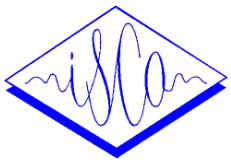

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

Normalised Segment Durations in a Syllable Frame

W. N. Campbell*
University of Edinburgh, Centre for Speech Technology Research
80 South Bridge Edinburgh EH1 1HN UK

Abstract

All segment durations measured in the phonetically balanced SCRIBE 200-sentence database were converted to standard normal form for each phoneme, with resulting distributions having zero mean and variance 1, to allow comparisons of the relative compression and expansion applied to different segments with regard to position in the syllable and in the utterance.

The data was divided into four subsets; segments in syllables from sentence-final position formed one group, then taking the plus-minus one sd cutoff as criterial, the remainder were divided according to membership of syllables assigned to long, intermediate and short classes.

Results are presented which show that segments in sentence-final position undergo greater lengthening in the ryme than in the onset, whereas segments that are lengthened sentence-internally, for stress and rhythmic reasons, are lengthened uniformly throughout the syllable. Segments in short syllables are similarly found to be shortened uniformly, regardless of position in the syllable and of phonemic distinction.

1 Introduction

Campbell 1989 [1] showed that 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 [3]. The advantage of the simpler 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 [2] in an analysis of their segment duration database showed that both consonants and vowels have similar positively skewed distribution densities. The main difference between the distributions of the different phoneme classes was in the maximum duration they could attain - in the positive end of the tail - not in the typical or most common duration, which for all segment types was 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 indicate less variability in the duration of segments in fluent speech than in the less typical modes such as readings of isolated words and citation form sentences. For a computer text-to-speech system, the requirement is more likely to be for the fluent variety 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.

*This work was supported by the Information Engineering Directorate/Science and Engineering Research Council as part of the IED/SERC Large Scale Integrated Speech echnology Demonstrator Project (SERC grants D/29604, D/29628, F/10309, F/10316, F/70471) in collaboration with Marconi Speech and Information Systems and Loughborough University of Technology.

2 Accommodating segments to syllable timings

In the text-to-speech system we are developing at Edinburgh, durations of syllables are predicted by a connectionist network trained to match syllable feature descriptors to the timings observed in a database of naturally occurring speech. Durations to be assigned to each segment within the syllable are determined as in the formula (1) below,

$$\Delta = \sum_{j=1}^{n} (\mu_j + k\sigma_j) \tag{1}$$

where: k is a constant term determined for each syllable by an iterative estimation procedure,
Δ is the duration determined for the whole syllable,
n is the number of segments in the syllable,
μ_j is the mean and σ_j the standard deviation observed in the database for segment$_j$.

A segmentally transcribed set of 200 sentences (the SCRIBE database) was used as the source of both segment and syllable durations. Means (μ) and variances (σ^2) were calculated for each phoneme, and the individual segment durations then normalised by z-transform as in formula (2) to a number (z) that reflects the amount of variation in terms of standard deviations about each segment mean.

$$z_{seg} = (observed\,duration_{seg} - \mu_{seg})/\sigma_{seg} \tag{2}$$

z can be positive or negative, typically in the range of plus or minus 3 for normally distributed data. but the positive skew observed in the SCRIBE segment data results in a range of -2 to $+5$, which difference has no significant effect on the result of the comparisons. A negative value of z is taken to indicate compression of the segment, and a a positive value expansion.

If expansion and compression are indeed uniform across the syllable, then this number will be equal to the constant term k used to determine the duration of the segments from the overall syllable duration in the TTS model. The purpose of the experiment is to determine the extent to which the expansion or compression of each segment, measured in this way, is uniform throughout the syllable.

3 Procedure

These sentences, read once in isolated word form and once as complete continuous sentences by an adult male speaker of RP English, provide almost total coverage of the permissible demi-syllables in English with almost all combinations of vowels and single consonants (in both initial and final position), as well as providing examples of consonant clusters up to length four. The two sets of speech data were segmented to produce files of phoneme labels with associated durations and diacritics

The MRPA transcriptions of sentence #3 *'Amongst her friends she was considered beautiful.'* are reproduced here in both isolated and continuous-speech form to illustrate the procedures used for determining word and syllable boundaries in the continuous speech data. Each character is followed by a diacritic and a duration in milliseconds.

isolated-speech version:
@ - 66, M - 51, UH 1n 120, NG - 61, K c 44, K b 12, S - 96, T c 51, T b 81, # - 506, H - 54, @ 1 279, # - 607, F - 90, R - 46, E 1 146, N - 84, D c 56, D b 12, Z - 152, # - 471, SH - 121, II 1 300, # - 676, W - 66, O 1 244, Z - 195, # - 51, 1 K c 53, K a 54, N s 114, S - 117, I 1 74, D c 35, D b 12, @@ - 88, D c 53, D b 19, # - 673, B c 49, B b 25, Y - 49, UU 1 67, T c 11, T a 30, I - 62, F - 100, @ - 30, L - 102.

continuous-speech version:
@ - 32, M - 41, UH 2n 15, UH - 48, UH n 17, NG - 56, S - 54, T c 21, T a 29, @ - 66, F - 129, R - 31, E 1 72, E n 41, N - 133, ZH - 35, SH - 74, I - 75, W - 17, @ - 32, Z - 62, K c 41, K a 32, @ - 31, N - 73, S - 109, I 2 63, D - 24, @ - 53, D c 63, B c 63, B b 19, Y - 22, UU 1 102, T c 20, T a 26, I - 46, F - 90, @ - 49, L - 82.

where the diacritics represent:
c: stop closure b: stop burst, a: aspiration, n: nasalised segment, s: syllabic segment. 1: primary stress, 2: secondary stress.

The two datasets were matched phonemically where possible and word boundaries (indicated by a '#' symbol) 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 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 no 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 criterial the plus-minus one sd cutoff for the averaged z value 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.

4 Results

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. 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.

	long syllables:			short syllables:			sentence-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

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. Tests of significance of the differences of these means showed the following:

peak and coda				onset and peak				onset and coda			
group	t	sig	df	group	t	sig	df	group	t	sig	df
short	1.13	ns	205	long	0.67	ns	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	ns	485								

4.1 Discussion

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 classess 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.

4.2 Conclusion

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.

References

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

[2] Thomas H. Crystal & Arthur House *Segmental Durations in connected speech signals: Current results* JASA 83(4), pp 1553 - 1573, April 1988.

[3] D. H. Klatt *Synthesis by Rule of Segmental Durations in English Sentences* pp 287 - 300 in **Frontiers of Speech Communication Research**, Lindblom & Ohman Academic Press 1979.