distribution - it is difficult to tell.

*Price* is well predicted in terms of syllable duration, but again the segment allocation of duration differs from the human readings, with more weight being given to the peak than the coda.

The next two words, *range* and *is*, seem to fit in between the two extremes, again closer to the measurements observed for continuous readings.

Both syllables of *smaller* are slightly shorter, perhaps because although the rules had no way of detecting that this word is the focus of the sentence, the human reader presumably did.

The next word, *than*, is considerably shorter - probably due to being marked as phrase-initial for syntactic reasons, but it is possible that the speaker chose either to group this word phonologically with the semantically related *smaller* and therefore not to put an intonational boundary in the sentence at that point at all. If the transcription is amended to remove this boundary, then the duration increases to 129 ms (after rounding the individual segment durations actually sum to 131 ms), but is still shorter than observed for the continuous speech in this reading

```
031331  134  141  -0.2  dh:24 @:38 n:69
```

and if the function/content status is changed to intermediate to reflect a more significant role in the sentence, then the overall duration matches better, but with more lengthening of the nasal in the coda than the fricative in the onset.

```
131331  154  141   0.2  dh:28 @:44 n:79
```

And so for each word in the sentence comparisons can be made, but to tell which differences are significant, and which words were stressed to what extent and for what reason in the original needs access to perceptual information and higher-level inferencing that cannot be achieved by numerical analysis alone. Campbell 1990a [5] measured the differences in timings of twelve speakers reading a passage at two different reading speeds and quantified the inter- and intra-speaker variation therein, then compared the output of the timing component with their timings only to find that there is as much variation between the durations observed for the speakers as there is between the output of the system and any one of the speakers. Needless to say, the human readers produced more natural performances in spite of any numerical equivalence that may have been found.

More work remains to be done on what perceptual significance can be determined for such variation in the timings.

Small differences such as in the durations of *than* above cannot be heard, but neither can it be claimed that they have no effect at all. Unfortunately, a test of the output that involves more than numerical comparison will also include output from other components of the text-to-speech system which may confound to the perceptual result with effects of their own.

# 7 Conclusion

Details of a two-level timing component for a text-to-speech system have been presented, showing that duration can be predicted at the syllable level according to rhythmic and structural constraints of the utterance, and later accommodated according to observed distribution probabilities at the level of the segment.

Two corpora of read speech provided the source of the observations; one providing prosodic information, the other segmental. It is hoped that as database resource grow, and larger and more representative sources of spontaneous naturally occurring speech become available, the routines developed and described here will allow the incorporation of even further knowledge into the system.

The problem in working with such large corpora is that statistical techniques have to be developed to facilitate the analysis and tagging of linguistically significant events. No description has yet been found, for example, of the events that trigger local changes in speech rate such as can be seen from a comparison of output from the system with durations in the corpus. It is hoped that the output of this model can be used not only as a tool in a text-to-speech system, but as a factoring device, to account for much of the more predictable variance in the durations observed in a database and thereby provide a more concentrated source of residuals from which further significant timing effects can be isolated. In this way both the knowledge incorporated in the program and our understanding of the timing effects of the language will grow together.

# Acknowledgement

An earlier, slightly fuller version of this paper appeared in the Edinburgh University Linguistics Department's Work in Progress 1990. I am particularly grateful to Dr Kurematsu and ATR for allowing me to continue with this research.

# References

[1] W N Campbell *Extracting Speech-rate Values from a Real-Speech Database*, pp 683 - 686, Proc ICASSP New York 1988a.

[2] W N Campbell *Foot-level Shortening in the Spoken English Corpus*, pp 285 - 288, Proc. FASE Edinburgh 1988b.

[3] W N Campbell *Syllable-level Duration Determination*, pp 698 - 701, Proc. European Conference on Speech Technology, Paris 1989.

[4] W N Campbell *Analog i/o nets for syllable timing*, in Speech Communication Special Issue on Neural Nets and Speech Vol 9 No 1, Elsevier Science Publishers B. V. (North Holland) 1990.

[5] W. N. Campbell *Timing Invariance in Read Speech*, pp 78 - 82 in Proc. ESCA Tutorial and Research Workshop on Speaker Characterisation, Edinburgh 1990a.

[6] W. N. Campbell *Normalised Segment Durations in a Syllable Frame*, in Proc. ESCA Tutorial and Research Workshop on Speech Synthesis, Autrans France. 1990b.

[7] W. N. Campbell *Evidence for a Syllable-based Model of Speech Timing*, in Proc. International Conference on Spoken Language Processing, Kobe Japan, 1990c.

[8] W. N. Campbell *Shortening of feet in longer articulatory units* Paper presented at the 120th Meeting of the Acoustical Society of America, Fall 1990.

[9] W. N. Campbell & S. D. Isard *Segment durations in a syllable frame* Journal of Phonetics, Special Issue on Speech Synthesis. In press.

[10] Carlson Granström & Klatt *Some Notes on the Temporal Perception of Speech*, in **Frontiers of Speech Communication Research**, Lindblom & Ohman, Academic Press 1987.

[11] J. Edwards and M. Beckman *Articulatory Timing and the Prosodic Interpretation of Syllable Duration*, pp 156 - 174 in Phonetica 45, 1988.

[12] IBM UK Scientific Centre/Lancaster University *Spoken English Corpus* 1988.

[13] J. Laver et al *ATR/CSTR Speech Database Project Status Report*, #1, October 1988.

[14] D H Klatt *Review of Text-to-speech Conversion for English*, JASA #82 (3) September 1987.

[15] D H Klatt *Synthesis by Rule of Segmental Durations in English Sentences*, in **Frontiers of Speech Communication Research**, Lindblom & Ohman, Academic Press 1978.

[16] I. Lehiste *Isochrony Reconsidered*, J. Phonetics 5, pp 253 - 265, 1977.

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