TOWARDS A MODEL OF SEGMENT AND SYLLABLE DURATION IN AUSTRALIAN ENGLISH

Janet Fletcher and Andrew McVeigh

Speech, Hearing, and Language Research Centre
Macquarie University

ABSTRACT - A corpus of nearly 6500 syllables and their component segments were analysed to formulate a model of segment and syllable duration for Australian English. Segments were grouped into four prosodic categories, unstressed, stressed, pitch accented and phrase final. Syllables were labelled and analysed according to their length (number of segments), prosodic context and grammatical function. Syllable duration was modelled using a three-layer neural network that was trained and tested on different portions of the database. Segment durations stretched or compressed to fit the network-assigned syllable duration frame. This relatively simple model was able to account for nearly 80% of the durational variance observed in the database.

INTRODUCTION

A number of factors are known to influence the measurable duration of a speech segment or syllable. Two important linguistic effects include stress and phrasal position. Furthermore, so-called segmental timing effects such as phonological identity of the segment and segmental context (e.g. consonant clusters) also influence the observed duration of phonetic segments. Correct modelling of these durational effects are deemed essential for the attainment of intelligible and natural-sounding synthetic speech. Many models of segment duration have been developed over the years. Among the most sophisticated are the models developed for segment duration in American English by Klatt (1979), and the companion model for Swedish, formulated by Carlson and Granstrom (1979). In a Klatt-type model prosodic factors are combined with segmental features to influence the durations of individual segments. It is assumed from the outset that each segment type has an inherent duration that is part of its distinctive phonetic properties. One often used illustration of this concept is the observed length difference among tense and lax vowels. It is also assumed that each segment has a minimum duration - i.e. a compression threshold, in addition to an inherent duration. In this type of model, prosodic and segmental timing rules are expressed as percentage changes on inherent duration to bring about either a duration increase or decrease. Inherent durations are usually calculated from /CVC/nonsense words inserted in carrier phrases of the type "Say /CVC/ again".

In an alternative modelling approach suggested by Campbell (1989), limited attention is paid to segment-specific timing factors. Rather, syllable duration is predicted on the basis of higher-level factors such as phrasal position, stress, and grammatical category of the word. The model is implemented as a three-layer neural net that is trained by back-propagation on syllable durations obtained from a database of spoken language (recorded stories). The above-mentioned linguistic timing factors act as input features to the neural net. The output of the net is a predicted syllable duration for a given set of linguistic factors. Segments are then "squeezed" into a syllable time frame on the bases of the "elasticity" hypothesis of segment duration. In its strongest form, this hypothesis states that the duration of each segment of a particular syllable can be estimated by adding the mean duration of the segment (calculated across the database) to a number of standard deviations (k) multiplied by the mean. The value of k is the same for each segment in the syllable, and is calculated irrespective of the prosodic characteristics of the particular segment. Campbell found that at an initial pass, this modelling procedure was able to account for 70% of the durational variance within the database under investigation. Previous application of the Klatt model to the same database yielded a similar result. In a later study, Campbell and Isard (1991) suggest modifying the "elasticity" hypothesis to take into account factors such as position in phrase.

This kind of model assumes the primacy of a syllable-type time frame over a segment-based time frame in reflecting prosodic and higher-level timing patterns. Furthermore, unlike earlier segmental-

timing models, this one is based on a large corpus of general speech data. The early Klatt and Carlson and Granstrom models were usually based on analyses of corpora that range from isolated nonsense words to words inserted in fixed carrier sentences. There is a current trend to supplement these historical durational paradigms with analysis of large corpora of general speech material (Carlson and Granstrom, 1976; Crystal and House, 1987). In the present study, therefore, the aims were threefold - to analyse global segment and syllable duration behaviour in a database. There is a basic lack of published data for Australian English. Apart from the vowel studies by Cochrane, (1970); Bernard (1967,1970), and Bernard and Mannell(1986), few durational analyses have been undertaken for segment durations in Australian English. Our final aim was to test the Campbell model of syllable duration prediction on the database.

METHOD AND MATERIALS

The database in this study consisted of 498 phonetically balanced and dense sentences (17,638 phonemes) from the SCRIBE (Spoken recordings in British English) corpus. The sentences were read by one male speaker of Australian English. They were recorded as part of the SHLRC-ANDOSL (Australian national database of spoken language) database project. A broad phonemic transcription of the database was performed. The *mu+* speech database analysis system (see Croot et al., this volume) was then used to group related segments into words, rhythmic feet and syllables. Information relating to whether the syllable is pitch-accented or is nuclear accented was included in the transcription, as well as boundaries at minor intonation breaks (intermediate phrases) and major intonation breaks (intonational phrases). An intonational transcription of the database was performed after Pierrehumbert (1981). Additional labelling functions were applied to classify each word as either a content or function word. The mean and standard deviation values for each segment type and syllable were calculated across the database in order to establish a yardstick for subsequent analyses. *mu+* routines were also written to extract vowel and consonant durations that fell into the four prosodic categories, pitch accented, final, stressed and unstressed. A similar analysis was performed on syllable durations.

RESULTS AND DISCUSSION

Segment duration

Table I. Duration of consonants (ms), standard deviation values and number of observations

	SHLRC			MIT			KLATT
	DUR	SD	n	DUR	SD	n	DUR
b	106	28	(71)	90	26	(91)	80
d	109	35	(35)	86	34	(66)	65
g	82	31	(32)	96	30	(56)	65
p	133	33	(69)	129	34	(73)	85
t	135	28	(66)	131	30	(60)	65
k	133	22	(91)	128	31	(83)	65
m	92	25	(54)	81	19	(59)	70
n	85	24	(39)	72	18	(51)	65
N	-	-	-	-	-	-	80
f	125	25	(66)	117	20	(87)	120
T	121	40	(17)	130	37	(16)	110
s	135	31	(139)	127	19	(99)	125
S	154	26	(18)	127	17	(41)	125
v	80	26	(14)	82	21	(12)	60
D	54	28	(12)	60	7	(2)	50
z	103	54	(78)	77	28	(3)	75
h	90	27	(58)	81	22	(30)	80
w	96	27	(50)	87	25	(98)	80
j	89	22	(12)	74	34	(16)	80
r	87	22	(27)	80	25	(83)	80
l	87	24	(44)	74	18	(88)	80

Table II. Duration of vowels (ms), standard deviation values and number of observations

	Bernard and Mannell (1986)	SHLRC Pitch accented vowels	SHLRC Final vowels	SHLRC Stressed Vowels
	DUR SD n	DUR SD n	DUR SD n	DUR SD n
i:	256 38 (170)	124 52 (116)	144 50 (74)	79 27 (471)
I	136 29 (169)	61 17 (163)	68 22 (33)	53 18 (539)
E	159 28 (170)	83 27 (183)	123 35 (24)	73 22 (194)
A	201 35 (170)	120 36 (150)	161 51 (17)	98 30 (153)
a:	315 41 (169)	175 50 (74)	224 56 (14)	145 32 (71)
V	165 29 (168)	87 24 (108)	119 32 (22)	83 21 (126)
O	182 28 (169)	105 32 (123)	136 41 (12)	89 24 (131)
o:	295 40 (169)	140 51 (107)	184 61 (20)	114 36 (137)
U	155 28 (169)	65 20 (15)	81 16 (2)	59 19 (43)
u:	268 40 (166)	112 48 (93)	151 60 (19)	73 31 (166)
@:	283 40 (167)	141 30 (62)	177 33 (12)	114 30 (49)
ei	303 39 (171)	145 40 (116)	195 55 (23)	128 34 (184)
ai	314 43 (170)	172 51 (106)	237 67 (21)	135 36 (214)
oi	285 39 (170)	170 53 (21)	178 43 (12)	156 37 (28)
au	317 41 (168)	174 52 (43)	217 66 (10)	142 30 (74)
ou	289 36 (171)	146 41 (76)	172 52 (29)	124 32 (138)
I@	282 40 (166)	180 77 (17)	270 52 (12)	118 34 (17)
E@	273 44 (166)	164 75 (27)	271 29 (6)	110 40 (38)
U@	265 42 (146)	-- -- --	-- -- --	-- -- --
@	-- -- --	-- -- --	80 42 (150)	-- -- --

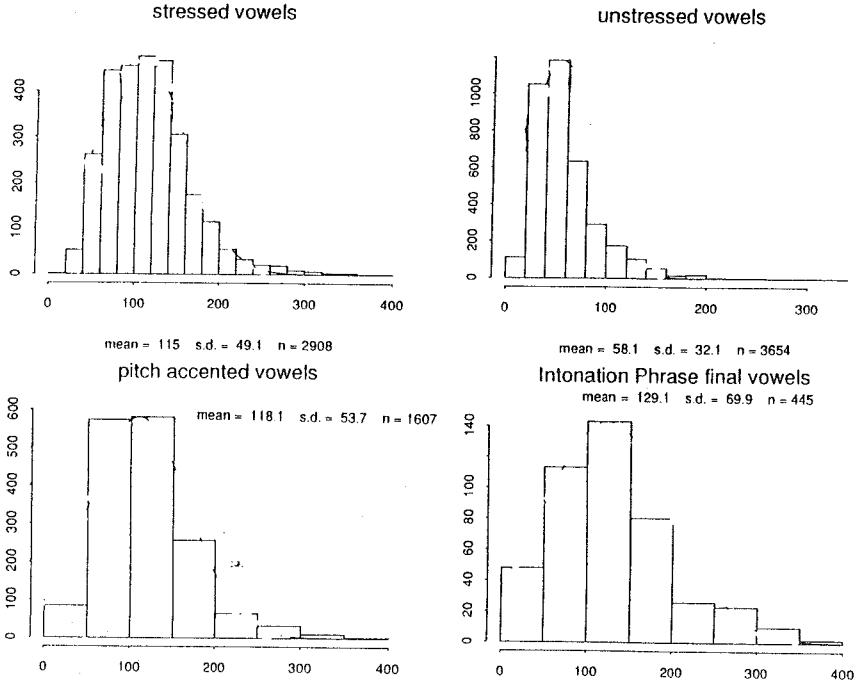

Figure 1. Durations of vowels (ms) in four different prosodic contexts

As part of the preliminary analysis of segment duration in the database, we calculated inherent durations of consonants. Following Carlson and Granstrom (1986) consonant durations were calculated for every consonant n a /C#'CV/ context, ie. in word initial pitch accented syllables for consonants. The results of this analysis are illustrated in Table I. As a general rule the data (Table I) show a similar basic durational structure to that observed for American English by Klatt(1979) and Carlson and Granstrom (1986).

The results of vowel duration analysis are shown in Table II and Figure 1. Data from Bernard and Mannell's (1986) reanalysis of Bernard's earlier studies are also listed in Table I. The data illustrated in Table II shows there is a notable disparity between the overall duration range of Bernard's data and the comparable set of results from the data base (final vowels). The duration values in the Bernard study are much longer than those extracted from the data base. This difference reflects the nature of the two corpora (restricted versus broad phonetic context). The influence of post-vocalic voicing on the vowel nuclei is also apparent in the /h_d/ tokens. Furthermore, the data in the present study were analysed from a single as opposed to several speakers. These differences aside, a similar patterning is observed with respect to inherent vowel duration differences. For example, tense vowels are generally longer than lax vowels across categories, although these distinctions are most evident in prosodic phrase-final contexts.

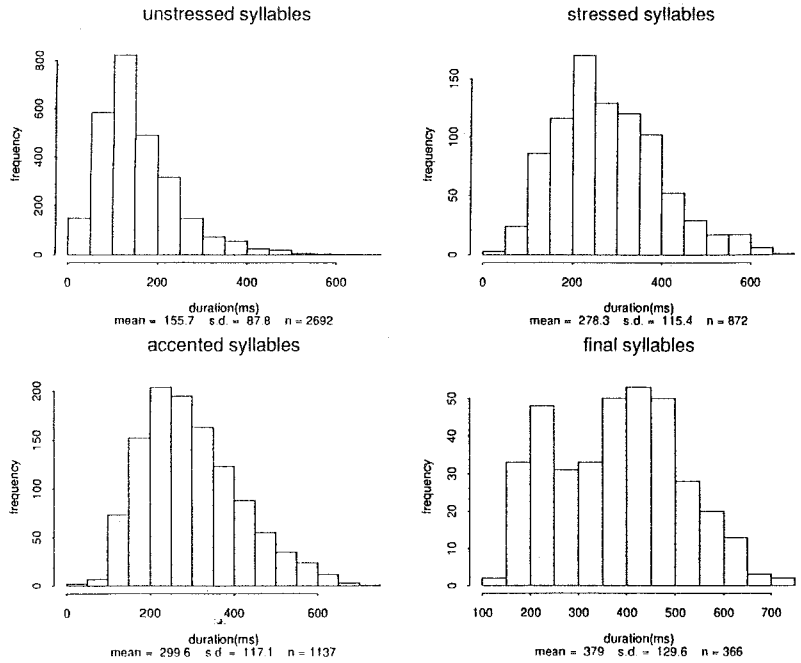

Figure 2. Durations (ms) of unstressed, stressed, pitch accented and phrase-final syllables.

In keeping with earlier studies on other dialects of English(e.g. Cooper and Danly (1981), final accented vowels in the data base are generally longer than pitch accented vowels, or unaccented vowels (figure 1). These differences are statistically significant at $p<.01$. Final vowels are also the most variable as a class, with the widest range of standard deviations. Pitch accented vowels have slightly lower standard deviations, presumably because of their greater frequency in the data base.

Syllable duration

Syllables were first analysed as a group. The overall mean duration was 202ms with a standard deviation of 118ms. These results compare very well with syllable duration data in British English described by Campbell (1989). The mean duration of syllables in his corpus was 202ms, with a standard deviation of 127ms. Figure 2 shows syllable durations for four prosodic categories - unstressed, stressed, pitch accented and intonational phrase final syllables. Predictably, final syllables are by far the longest in this corpus, and unstressed syllables the shortest.

The model

A three-layer neural net (Campbell, personal communication) was trained on the syllable durations from the syllable database. The neural net in this study incorporated 14 input nodes (one for each feature value) 7 hidden nodes and one output node. Our version of Campbell's original model takes the following factors into account: 1) number of phonemes in broad transcription of a syllable; 2) nature of syllabic peak - whether it is a tense or lax vowel, a diphthong, or a sonorant consonant; 3)position of syllable in phrase; 4) degree of prominence of syllable (i.e. whether it is unstressed, stressed or pitch accented; 5) function/content role of the word in which the syllable occurs.

The model was trained on 2488 syllables and subsequently tested on a further 2000 syllables. Regression analysis of the two data sets - the original and test sets - produced a correlation of .78. Analyses of the original and training data sets produced a correlation of .81. The model was subsequently tested on a further 2000 syllables. Regression analysis of the two datasets - original and test revealed a correlation of .78, with an offset of 41ms and a slope of .75. - original and training - correlation is .81, an intercept of .41 and a slope of .77.

In keeping with Campbell's (1989) study, the syllable data were converted to log durations. Apparently, log durations reflect a better overall distribution than raw duration data. The following results were obtained. Regression analysis yielded a correlation of .8 for the training set of log duration data and an intercept of 34.8, and slope of .78. Analysis on the test set resulted in a correlation of .78, intercept of 35.9 and slope of .76. As the log conversion did not improve on the original result, it was decided to model the data in raw durations (ms).

As the second step in the modelling procedure, segment "elasticities" were calculated for each consonant and vowel in the database using the following formulae:

$$\Delta = \sum_{i=1}^{n} (\mu i + k\sigma i)$$

$$\Delta = \sum_{i=1}^{n} (\mu i + 0.75^{(n-i)} k\sigma i)$$

Δ represents the predicted syllable duration given by the net, μ and σ the mean and standard deviation values of the particular segment type. At a first pass, formula 1 yielded acceptable durations in unstressed syllables but not in pitch accented or phrase final syllables. Subsequent application of Campbell's modified formula (2) in which lengthening in final syllables was taken into account, yielded closer fit with observed segment durations.

CONCLUSION

The segment duration data reported in this study support many of the results of earlier traditional vowel duration studies of Australian English, although the magnitude of durations for vowels in particular are somewhat reduced in range in the general speech data base analysed in this study. Moreover, final lengthening and accentuation influence both segment and syllable durations this corpus, in keeping with other dialects of English. However, there is an obvious need to investigate these timing effects across a larger database consisting of several speakers. This analysis is currently under way.

The results of syllable-modelling in the database lend support to Campbell's original hypothesis that a large proportion of the durational variance (although not all) can be accounted for by a relatively simple syllable-level view of speech timing. In other words explicit modelling of traditional segmental effects, such as shortening in consonant clusters or vowel aperture related length differences noted in the above section on segment duration, may not necessarily yield a better durational "fit" than a relatively simple model of syllable timing.

REFERENCES

Bernard, J.R. (1970): On nucleus component durations. *Language and Speech*, **13**, 37-58.

Bernard, J.R. and Mannell, R.H. (1986): A study of /h-d/ words in Australian English. *Speech, Hearing and Language Research Centre Working Papers*.

Campbell, W.N. and Isard, D.(1991): Segment durations in a syllable frame. *Journal of Phonetics*, 19, 37-47.

Carlson, R. and Granstrom, B. (1975): A phonetically oriented programming language for rule description of speech. In G. Fant (ed.) *Speech Communication*, vol. 2 Stockholm: Amqvist & Wiksell, 245-253.

Carlson, R. and Granstrom, B. (1986): A search for durational rules in a real-speech data base. *Phonetica*, **43**, 140-154.

Cochrane, G.R. (1970): Some vowel durations in Australian English. *Phonetica*, **22**. 240-250.

Cooper, W. and Danly, M. (1981): Segmental and temporal aspects of utterance-final lengthening. *Phonetica*, 38, 106-116.

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

Klatt, D. (1976): Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59, 1208-1221.

Klatt, D. (1979): Synthesis-by-rule of segment durations in English sentences. In *Frontiers of Speech Communication Research*. Eds. B. Lindblom and S. Ohman. New York: Academic, 287-300.