SPEECH-RATE VARIATION AND THE PREDICTION OF DURATION

W. N. CAMPBELL

IBM (UK) Scientific Centre . Winchester . England

ABSTRACT

A comparison between the output from a set of duration rules based on Klatt '76 and measured durations in a text allows quantification of speech rate at a local as well as a global level. The rules account for known correlates of duration change, such as stress, phonetic and phrasal context, and inherent differences in the durations of each segment, but make no allowance for local changes of rate within a text. The degree of fit of the output from such a system to the observed durations in the text provides a guide both to the accuracy of the rule-set and to the rate-related variation within that text. Statistical procedures can be applied to reduce the rule-related error and thereby strengthen both the predictions of the rules and the quantification of the rate variation. This paper describes research in progress.

INTRODUCTION

Speech rate is known to be a variable affecting timing in a speech signal, but one that is difficult to quantify. Absolute measures of duration in a text tell little about the relative lengths of segments, and account must be taken of all other factors involved if relative values such as 'long', 'short', 'fast', or 'slow' are to be applied.

Simple measures of speech rate, such as 'words-per-minute', and 'syllables-per-second' account well for variation at a global level, but are inadequate to describe local changes in rate, due to the effects of differences in the structure of words and syllables. Words can be mono- or poly-syllabic, and syllables themselves can vary greatly in the number and type of segments occurring in onset, peak and coda positions. A measure of a small number of words or syllables, expressed as a rate in counts per unit of time, will be affected by the complexity of the component units, structure of syllables or syllabicity of words, such that a string of simple units will yield a higher rate than the same number of more complex ones. This effect is reduced somewhat as the number of units increases and a more balanced distribution occurs, but will always be present as a corrupting factor in the accuracy of the rate measurement. It is likely that text type, with stylistic differences in lexical choice, will be a strong determiner of 'rate' in measures such as these, and a more text-independent method is required

Segments would seem to be a better unit for such measurement, but as yet there is no satisfactory method of determining segment boundaries for automated measurement, and the number of decisions required for measurement by hand of a passage of text long enough to provide statistically adequate results would be both uneconomical and error-prone. In the present study, a compromise is reached in the choice of syllables as basic unit, so boundary decisions are reduced and enough measurements can be taken to allow statistically valid conclusions to be made. A method of normalising for differences in syllable structure is proposed.

THE DATABASE

A database of five thousand syllables was prepared from recordings in the Spoken English Corpus [1] which have been prosodically transcribed, tagged for part of speech and punctuated. These were measured for duration and transcribed phonemically. Samples chosen were one long text, a twenty-minute broadcast of a short story by Doris Lessing, read by Elizabeth Bell, of approximately four thousand syllables, and two shorter texts of approximately five hundred syllables each, one Open University lecture on philosophy, and one news extract, for cross-checking.

Syllables were measured in milliseconds from recordings digitised at 10k Hz (4.5k lowpass filtered) with the IBM UKSC SAY speech analyser [2, 3] using interactive graphic display at one-thousand samples per screen-width, and simultaneous auditory replay of the waveform. Hard copy of both the waveform and gain plots were retained for reference purposes. In the case of ambisyllabicity, the clearest boundary in the acoustic waveform was selected, and the phonemic transcription, later to be used as input to the rule system, marked accordingly.

FITTING THE RULES

'The rules operate within the framework of a model of durational behaviour which states that (a) each rule tries to effect a percentage increase or decrease in the duration of a segment, but (b) segments cannot be compressed shorter than a certain minimum duration. The model is summarised by the formula

$$DUR = [(INHDUR - MINDUR)*PRCNT]/100 + MINDUR$$

where INHDUR is the inherent segment duration in ms, MINDUR is the minimum duration of a segment if stressed and PRCNT is the percentage shortening determined by the rules.' (D. H. Klatt [4])

An iterative process was used to match the rule set (originally designed for American English, based on the durations of a single male speaker, and taken from CVC words in frame sentences uttered in a controlled environment) to the durations required for the prediction of British English and for this particular speaker-text pair. The phonemic transcription of the test text was used as input to a computer implementation of the Klatt [5] rules for duration prediction, and the resulting segment values summed to the syllable level. These were compared with the measured durations according to the factors underlying the rules, which were in turn adjusted accordingly. The same input was passed through the improved rules and the process repeated until the output stabilised.

Segment durations were first adjusted, by sorting 'fit' for each syllable, expressed as a percentage of predicted duration to observed, by nature of the segments appearing. Thus, /f/ for example, although assigned an inherent duration of 120ms and a minimum of 60ms in the Klatt rules, was found to be appearing in syllables that were consistently overpredicted, and by reducing its inherent duration in the rules to 95ms, and its minimum to 50ms, a better overall fit was observed. An exact fit is not to be expected since the rules make no allowance for speech-rate variation, other than offering a single variable ('PRCNT') that can be reset to change the overall rate of duration. The variance observed in the fit for any individual factor will never be reduced below the variance of the underlying speech rate changes, but can only be minimised.

The original rules assume that the minimum duration of an unstressed segment is half that of the segment in a stressed position. On further analysis of segment fit according to stress, it was found that a better prediction could be achieved by specifying absolute minima separately for the two situations; thus /k/ for example, while 65ms and 50ms for inherent and minimum in the Klatt rules, was found to fit better if specified as 65ms and 35ms, with an absolute minimum (for the unstressed position) of 15ms. The full table of final values with the original defaults is shown in Fig 1. These represent an intermediate stage in an iterative process, and are not presented as statements about individual segment durations per se.

With these segment defaults fitted to the sample text, the values specified in the rules for modifying PRCNT were similarly adjusted so that the best fit could be obtained. In summary, clause and phrase medial syllables were found to be overpredicted, and both initial and final syllables underpredicted; clause final syllables considerably so. An extra rule was included to cover the case of phrase-initial syllables, which are not accessible through the framework of the original rules [6].

QUANTIFYING SPEECH-RATE

With the rules matched to the text at a global level through statistical analysis of averaged results, differences in output can be examined at the local level. Since there is no speech-rate information in the rule-set, differences will contain a quantification this, contaminated by noise from measurement and prediction error.

There will inevitably be a certain amount of error in hand-measurement of several thousand syllables, no matter how precise the equipment, but since the totals are cumulative, and sums can be simply checked against overall durations for stretches of the text, it can be assumed that the majority of errors will lie in boundary determination. These can be overcome by smoothing with a three-syllable moving-average window since any over- or under-measurement in an individual syllable should be compensated by a corresponding under- or over-measurement of its immediate neighbours.

Errors in prediction will be systematic by definition, and therefore susceptible to detection by statistical methods. They can be determined by examining the measures of fit according to criteria not included in the rule-set and implementing new rules to cover any regularities found.

	inh	m+	m-	orig	min		inh	m+	m-	orig	min
i	165	85	25	160	50	eɪ	220	110	35	190	70
ɪ	130	60	25	130	40	aɪ	250	115	60	250	90
ɛ	165	60	30	150	60	ɔɪ	220	110	35	280	110
æ	205	100	35	230	60	əʊ	220	110	25	220	70
ʌ	185	75	45	140	50	ɑʊ	220	110	50	260	100
ɑ	265	125	55	240	80	ɪə	235	110	50	260	100
ɒ	190	110	35	240	80	ɛə	270	100	50	270	100
ɔ	235	100	55	240	100	ɔə	230	100	50	230	100
u	185	110	30	210	60	tʃ	100	75	30	70	50
ʊ	105	40	30	160	50	dʒ	95	70	20	70	50
ə	100	55	25	120	40	p	70	40	20	85	50
ɜ	170	75	20	180	60	t	60	30	15	65	40
h	55	40	10	80	20	k	65	35	15	65	55
m	80	45	15	70	60	b	70	40	15	80	50
ɱ	120	60	25	170	110	d	70	40	15	65	40
n	90	40	15	65	35	g	70	40	15	65	50
ŋ	90	50	25	80	50	f	95	50	25	120	60
ṇ	170	80	50	170	100	v	65	50	20	60	40
l	85	45	25	80	40	θ	75	35	25	110	40
ɫ	85	55	15	90	70	ð	55	25	10	50	30
ḷ	180	90	45	160	110	s	120	50	30	125	50
r	80	40	25	80	30	ʃ	105	55	25	125	50
j	85	50	10	80	40	z	80	35	15	75	40
w	85	30	30	80	60	ʒ	90	35	20	70	40

Figure 1. Duration values (ms) used for British English text: Default inherent, stressed and unstressed minimum durations for modified Klatt rules, with originals.

The quantification of speech-rate is thus not a single, simple process, but an iterative one, with accuracy (and therefore confidence) increasing at each iteration. It can be expressed as a ratio ('SPRATE') of predicted rate in syllables/second to observed rate in syllables/second calculated from smoothed data. The above objection to syllables per second as a measure of speech-rate is overcome by comparing like with like in the present method. Thus

SPRATE(%) = (SMOOTHED PREDICTED
RATE/SMOOTHED OBSERVED RATE) * 100

PRELIMINARY RESULTS

At the current iteration, sprate mean is 100.2% for 3959 syllables [7], indicating an almost exact overall fit between the predicted and observed durations, but with a standard deviation of 19.18 that is partly accounted for by the lack of rate information. Of the other factors contributing to this variation, no significant effects could be found for e.g. the type of syllable structure, the position of the syllable in the word, or the position of that word in the phrase or clause. Part of speech, however, appeared to be a significant factor, with sprate results for selected categories as below,

syllable type	mean	s.d.(est)	n
lexical verbs	102.7	0.7	626
nouns	98.5	1.0	750
adjectives	92.9	1.2	339
adverbs	102.6	1.6	184

which shows that while verbs and adverbs are slightly overpredicted by the rules, nouns and especially adjectives would generally appear to be spoken more slowly than the rules predict.

An earlier iteration showed that polysyllabicity, instead of being in the domain of the word as the original rules predict, gives a better fit if measured in feet, and adjustments made according to the number of the unstressed syllables that follow each stressed syllable.

A category that needs further examination is that of stressed but unaccented syllables which are 'prominent but have no pitch movement' [8]. By default, these are treated as stressed, but on examination of the results, sprate, which is 100.3 for unstressed syllables and 98.3 for stressed, is 106.7 for stressed-but-unaccented, showing slight underprediction of stressed syllables, but greater overprediction of the intermediate category.

Of perhaps greater interest though, is the fit of sprate to the perceived speeding up and slowing down in the presentation of the text by the reader. Taking the mean sprate values for all syllables in the tone-group, we find the following

93.9 Walking down the path with her, he blurted out
112.1 'I'd like to go and have a look at those rocks down there.'
86.4 She gave the idea her attention.

102.8 The water was pushing him up against the roof.
98.7 The roof was sharp and pained his back.
120.2 He pulled himself along with his hands, fast, fast.
97.8 and used his legs as levers.

99.8 They sat down to lunch together.
128.2 'Mummy, I can stay under water for two minutes, three minutes at least.'
84.1 It came blurting out of him.

where an increased sprate indicates overprediction in the rules or, conversely, a speeding up in the text. Here, examples have had to be chosen to include textual clues to the rate, but listening to longer passages confirms that rate correlates well with sprate. Further iterations will allow more confident examination at levels lower than the sentence.

DISCUSSION

As Fig 2. suggests, a small random or high frequency error summed with a more slowly changing effect does little to hide its rhythms. Speech rate cannot be expressed as a simple sine wave, but its effect on the prediction of segment duration by rule can perhaps be seen in this way and until its processes are understood, no predicted durations can match observations from a real text - the lows cannot be slow enough nor the highs fast enough. Until rate information is superimposed on phonetic and phrase-level information in a systematic manner, the output will be flat and if in a computer text-to-speech system, 'robotic'.

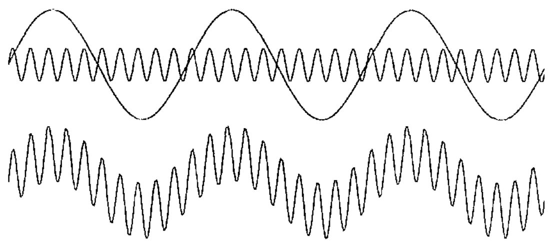
Figure 2. Interacting datasets.

The above method provides a quantification of speech rate that reveals both local and wider-range domains. Being an iterative process, it provides for an improvement of the rules for prediction of duration in a text while at the same time revealing processes within the text that govern changes in rate at the more local level.

REFERENCES

[1] *Spoken English Corpus* : Lancaster University and IBM UKSC.

[2] *IBM UKSC Reports* No.135, June 1985, and No.145, January 1986.

[3] W N Campbell : *A Search for Higher-level duration rules in a Real-Speech Corpus.* Proc Conf Speech Tech Edinburgh 1987

[4] D H Klatt : *Synthesis by rule of Segmental Durations in English Sentences* in Frontiers of Speech Communication Research edited by Lindblom & Öhman, Academic Press 1979 (pp 287-299).

[5] D H Klatt : *Linguistic uses of segmental duration in English* pp 1208-1221, JASA 59 1976.

[6] W N Campbell : *Extracting Speech-Rate Values from a Real-Speech Database.* ICASSP 1988 (forthcoming)

[7] *Glim 3.77* update 1 (copyright) 1985 Royal Statistical Society, London

[8] L G Taylor & G Knowles : *Manual of Information to Accompany the SEC Corpus.* UCREL University of Lancaster 1988