Segmental elasticity and timing in Japanese speech

W. N. Campbell

ATR Interpreting Telephony Research Laboratories

April 17, 1991

Abstract

Variation of duration in natural speech can be largely accounted for by consideration of the phonetic and structural features of the spoken utterance. However, if such variation was simply a function of the various contexts applying to inherent segmental timings, it could in theory be accounted for by a finer and finer statistical analysis until, in the ultimate case, a value was known for each phone in any context.

However, because speakers can vary their rate of articulation, and because the stress given to a particular phone-sequence can vary widely according to semantic and pragmatic context, it is unlikely that such an approach would be practical for a speech understanding or synthesis system. Higher-level control can be achieved by an understanding of the linguistic processes governing the variations in rate, and a knowledge of how the consequent timing changes are implemented.

The absolute length realised by a segment or syllable can be considered as a contextualisation of the lengthening applied to it in a given environment. The articulatory characteristics of a segment determine its length characteristics, but its linguistic environment determines its lengthening. A technique is described that employs a pure numerical value of lengthening of speech segments, separate from the phone-specific durational features, and results are reported of an analysis of the timing characteristics in a database of 503 Japanese sentences. The study focusses on two points that illustrate the benefit of a linguistic as well as a statistical description of the timing proesses.

1 Introduction

Studies of timing based on measures from an acoustic representation of speech must account for considerable variance in the data. As both Nooteboom and Cutler point out elsewhere in this volume, much of

the variability in speech timing is linguistically determined. However, much of it also results from articulatory constraints on the production of the different speech sounds. In order to separate the two, to allow for the separate study of each, a method of normalisation is proposed.

Differences in the timing of speech sounds can be quantified in raw millisecond terms, as fractions or percentages, or in terms of an observed range of distributions. The first method, though, fails to take account of any inherent differences in the duration of different phone types and is blind to the fact that, for example, an increase in duration of 50 milliseconds will be relatively greater for a vowel that has a typical duration of 50 milliseconds than for one that has a typical duration of 150 milliseconds. Expressing this increase as a fraction affords a relativisation of the difference but still doesn't enable a simple comparison of the lengthening of phones of different types, unless their typical durations happen to be similar.

A better comparison can be achieved by expressing lengthening in terms of standard deviations from a mean, i.e., as z-scores. Thus, raw millisecond durations can be transformed as in (1) into z-scores that will normally be in the range of plus or minus three standard deviations (SD). With this normalised measure of length, comparison can be made not only between phones of different types, but also within each type, as a positive value of z represents lengthening, and a negative value shortening relative to the mean duration observed for all tokens of that type in the database. With normally distributed data, 68% of the tokens can be expected to fall within ± 1 SD, and 99% between ± 3 SD.

$$z - score = raw - duration_{token} - \mu_{type}/\sigma_{type} \quad (1)$$

where μ_{type} is the mean duration observed from all tokens of that phone type,
and σ_{type} is their standard deviation.

In the next section we will examine the suitability of such a measure for speech data, and in following sections will see how it can shed light on the lengthening processes in Japanese speech.

2 Segment distributions in the Japanese database

Data from a single male speaker (a professional broadcaster) reading 503 separate sentences was used. These sentences were chosen from 10,196 Japanese newspaper and magazine sentences to present a phonetically balanced cross-section of the commonly occurring sentence types. Figure 1 shows the range of durations observed for this speaker. Analysis of the segment duration distribution densities shows them not

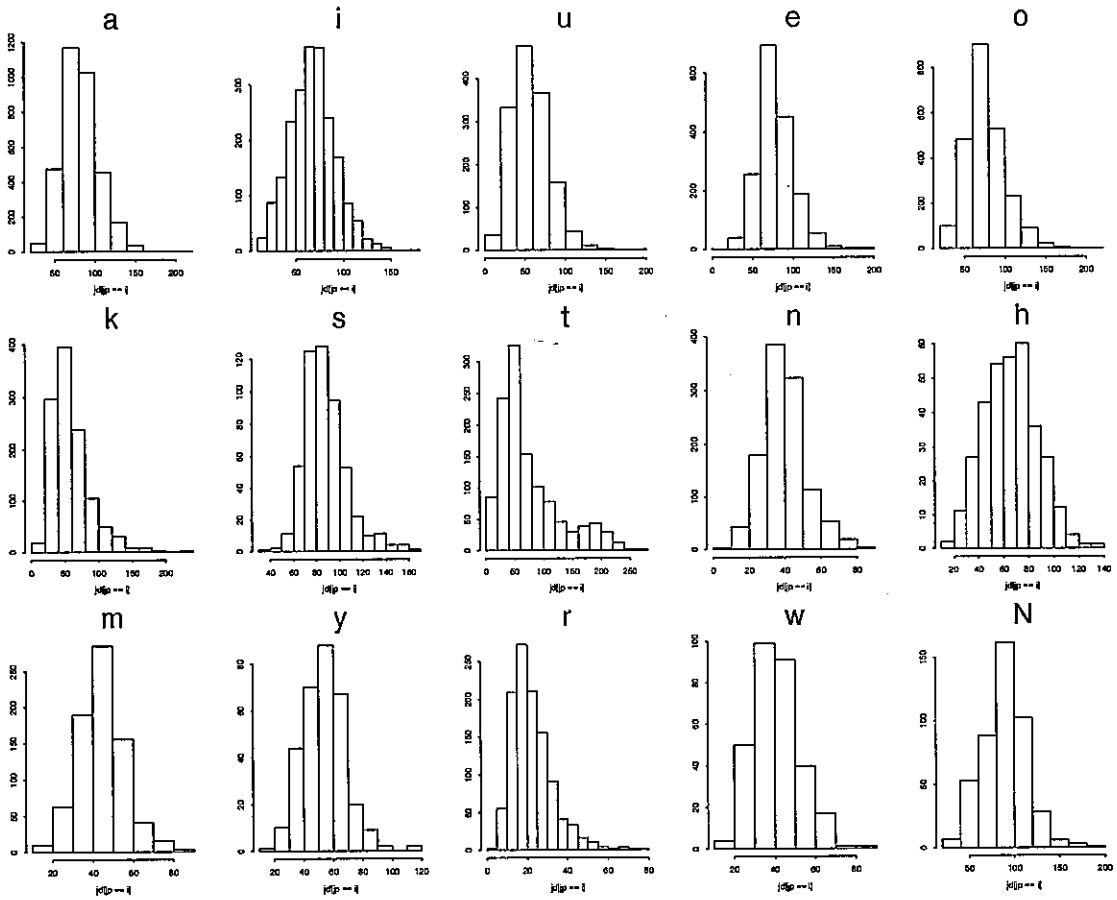

Figure 1: Histograms of the phoneme durations.

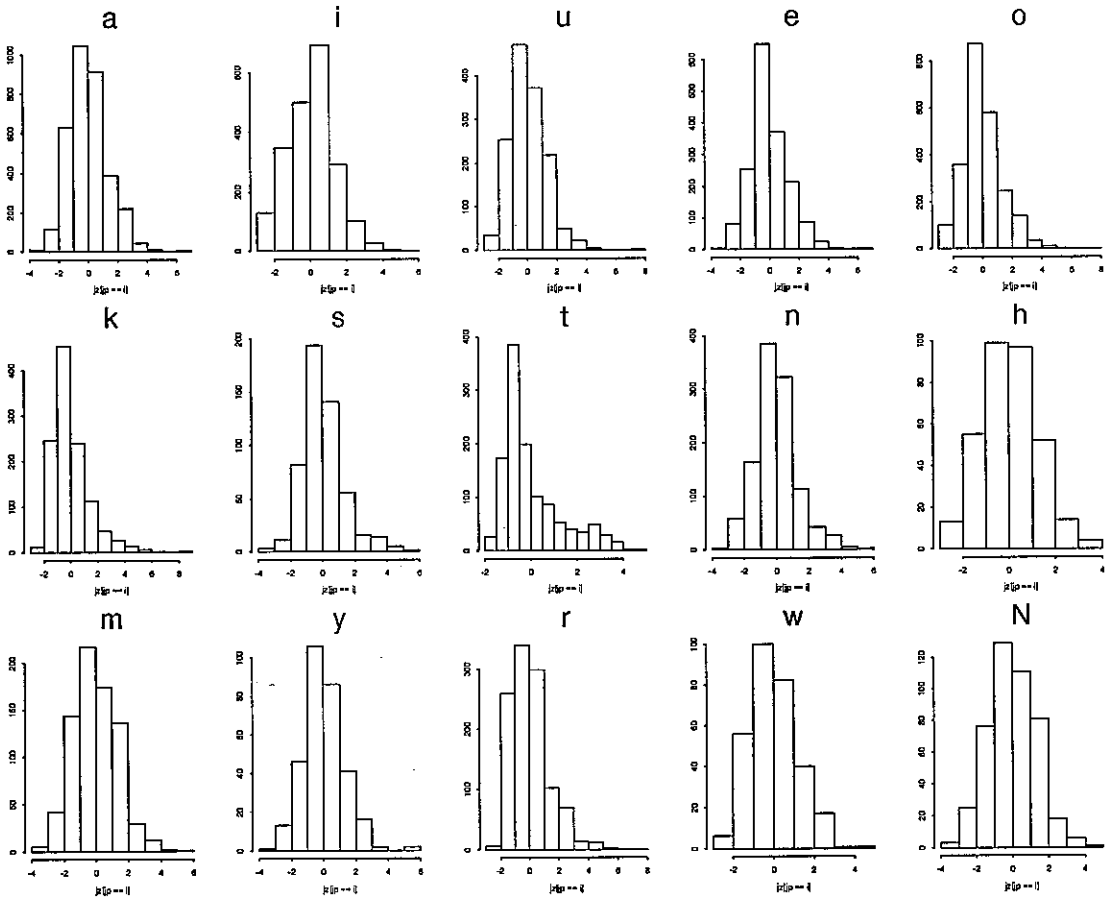

Figure 2: Z-scores for the same phonemes.

in fact to be normally distributed (gaussian), but to be similarly positively skewed. This indicates that greater lengthening is taking place than shortening, perhaps because while no phone's duration can be reduced as far as zero, many can be sustained, albeit in exceptional circumstances, for as long as a second. In fact, durations of longer than a quarter of a second were not found in the current data. What are the repercussions of such skew? Z-scores become shifted in the positive direction, in the range from -4 to +6, (see Figure 2). so that comparison between different extremes becomes less valid. In other words, it is better not to compare degree of lengthening with degree of shortening, but comparisons of different degrees of either lengthening or shortening are not affected. A comparison with the quantiles of a standard normal distribution (Figure 3) shows that with the exception of the plosives, the lines fall similarly close to the main diagonal so approximate normality can be assumed. In the rest of this paper, I shall look in more detail at the distributions of the /a/ and /w/ phonemes.

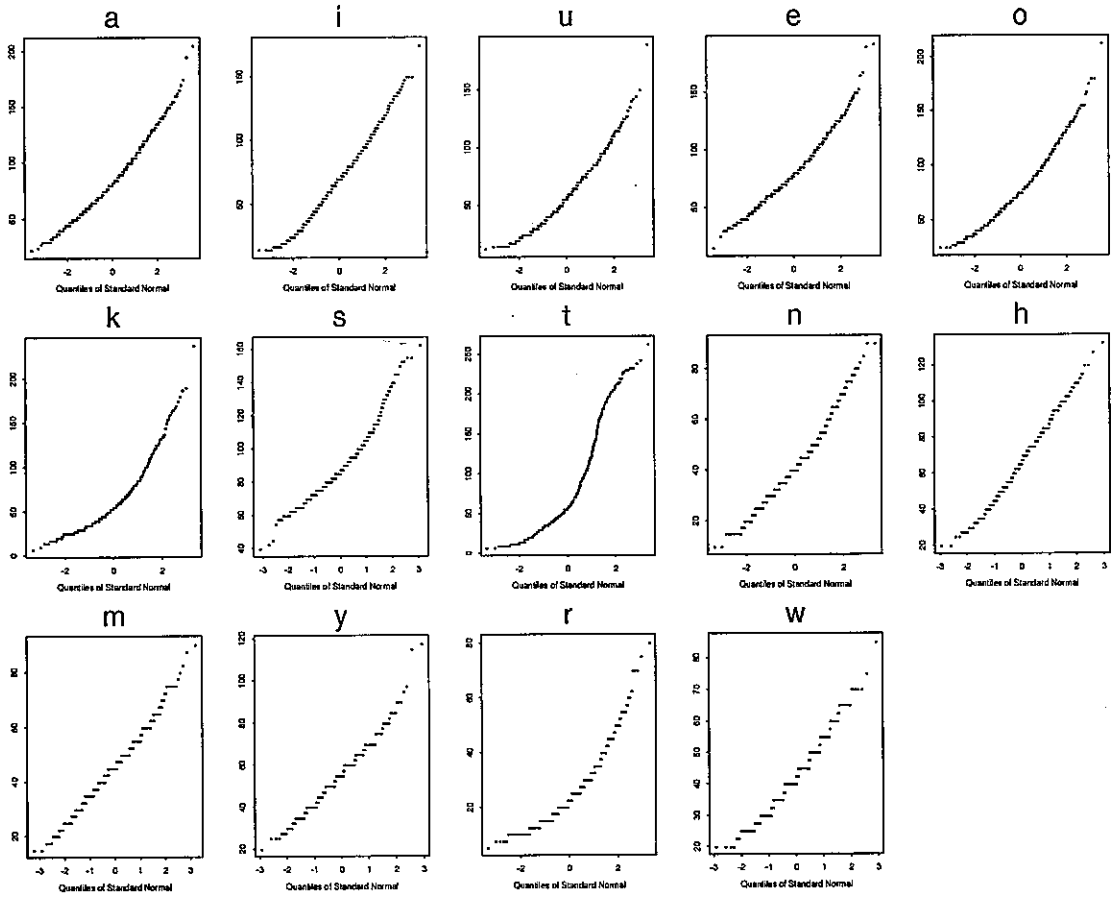

Figure 3: Comparison with quantiles of a standard normal distribution.

Table 1: Average vowel durations

vowel	mean	S.D.	n
/a/	83.86	22.54	3380
/i/	69.82	23.53	2095
/u/	58.32	23.02	1426
/e/	80.04	21.18	1697
/o/	77.71	23.59	2359

(all durations in milliseconds)

Table 2: Average vowel z-scores.

vowel	mean (z)	S.D.	n
/a/	-0.00021597	1.268331	3380
/i/	0.00030548	1.275097	2095
/u/	-0.00039971	1.248374	1426
/e/	-0.00064231	1.294195	1697
/o/	-0.00011021	1.293226	2359

3 The vowel /a/

The mean durations observed for the vowels /a/, /i/, /u/, /e/, /o/ in the database varied from 58 to 83 milliseconds, but their standard deviations were all in the range of 21 to 23 milliseconds. Numbers of tokens ranged from 1426 for /u/ to 3380 for /a/. Table 1 shows the details, and Figure 4 illustrates how this variability in mean duration is factored out by z-score normalisation. It is clear from the boxplots in Figure 4 that the variance in the millisecond durations of the vowels is considerably, though not completely, reduced by the transform. Median z-scores differ (especially for the /i/, which shows much less skew) but their means are all close to zero, as Table 2 shows: The boxes are drawn with horizontal lines indicating the 25th, 50th and 75th percentiles. Vertical lines extend above and below the boxes to one-and-a-half times the upper and lower interquartile ranges respectively. The width of each box is proportional to the log of the number of tokens in that sample, and the notches indicate significance at the 5% level in the difference of the distributions if they show no overlap. The vowel /a/ has been selected for this study not just because it is the most commonly occurring in the corpus, but because it also illustrates some interestiong lengthening characteristics that show the disadvantages of simple statistical modelling of average duration ranges for a given context, as will be shown in the next sections.

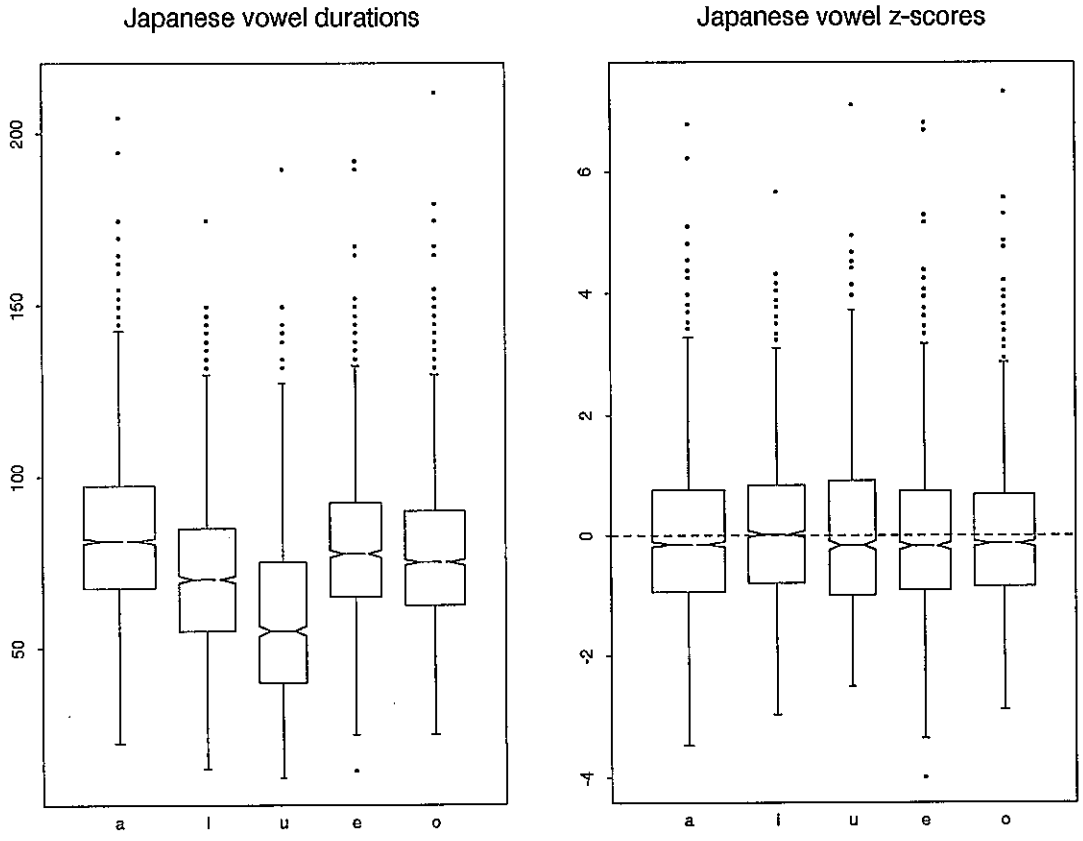

Figure 4: Durations and z-scores for the five Japanese vowels.

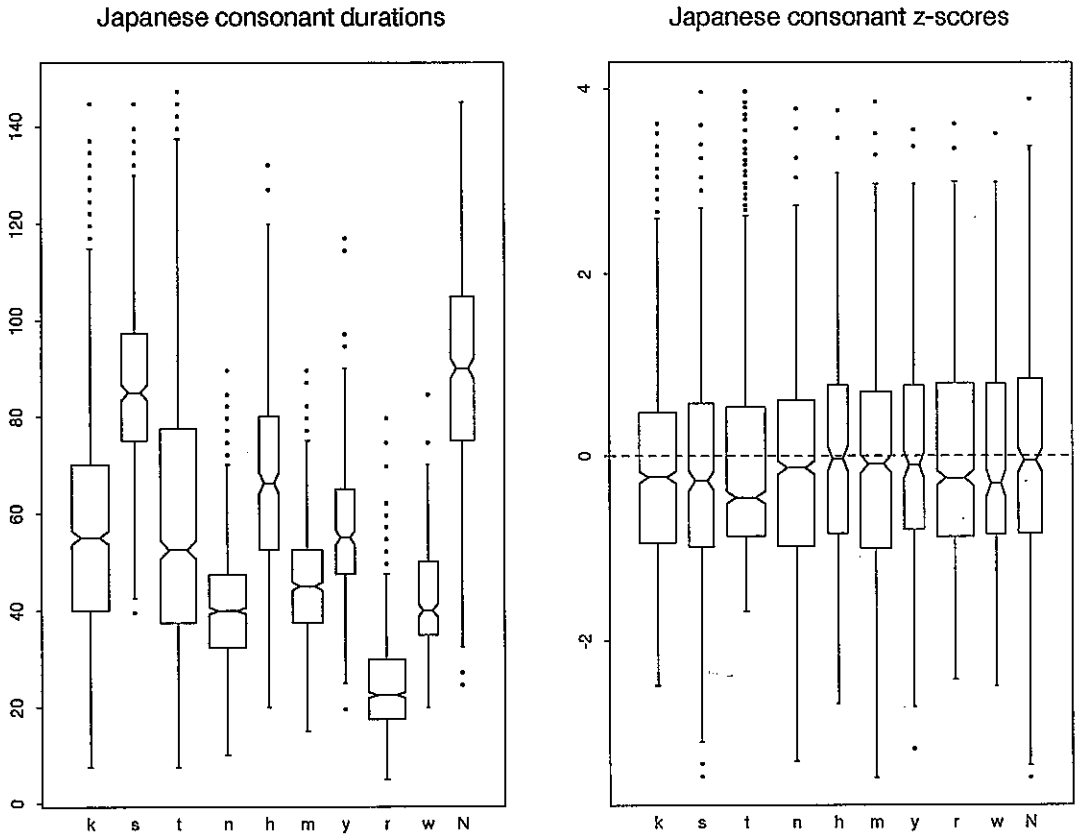

Figure 5: Durations and z-scores for the ten Japanese consonants..

4 The consonant /w/

As with vowels, so with consonants, the normalisation process reduces the segment-specific durational differences to an almost insignificant level (Figure 5). In the case of consonants it can be seen more clearly that much of the variance of the durations is also factored out by the transform. What does this variance signify, and why should it be ignored? Differences in manner and place of articulation are reflected in differences in their typical durations. For example, an alveolar flap, /r/, can be produced by little more than a brief contact of the tongue tip against the alveolar ridge, but a voiceless alveolar fricative, /s/, on the other hand, needs careful adjustment of the pressure of the tongue against the roof and sides of the mouth, and therefore typically takes longer to produce. While these timing effects are of great interest in themselves, they interfere in the visualisation of the effects of higher-level, more linguistically motivated timing processes, and so are factored out.

Table 3: Consonant durations for Japanese.

consonant	mean	S.D.	n
/k/	60.04	28.41	1152
/s/	89.05	18.55	521
/t/	75.32	51.91	1187
/n/	41.38	12.24	1119
/h/	66.71	21.19	334
/m/	45.98	11.35	763
/y/	56.24	14.61	313
/r/	23.98	10.12	1105
/w/	42.84	11.39	303
/N/	90.99	24.38	450

Table 4: Consonant z-scores for Japanese.

consonant	mean (z)	S.D.	n
/k/	-0.00042534	1.33890	1152
/s/	0.00061420	1.31803	521
/t/	-0.00018534	1.27936	1187
/n/	-0.00043789	1.30080	1119
/h/	-0.00020958	1.22431	334
/m/	-0.00020969	1.28968	763
/y/	-0.00015974	1.28181	313
/r/	0.00030769	1.31090	1105
/w/	-9.90099e-05	1.25214	303
/N/	-0.00015555	1.28715	450

Table 5: Verb and Noun phone distributions.

	N	a	aa	ai	b	ch	d	e	ee	f	g	gy	h	hy
verbs:	19	1080	5	8	50	9	231	508	2	13	90	0	53	0
nouns:	362	1119	30	0	210	142	129	620	115	56	246	29	222	27

	i	ii	iy	j	k	ky	m	my	n	ny	o	oo	p	py
verbs:	508	11	8	22	246	2	236	0	268	0	265	64	1	0
nouns:	1098	29	34	244	670	98	320	7	203	26	815	436	90	5

	r	ry	s	sh	t	ts	u	uu	uw	w	y	z
verbs:	501	0	158	43	315	37	571	3	4	20	72	52
nouns:	373	39	256	277	336	87	643	208	18	44	152	87

The choice of /w/ for closer investigation in this paper was motivated by just such a linguistic process. The phoneme /w/ occurs less frequently in the lexical words of Japanese, but commonly in the particle *wa*. Out of 303 instances of /w/, 221 (73%) were in the word *wa*, while only 20 occurred in nouns, 18 in verbs, and the remaining 44 in adjectives, adverbs etc. Section (6) will discuss this in more detail and show the implications that such a biased distribution can have on the statistics of related segments; in this case, /a/. Of the 3380 /a/ tokens, only 753 occurred in particles, while 1030 occurred in verbs, 1119 in nouns, 132 in adverbs, and 121 in adjectives, showing a much more evenly balanced distribution.

5 Part-of-speech & timing

Nouns are by far the most commonly occurring part of speech in the corpus, accounting for 9902 phones, with verbs being second (5475 phones), and particles third (4452 phones).

If we factor segment durations by part of speech (Figure 5), we find that they are not evenly distributed, but show segments in nouns to be longer,, and those in verbs and particles to be shorter. It may be possible to account for this by consideration of the information content of the words (the function/content distinction), with nouns perhaps carrying more meaningful, or less predictable, information than other parts of speech, but let us look more closely at the segmental distributions. There are long as well as short phones in Japanese. These have been excluded from earlier illustrations for reasons of simplicity, but become relevant here. If we look at the distribution of these phones amongst nouns and verbs in the corpus, we find that it is unusual for long segments to occur in verbs, even if the almost two-to-one difference in frequency of these parts of speech is taken into account. Figure 6. shows the durations of these segments and includes the

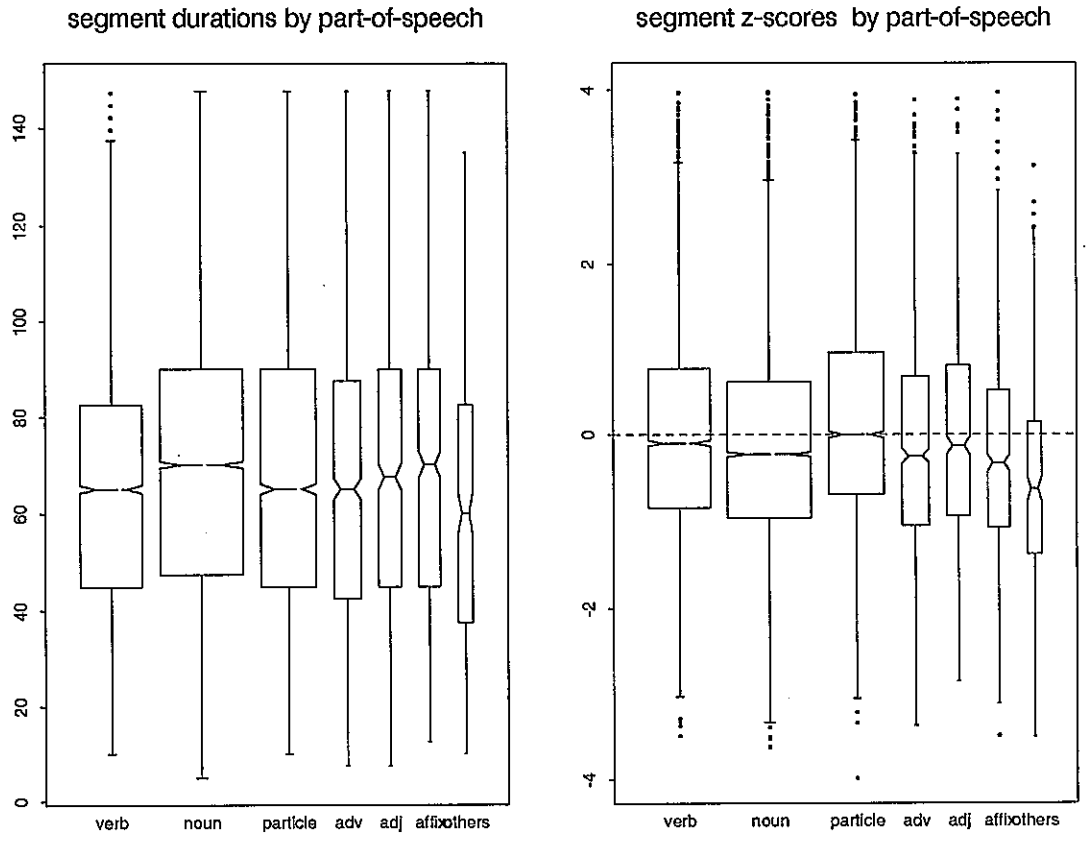

Figure 5: Durations of segments by part-of-speech.

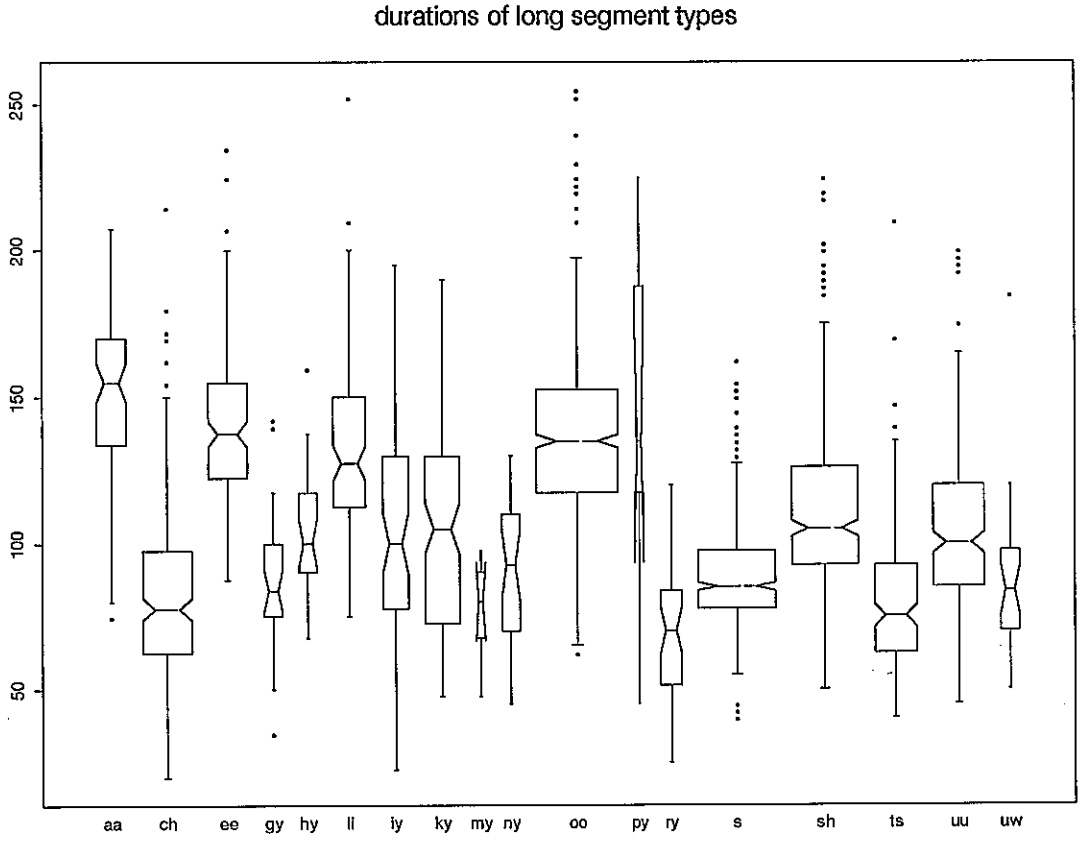

Figure 6: Durations of long segment types.

longest 'short' segment (/s/) for comparison. The mean duration of these 'long' phones is 109 ms (SD = 35 ms, n=2741), compared with a mean of 66 ms for the 'shorter' phones (SD = 29 ms, n=19883). The reasons for such differences in distribution, perhaps related to word frequency, need not concern us here. But the imbalance explains the anomaly shown in Figure 6, of segments in nouns being longer than those in verbs in millisecond terms, but *shorter* in normalised terms. It turns out that segments aren't necessarily longer in nouns, but that nouns have more longer segments. If anything is 'longer', it appears to be the particles. There are at least two possible reasons that would explain the lengthening of particles in Japanese. One is that they occur, by definition, phrase-finally, and are therefore subject to the phrase-final lengthening that has been observed in many languages. The other is that they serve to signal the emphasis of the phrase they govern, and thus signal contrasts within a sentence [Ono 1973, pp. 336-7]. As such, it is the particles that signal much of the 'meaning' of a sentence.

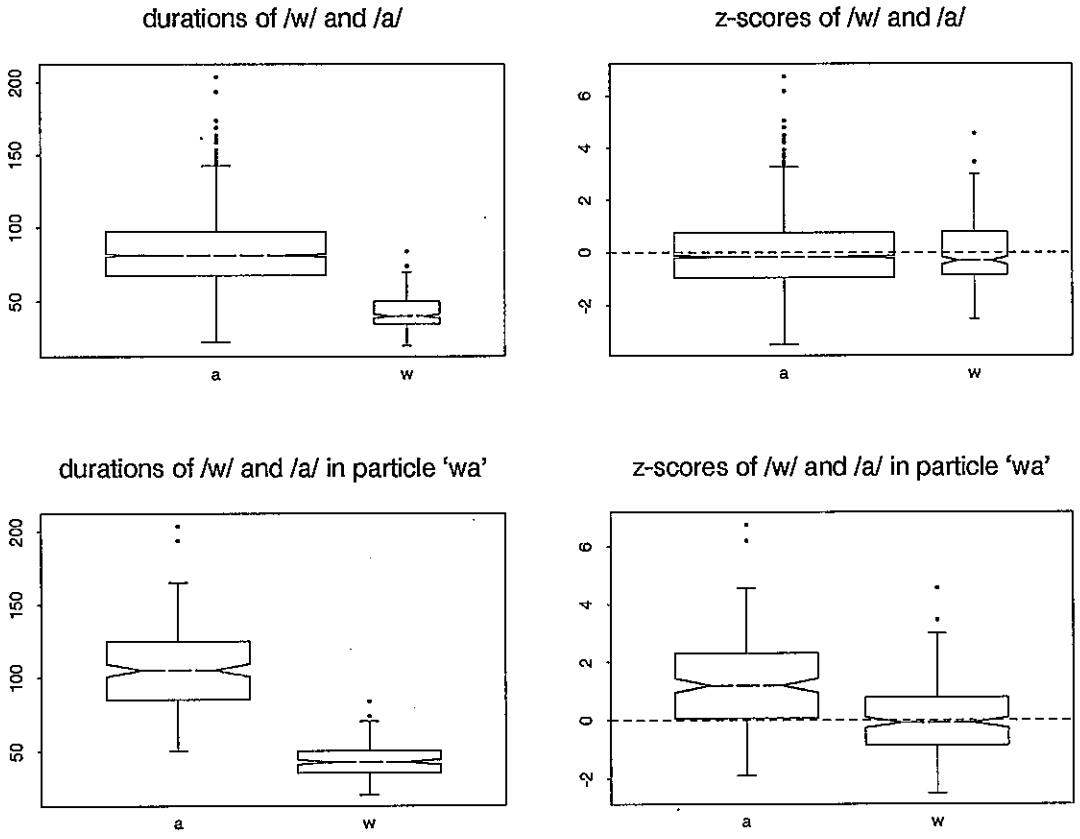

Figure 7: Durations and z-scores for 'wa'.

6 Particles and lengthening

This brings us to *wa*. Figure 7 shows the durations and z-scores for the two phones both independently and in the particle 'wa'. In both cases tha duration of /a/ is longer than that of /w/, as would be expected, but their lengthening appears anomalous. In the general case, both sets of z-scores have a mean of zero and exhibit some positive skew. However, in the 'wa' case, the /a/ appears to be at least one SD longer than the /w/. How can this be explained? The biased distribution of these phones has already been noted. Since by far the majority of /w/ tokens are found in the 'wa' (particle) environment, their 'normal' state will be long, lengthened by dint of being both phrase-final and information-carrying. This lengthening is reflected in the duration of the associated /a/. However, because the most common state of the /w/ phone is long, the mean duration for this phone will be influenced by this lengthening. Other /w/ tokens in less biased contexts will appear 'short' in comparison. If there is an 'inherent' length for a

phone, then that of the /w/ is likely to be less than its statistical 'average' duration.

This situation serves to warn us that the statistical procedures being used to model timing processes in speech only reflect the distributions in the corpus they are applied to, and tell us little about whatever 'inherent' characteristics individual phones may have. Sufficiently careful factoring will control for this, but over-generalisations are likely to result from inadequate analysis.

7 Sentence-final lengthening

Following the trail of the /a/ in this context, we come to the phenomenon of sentence-final shortening in Japanese. This has been reported [Kaiki, Takeda & Sagisaka 1990, pp 18-20] from an analysis of the same speech database, and is in contrast to the more commonly found clause-final lengthening of other languages, and phrase-final lengthening of Japanese. Does Japanese really shorten in sentence-final position, or is this a statistical artifact of the same order as the 'non-lengthening' of the /w/ in 'wa'?

Verbs usually occur sentence-finally in Japanese, and as information-carriers, might be expected to be longer. On the other hand, they may be more-or-less predictable from information earlier in the sentence and thereby forego any content-related lengthening. This is speculation, but an analysis of the phone durations may provide some evidence. Vowels in sentence-final position show interesting lengthening characteristics. Figure 8 shows that only three types predominate; /a/ (n=199) being the most frequent, followed by /u/ (n=94) and /i/ (n=34), with only one occurrence of /e/. This figure does not account for all sentence-final occurrences because some segmentation difficulties resulted in moraic final segments that have been excluded for reasons of simplicity (/ku/(16), /su/(34), /ru/(51), /ai/(10), /ii/(13), etc.). One thing is obvious — sentence-final vowels may show little difference in their raw durations, but the *shortening* undergone by the /a/ in this position is significantly different from that of the other types which appear normally distributed about their mean (/i/), or even lengthened (/u/). What is happening to the /a/?

Looking at the penultimate phones, i.e., the consonants preceeding the /a/, we find that the plosive /t/ predominates (/t/(150), /d/(39), /k/(6), /s/(1)). This is the past-tense verb ending '-ta'. The z-scores show all these phones to be lengthened relative to their means in these sentence-final morae. So why is the /a/ shortened? Three effects may be taking place here: a) the preceeding plosive has a shortening effect on the following vowel, b) the past-tense marker may be informationally redundant because of textual clues (time-marked events) earlier in the sentence, and c) the frequent lengthening already noted for /a/ in particles (not just in 'wa', but also in corresponding 'ga' etc., (n=753))

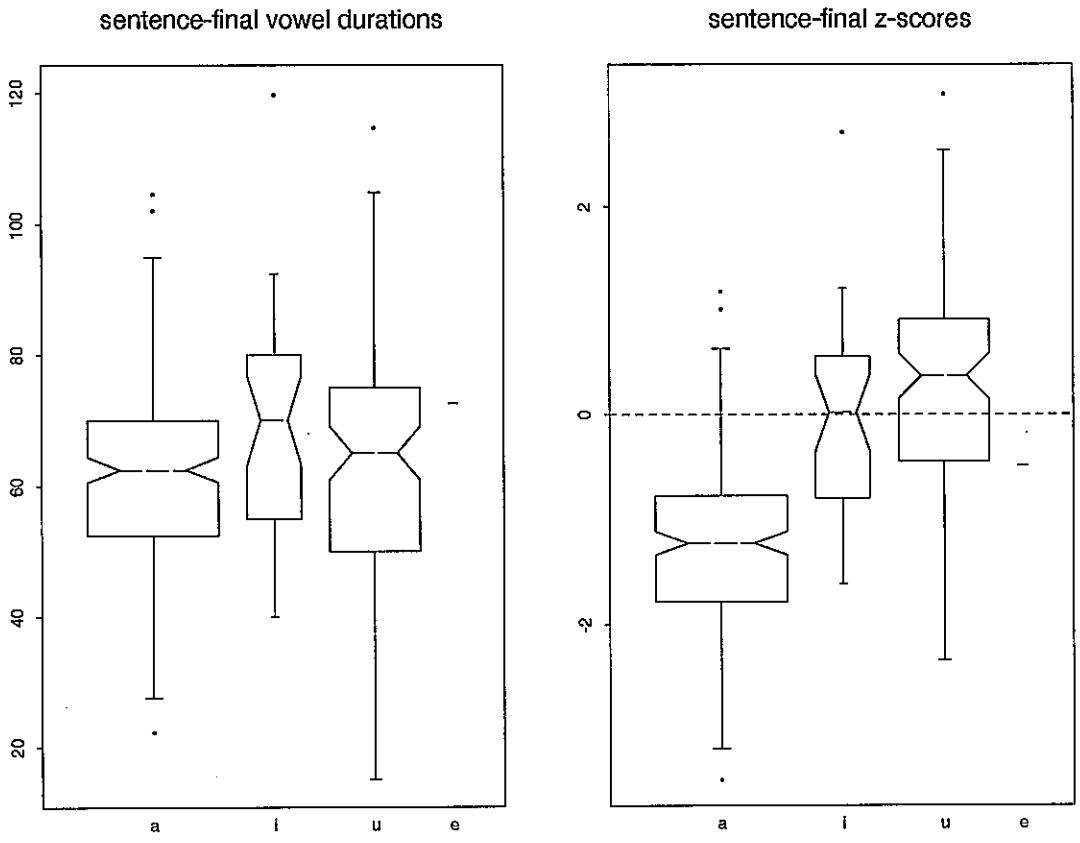

Figure 8: Durations and z-scores for sentence-final vowels.

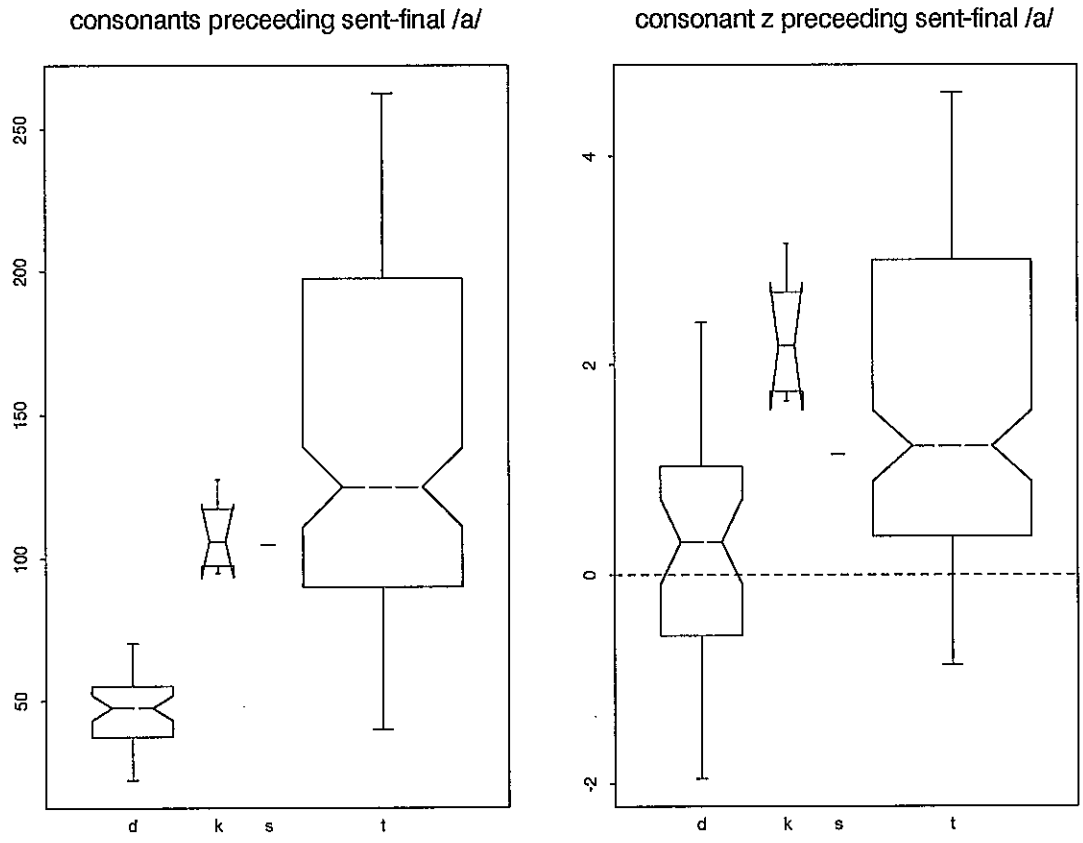

Figure 9: Durations and z-scores for sentence-final consonants before /a/.

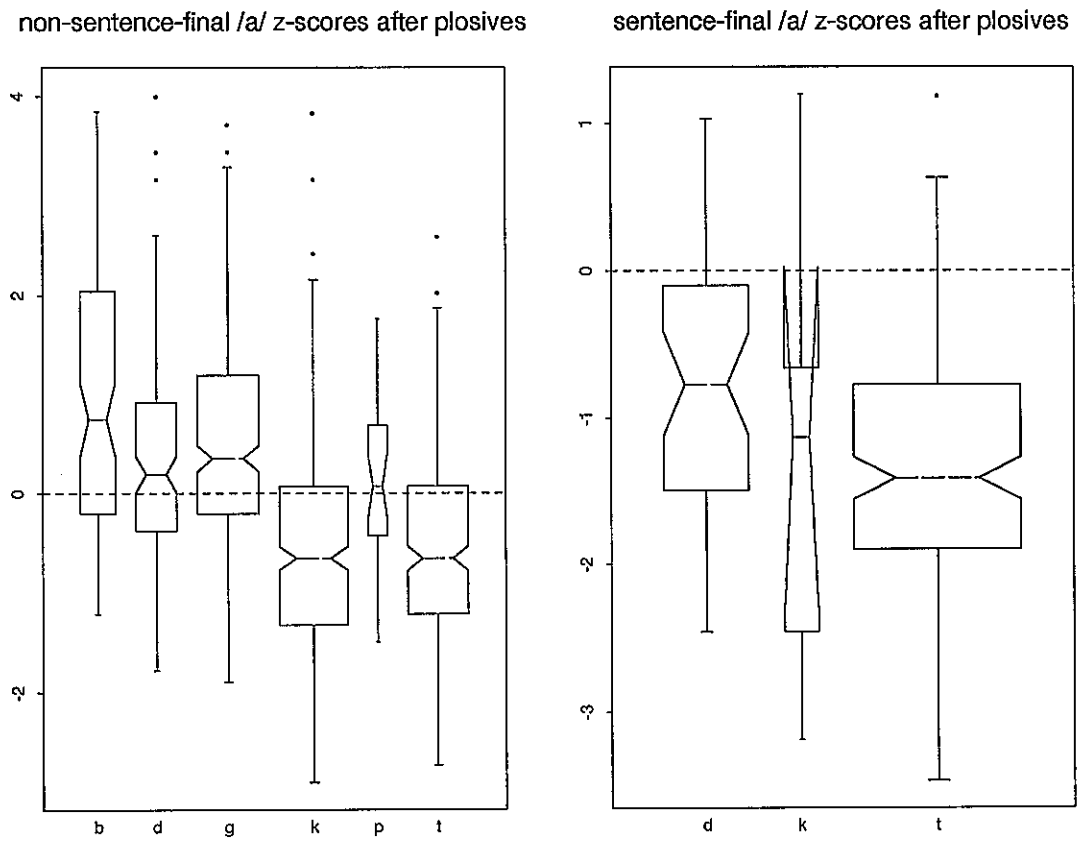

Figure 10: Plosive lengthening effects on following /a/.

biases the results for non-particle /a/, which as a result appears relatively short in this context. We can show evidence for the first, but only speculate about the second and third. Figure 10 shows the shortening effect of unvoiced plosives on following /a/ in non-sentence-final position (n=1140), compared with the effect in sentence-final position (n=195). In all cases it appears that there is some further shortening in the sentence-final context The difference due to sentence-final position is markedly less than the phonetically-motivated difference due to the previous unvoiced plosives. It may well be that the shortening reported for sentence-final vowels results primarily from the predominance of the /t/-/a/ combination in this position in Japanese sentences, rather than any context-specific shortening effect operating from situational features alone. This idea is also supported by the lengthening undergone by the consonants in the sentence-final mora, complementing the /w/-/a/ combination reported above. Going further, although the typical duration of /t/ in verbs (z = -0.7) is much less than that of /t/ in other environments (noun: z = 0.58, part: z

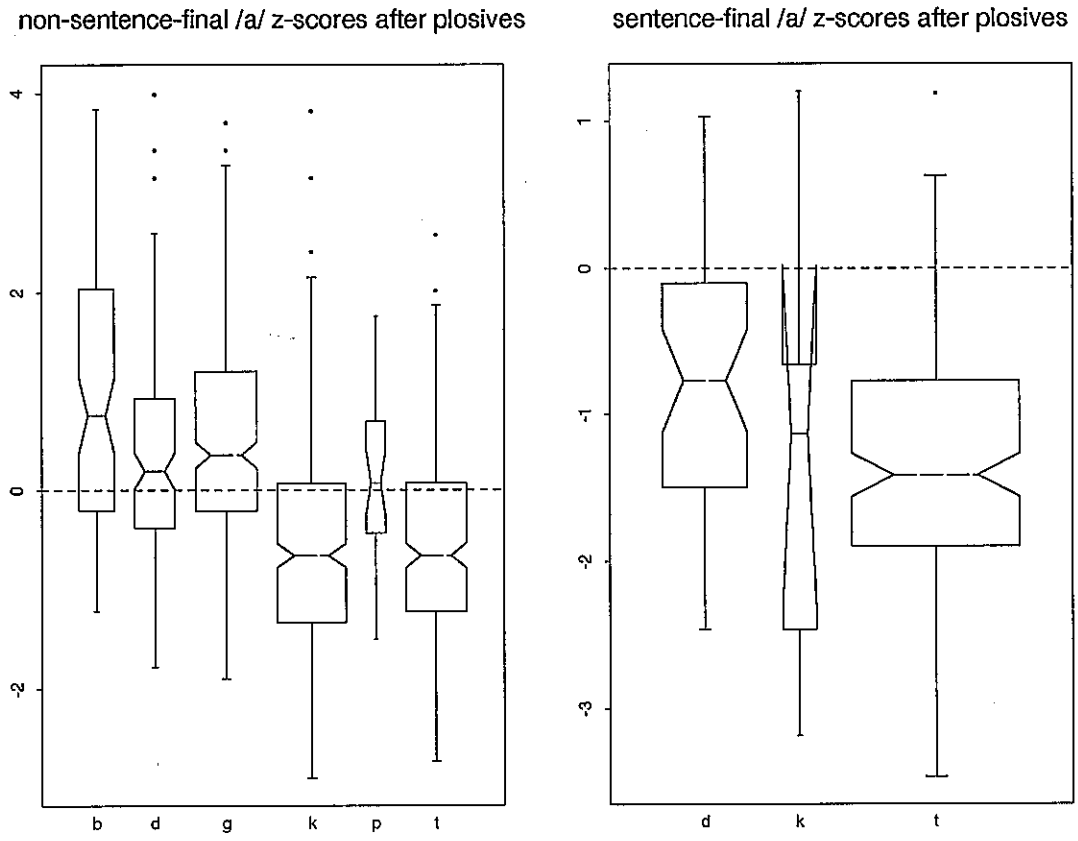

Figure 11: Lengthening of vowels after plosives.

Table 6: Mean z-scores and durations for /t/ by part-of-speech.

	mean (z)	SD	mean (ms)	SD	n
verb:	-0.70	0.60	46.93	24.52	315
noun:	0.58	1.58	98.89	64.34	336
part:	0.21	1.21	83.89	48.90	447
adv:	-0.32	1.00	62.28	40.62	57
adj:	-0.55	0.79	52.91	32.18	9

= 0.21,, see Table 6.), the /t/ in sentence-final position is found to be considerably lengthened (Figure 12). The anomalous non-lengthening of /ru/ in this Figure, (so segmented because of the difficulty of determining a phone boundary in the sonorant-sonorant sequence) can similarly be accounted for by the fact that 51 out of 52 occurrences are in sentence-final position, so again the 'average' will be that of the (presumably lengthened) majority. It becomes clear from this closer analysis, that the predominance of /a/ in sentence-final position, which is presumably a consequence of the verb's position in the Japanese sentence, combined with the distribution of past-tense verbs in this newspaper/magazine-sentence database, results in the appearance of shortening, even though the other types of (less numerous) segments appear to be lengthened.

8 Implications for speech synthesis

As bigger and bigger speech corpora become available, algorithms to predict the timing characteristics of speech are being based more and more on statistical analyses of these data. They result in rules, or operations, that apply to segments in given situations based on averages observed for similar segments in similar situations. But the nature of the similarities is not always defined. Vowels are vowels, and in the discussion of the shortening of *the final mora of a sentence* (ibid, p 19), there was no mention of *the vowel /a/* as a special case. There is perhaps a danger that over-generalisation could result in *all* mora in sentence-final position being shortened as a result of the exceptional but numerous occasions of the shortening of one particular type.

It is likely that statistical models of timing bias their average durations for a phone toward the durations of the most numerous occurrences. In the case of /w/, this can result in a 'false' average, which in conjunction with the possible lengthening effect noted for particles (based on the durations of the other segments occurring in these parts of speech) could result in over-long /w/ phone durations in all situations.

A similar generalisation was reported for the durations of phrase-initial syllables for English (Campbell 1989), based on statistical anal-

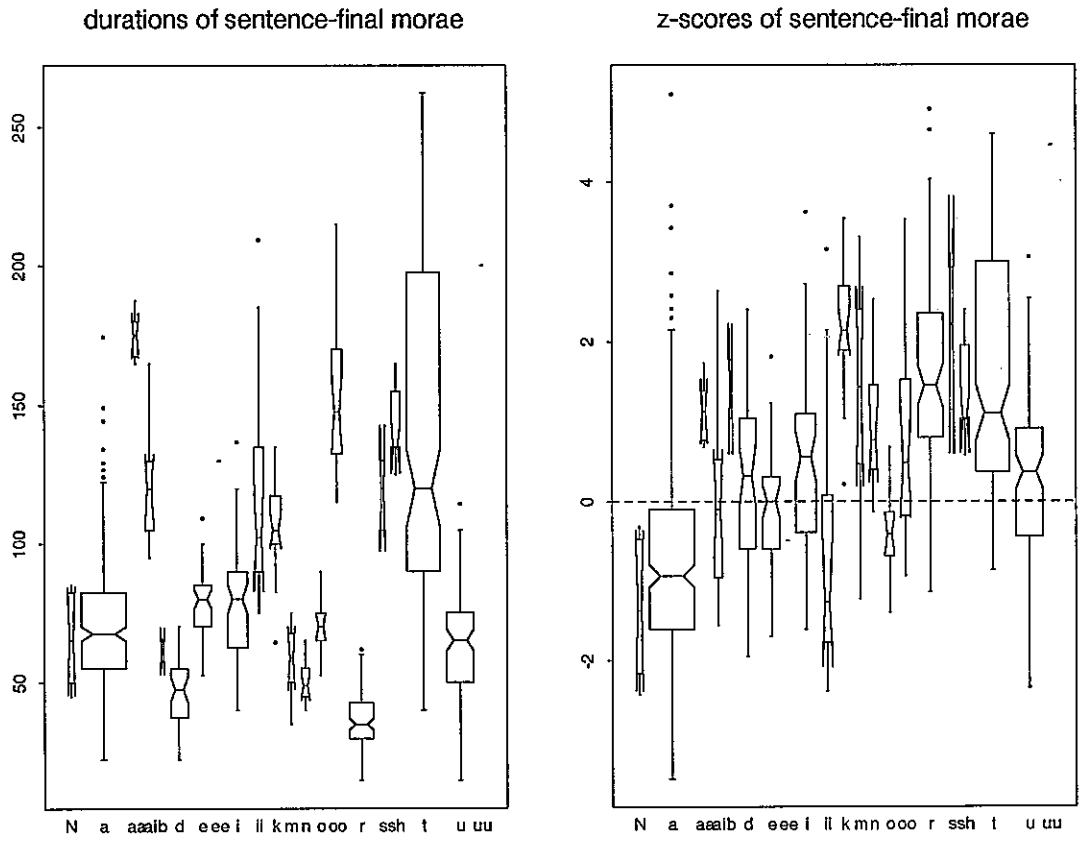

Figure 12: Lengthening of segments in sentence-final morae.

ysis of the durations of the syllables in a long (20-minute) passage of broadcast text. Shortening these initial syllables by 20% resulted in a better fit to the data, but It transpired to be an optimisation caused by the uneven distribution of determiners in English (with the word 'the' being far more common in initial position). The timing characteristics of the /dh/ phone in onset and coda positions are quite distinct, and by later modelling these separately, not only was the word 'the' well predicted, but the (far less frequent) non-'the' phrase-initial words were no longer being shortened as a result of the over-generalisation.

The attempt to make linguistic sense of observations of the data is a tedious, but not unnecccessary task.

9 Conclusion

Sentence-final shortening and noun lengthening have been reported in the literature for Japanese, as has phrase-initial shortening for English. To what extent are these effects statistical artifacts of the particular phone and word distributions of a particular corpus? Unfortunatley until more and bigger corpora become available this will be difficult to answer, but even then there is the danger that language-specific peculiarities will be manifest in the data, that lead to generalisations beyond their scope. Two examples have been illustrated here, more will no doubt be found.

Looking at durations in terms of length, normalising by factoring out the effects of individual phone differences has proved to be a useful way of examining the timing characteristics of speech, but it has been shown that the normalisation process itself, being statistically based, is not always immune to the same dangers.