Moraic and syllable-level effects on speech timing

音声タイミングにみられるモーラと音節の影響について

W. N. Campbell & Y. Sagisaka
ウイルヘルム N. キャンベル 匂坂 芳典

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

ABSTRACT:

A large database of spoken Japanese was examined for acoustic clues to the presence of a timing control at the level of the mora or syllable. Normalised measures of lengthening (z-scores) were employed to factor out the effects of any phoneme-specific durational differences, and analyses of variance were performed on mean lengthening values of both consonants and vowels factored according to previous and following contexts. A tendency to normalise the duration of the mora by lengthening segments in shorter mora and shortening segments in longer mora was found. A similar lengthening was found in segments around non-CV moraic units, that also served to normalise the distance between the surrounding morae.

あらまし:

モーラもしくは音節レベルのタイミング制御の存在を示す音響的な手がかりを求めて、大規模な日本語音声データベースを用いた分析を行った。音韻固有の継続時間長の特性を除去するため、正規化した時間長伸縮尺度（Z 値）を用い、隣接音韻環境にもとづく子音、母音の平均伸縮尺度値の分散分析を行なった。この結果、モーラを構成する各音韻長は、短い平均値となるモーラについては伸長し、逆に長い平均値となるモーラでは短縮し、モーラ長を均一に保つ傾向が見られた。同様な伸縮傾向はモーラを構成しない隣接音韻間にもみられ、隣接するモーラ間の間隔を均一にするように働くことが判明した。

1 Introduction

Japanese is often referred to as a mora-[1] or syllable-timed language. In order to examine the effects of such a supposed timing constraint on acoustic segmental durations, measures from a large database of read sentences of Japanese were examined.

Specifically, the following question is asked:

- Is there a tendency towards even spacing of the vowel onsets even if consonants of different articulatory complexity intervene?

There are considerable length differences between the different types of mora, largely as a result of differences in the intervening consonants, so if such a tendency is found, then evidence for a higher level of timing control in the domain of the mora or syllable can be assumed.

Further to the above, it is of interest to know whether such higher-level control can be thought of as taking place in a syllable or a mora framework. The Japanese orthography (kana alphabets) has separate symbols for each CV pair of sounds, suggesting that spoken Japanese may be similarly structured. A secondary question of the paper is then:

- Is there evidence for moraic articulation from the acoustic signal, or is the phoneme or syllable to be considered a more appropriate unit of articulatory timing?

Previous studies showing temporal compensation have employed actual measures of duration [3][4][5]. Rather than attempt to make sense of the actual observed durations of so many individual tokens, which can be subject to variance from many interacting features, a measure of lengthening, in standard units, is employed. Means taken over large numbers of samples further reduce the effects of individual or local peculiarities.

2 Normalised lengthening

The database consists of 503 sentences of varying length and complexity, extracted from Japanese magazines and newspapers and read by a professional announcer trained to

[1]The mora as defined here is a unit consisting of a single vowel, a CV pair, a CCV cluster (where the second C is an approximant), or a syllable-final nasal, corresponding to a single character in the Hepburn system of romanisation for the Japanese Kana alphabets

produce standard Japanese in the received NHK broadcasting style. The recordings were digitised at a sampling rate of 20k Hz and spectrograms produced from which trained transcribers determined segment boundaries and generated a computer-readable file of segment labels and durations [2].

Segmentation produced labels of contoid and vocalic segments which were then tagged and grouped into mora-, word-, phrase-, and sentence-sized units. In order to determine whether there is a measurable effect of moraic organisation on the duration of the segments in spoken Japanese, the durations were normalised and comparisons performed between the *length* as determined from absolute millisecond measures, and the *lengthening* as determined by z-score normalisation of the durations of all segments in the database for each phoneme class.

To determine the z-score for each segment, means and standard deviations were first calculated for each phoneme type, then the means subtracted from the individual durations and the differences expressed in terms of the standard deviations. For normally distributed data, this would result in a distribution of z-scores such that 68% fell between ± 1, and 99% between ± 3. In this way the amount of lengthening or shortening undergone by each segment, in comparison with other segments of the same type in the same database, can be expressed as a pure unitless number, independent of any durational peculiarities that may result from differences in manner or place of articulation. This measure of *lengthening* is independent of any phoneme-specific features of *length*.

If there is an effect on the durations for moraic grouping in articulation, and if the sequence of morae were to be cyclic or rhythmic, then it could be expected that inherently longer segments (in terms of articulation time) would require some compensatory compression or shortening of the following vowel in a CV pair, and inherently short segments on the other hand would undergo a corresponding expansion or lengthening, so that the overall duration of the CV unit approaches or tends to approach a standard timing.

3 Segment-level effects

Looking first at the lengthening (z-scores) of the central consonant (C_2) and vowel (V_2) in a $C_0V_0C_1V_1C_2V_2C_3V_3C_4V_4$ sequence, we find from an analysis of variance that by factoring the vowel z-score by type of consonant in the same mora (C_2) we get the strongest effect (F 33 12593 = 101.5). Next in strength is the effect of the following consonant (C_3: F 34 12592 = 82.56). Third comes the effect of the vowel in the following mora. The order is nearly the same for the consonant (C_2), but in this case the effects from the following mora (C_3V_3) appear stronger than those of the immediately following vowel in the same mora (V_2).

Effects on the vowel:

effect	df1	df2	F
v2 <- c2	33	12593	101.506
v2 <- c3	34	12592	82.567
v2 <- v3	17	12609	78.291
v2 <- v1	11	12615	20.486
v2 <- v4	17	12609	8.904
v2 <- c4	34	12592	6.576
v2 <- v0	16	12610	3.291
v2 <- c1	34	12592	3.126
v2 <- c0	34	12592	2.093

Effects on the consonant:

effect	df1	df2	F
c2 <- v3	17	9977	58.195
c2 <- c3	34	9960	40.016
c2 <- v2	13	9981	35.473
c2 <- v1	8	9986	16.275
c2 <- v0	16	9978	12.345
c2 <- c1	34	9960	9.721
c2 <- v4	17	9977	6.317
c2 <- c0	34	9960	4.098
c2 <- c4	34	9960	2.911

There is clearly a strong effect both from the consonant (C_3) and from the vowel (V_3) of the following mora, but it is probable that the latter (V_3) will have been subject to strong influences from its own (preceeding) consonant (C_3). Thus it would appear that in the flow of spoken Japanese, it is the consonants that cause the greatest effects on the timing, the vowels showing less variation both in articulatory (manner and place of articulation), and in durational characteristics. However, it is interesting to note, and will be discussed in more detail in the next section, that both accommodate to the differences in the other and combine to form a more regularly spaced interval than would be expected by prediction from their means alone.

The next level of influence is from the preceeding mora (C_1V_1), with a similar stronger effect on the vowel than on the consonant. Weak but significant effects were also detected from the next following mora (C_4V_4), suggesting that if there is a direction in the influences on timing it may by anticipatory rather than retroactive. This could be motivated phonetically by a need to prepare gestures more carefully in advance; the position from whence the articulator came being less important to the generation of the sound than the position to which it has to go.

4 Mora-level CV compensation

Vowel and consonant length are next examined according to groupings with both preceeding and following segments. In order to examine the effects of CV versus VC grouping, two sets of data were prepared. The phoneme string of the annotated speech was divided into both CV and VC units, wherever such a grouping was motivated. Exceptional cases of a single vowel or a nasal mora (counted here as neither C nor V, but N) were excluded. In order to determine the duration of the resulting unit, the durations of the two segments were added, and to determine the lengthening of the unit, their z-scores (z) were added. Addition of the individual C and V z-scores yields a better representation of unit lengthening than, for example, taking their averaged value; in the case of consonant z being positive and vowel z being negative, or vice versa, their (contradictory) effects would

be cancelled out, but in the case of their having the same sign, the effect is strengthened.

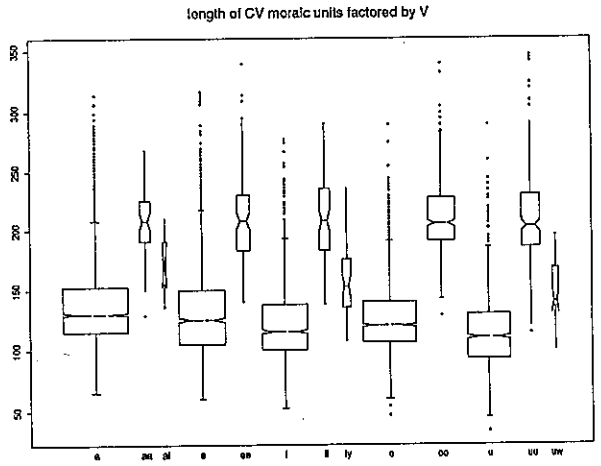

Figure 1: Consonant-Vowel (mora) unit durations factored by vowel

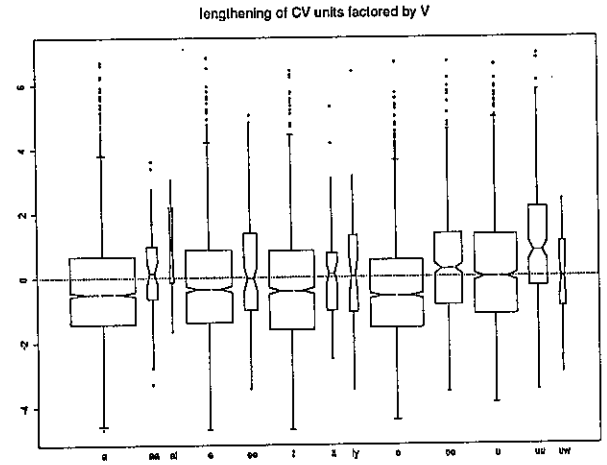

Figure 2: Consonant-Vowel (mora) unit z-scores factored by vowel

Figure 1 shows the durations of the mora in milliseconds, factored according to the vocalic part. Figure 2 shows the lengthening (summed z-scores of both C and V segments) in the mora. By factoring for the vowel, we are in fact looking at the differences in consonant lengthening, as the mean z-score for the grouped vowels will by definition be zero.

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 the 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.

These and the following Figures allow an important comparison to be made between the length of a mora, as measured in raw milliseconds, and its lengthening, reflected in the z-scores. It will be noticed that although considerably longer in millisecond terms, the morae with long vowels (with the exception of /uu/) do not show particularly different lengthening values, phoneme-specific durational differences having been factored out. Thus the fact that /uu/ mora show greater lengthening can be attributed to a compensatory effect in the consonants, rather than to any inherent length differences in the vowels.

There is little variation in the z-scores of the vowels, as shown in Figure 2., but it will be noticed that the inherently shorter /u/ appears to have a lengthening effect on its associated consonants, and the longer /a/, a shortening effect. Analysis of variance shows these differences to be significant ($F\ 12\ 9830 = 20.48\ p < 0.001$). That 'long' vowels have a lengthening effect on the consonants will be discussed in section 5.2.

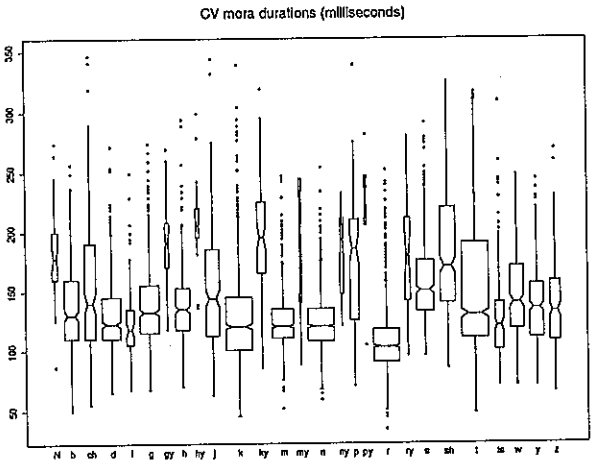

Figure 3: Consonant-Vowel (mora) unit durations factored by consonant

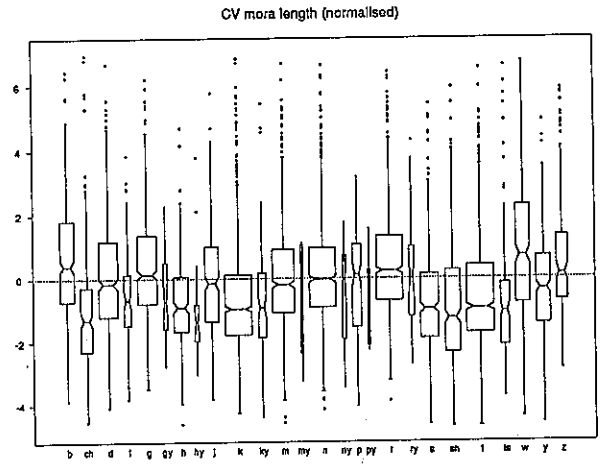

Figure 4: Consonant-Vowel (mora) unit z-scores factored by consonant

If we factor according to the consonants associated with these vowels ($F\ 26\ 9816 = 34.35$), it can be seen that there is a negative correlation between the length of the CV (moraic) unit and the shortening undergone by its components. This is particularly evident in longer mora such as those including /ky-/, /gy-/, /t-/ and /sh-/, in which the more precise placing and pressure of the articulators for the stops and fricatives results in longer consonants. These show significant shortening in their z-scores, and conversely the shorter

mora such as those including /r-/, /b-/ and /z-/, with no aspiration, and often just a lateralisation or rhotacisation of the vowel, appear lengthened. Exceptions to this pattern are the /w-/ group which appears almost exclusively in the word *wa* in phrase-final position and is therefore being lengthened for other known reasons, and the /k-/, /ts-/ morae in which the following vowel (predominantly /u/) is frequently devoiced.

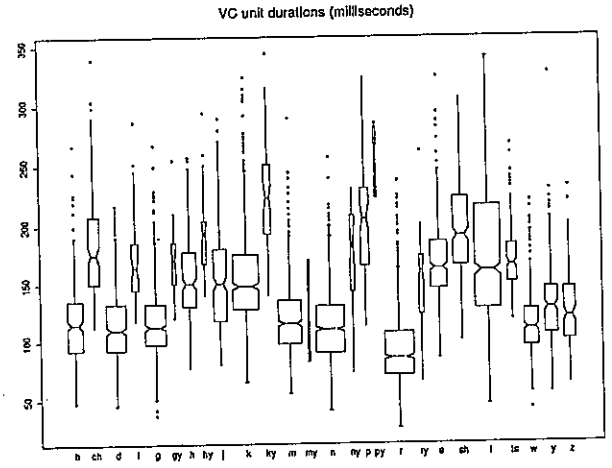

Figure 5: Vowel-Consonant (non-moraic) unit durations factored by consonant

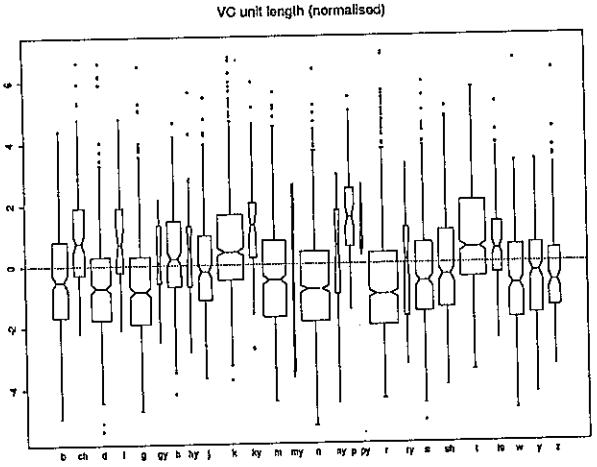

Figure 6: Vowel-Consonant (non-moraic) unit normalised scores factored by consonant

Figures 5 and 6 show the corresponding durations and lengths of the non-moraic VC groupings. A similarity in the directions of the variances can clearly be seen from these two figures, and reflects the positive correlation of the vowel durations with the durations of the following consonants.

A correlation of the means of the durations of the CV units with the means of their values of lengthening, factored by C, yields r = -0.432, compared with a closer positive correlation (r = 0.847) for the VC units similarly factored. In the case of vowels grouped with preceeding consonants (moraically motivated CV grouping), compensation appears to be taking place in the length of the vowel to better approximate a prototypical mora timing.

4.1 CV or VC?

Of particular interest here is the dichotomy of a positive correlation in the VC direction and a weaker negative correlation in the CV direction. Can this be resolved in view of the moraically motivated CV grouping, or must we revise our view of this link between orthography and speech and accept a segment-based description of the spoken language?

The stronger positive correlation can perhaps be explained in terms of anticipatory gestures related to the following phonetic context; if more careful positioning of the articulators is indeed required for the longer consonants [6], then it would seem that the preceeding vowel is being sustained longer while these organs move into position.

In the case of English, however, the vowel is longer before a voiced stop and shorter before an unvoiced one. In Japanese, the reverse appears to be the case - a vowel is typically *longer* before an *unvoiced* stop, and shorter before a voiced one. Since it would be unreasonable to conclude from this that the phonetics of the two languages are radically different, we must look for another explanation of the phenomenon.

	unvoiced stop:			voiced stop:		
	mean	sd	n	mean	sd	n
before:	111.3	84.50	1754	87.7	70.66	1486
after:	73.5	31.26	2165	89.7	29.43	1667

The answer lies in the CV grouping; the stop that follows the vowel is not in the same syllable, or articulatory unit. For English, the difference has been partly explained by a process of durational accommodation within the syllable [1]; equivalent lengthening effects are not found in vowels followed by a voiced stop thet starts a following word - i.e. one that is not in the same syllable.

Similar accommodation appears to be taking place in Japanese, resulting in a more stable timing pattern for the CV units; in Japanese, the vowel is more closely linked to the *preceeding* stop, and compensation is therefore in the opposite direction. Thus in spite of apparently opposite lengthening results, both the phonetics and the durational phonology of the two languages can remain consistent.

Visual comparison of the spread of CV unit versus VC unit durations, as plotted in figures 3 & 5, shows that this compensation results in a distribution much closer to the mean, with less variance in the moraic units (SD = 43 ms) and considerably more spread in the durations of the VC units (SD = 50 ms).

5 Gemination

Another test for the presence of a higher-level timing framework is the case of gemination. This doubling of segments is signalled in written Japanese by the insertion of a small 'tsu' before a geminate consonant, which is typically but not always an unvoiced stop, and by the insertion of a 'bar' after a lengthened (or 'long') vowel.

5.1 Doubled consonants:

In spoken Japanese, 'doubling' of consonants is generally realised by glottalised offset of the vowel, extension of the period of closure, and increased articulation of the release or frication.

5.1.1 Effects on consonant length:

Substantial lengthening is evident in the geminated consonant, which shows a mean three times that of the control group. In a similar compensatory manner, some shortening is observed in the duration of the consonants in the morae before (t 9217 = 3.41) and after (t 9100 = 2.57) the geminate. These trends are closely reflected throughout the percentile durations. Since the consonant lengthens to at least the average mora duration, this is sufficient reason alone to accept the geminate as an independent mora for timing purposes.

mean durations in milliseconds:

	consonants			vowels		
	mean	sd	n	mean	sd	n
control:	53.8	26.9	8940	79.9	29.6	10132
geminate:	151.0	39.7	420	70.2	26.2	420
before:	48.3	23.4	279	91.8	19.7	290
after:	48.3	24.0	162	78.3	28.5	186

5.1.2 Effects on vowel length:

The vowel in the mora immediately preceeding the geminate is significantly lengthened (t 10420 = 6.84) and the vowel in the geminated mora correspondingly shortened (t 10550 = 6.61), but no effect was found for the vowel of the mora following. These vowel effects are consistent with accommodation effects for stops as noted above, but the shortening of the consonants would be consistent with the idea of a higher-level timing framework.

These tendencies are not found in compund consonants. Some consonants, typically stops but also nasals, /h/ and /r/, compound with /y/ to form CCV morae. In contrast to the geminates, these show a shorter mean duration of 80 ms (SD = 32 ms, n = 557) and fall within the range of single mora durations with their correspondingly shortened following vowel.

5.2 Long vowels:

Some vowels are lengthened as a result of 'doubling' (the term is used here to avoid confusion with 'lengthening'), some simply appear doubled in a phoneme sequence but follow another vowel as a separate mora in their own right. Differences between these two classes of vowel and their associated consonants are examined next.

Doubled vowels in the database had a median duration of 127.5 ms (lower quartile: 105.5 ms, upper quartile: 150 ms) as opposed to a median duration of 80 ms (lower quartile: 62.5 ms, upper quartile: 100 ms) for 'single' vowels. These durations are not in fact double, but closer to a 50% increase.

Doubled vowels:			V-mora vowels:			CV-mora vowels:		
mean	sd	n	mean	sd	n	mean	sd	n
123.3	40.6	1280	93.5	29.1	1609	82.8	32.6	11695

Consonants in the same mora as these doubled vowels had a median duration of 77 ms (lower quartile: 54 ms, upper quartile: 105 ms) as opposed to 47.5 ms (lower quartile: 35 ms, upper quartile: 67 ms), again showing a similar increase.

consonant z-scores:

preceeding V-mora:			following V-mora:		
mean	sd	n	mean	sd	n
0.187	1.01	2093	0.278	0.97	2081

This is in contrast to the tendency seen above for consonants followed by long vowels to be shortened in compensation. If the mora-level timing hypothesis is accepted, then this lengthening could be taken to compensate for the 'inadequate' lengthening of the vowel and serve to bring the following vowel onset into better alignment.

The different case of vowels following as a single mora (V-mora vowels in the table above) shows significantly more lengthening than that of those in CV morae (t 13302 = 12.48), but not as much as is found in the doubled vowels: median = 90 ms (lower quartile: 75 ms, upper quartile: 110 ms)(t 2837 = 29.81). Here too some compensation may be taking place, as the consonants in the preceeding and following morae are significantly lengthened. No effect was measurable in the lengthening of the surrounding vowels.

6 Syllable or mora?

A question of particular interest is whether the mora or the syllable can be considered a basic unit of timing. A possible clue to this may be found in the effect of the mora-final nasal on the timing structure. In the orthography it is a separate character of the same status as CV units, so if the sound has a timing slot of its own, and takes up approximately a mora's worth of duration, then the mora can be considered prime; if on the other hand there is evidence of compensatory shortening, then the preceeding mora can simply be considered to have a nasal, consonantal offset, and a syllable-based description of the timing can be considered better motivated.

The database contained 466 occurrences of this nasal segment /ɴ/ in clearly measurable settings. Comparing the percentile durations with those of phonetically related /m/ and /n/ segments which only occur mora-initially,, we find more than twice the duration for this nasal mora.

	min	25%	50%	75%	ma
N:	25.0	75.0	90.0	107.5	185.0
n:	10.0	32.5	40.0	47.5	90.0
m:	15.0	38.1	45.0	52.5	90.0

Vowels and consonants in the mora immediately before and after each /ɴ/ were tagged accordingly ($C_1V_1NC_2V_2$), leaving the distanced mora as controls. A significant difference was found in the lengthening undergone by both vowels

and consonants in the three groups thus formed. Both the vowels (V_1) and the consonants (C_1) preceeding the ɴ are lengthened, and the following consonant (C_2) shortened. No significant effects were found for the following vowel (V_2).

Looking at the mean durations for vowels alone, there would appear to be no difference in lengthening, but the z-scores show that significant lengthening does in fact take place before the nasal.

```
mean durations in milliseconds:
          vowels                        consonants
         mean   sd     n          mean   sd     n
others:  79.7   29.6   10285      57.3   33.8   9089
V1:      82.2   17.9     409 C1:  63.1   32.3    378
V2:      80.6   32.5     334 C2:  63.9   33.9    334

Vowel durations:     F(2 11025)= 1.61 (n.s.)
Consonant durations: F(2 9798) = 10.90  p < 0.001

z-scores:
          vowels                        consonants
        mean    sd     n          mean   sd     n
others: -0.0021 0.9  10285       -0.002  1.0   9089
V1:      0.122  1.0    409  C1:  0.122  1.1    378
V2:     -0.007  0.9    334  C2: -0.135  0.9    334

Vowel z:     F(2 11025) = 37.52  p < 0.001
Consonant z: F(2 9798)  =  6.17  p = 0.002
```

Examination of the quantiles explains this apparent dichotomy; the control set includes a few extremely long vowels that bias the mean, but shorter vowels, up to somewhere between the 50th and 75th percentile are clearly longer in the 'before' case.

```
          min   25%   50%   75%  max
Control:  12.5  60.0  75.0  95.0 255.0
V1:    :  30.0  71.9  82.5  95.0 150.0
V2     :  20.0  60.0  75.0  95.0 197.5
```

This lengthening of both the preceeding segments suggests another form of accommodation, again serving to separate the vowel onsets, perhaps to make up for the shorter duration of a mora with only one segment. This alone cannot suffice as evidence for mora timing, but does support the hypothesis that the mora, not the syllable, is the prime unit of timing in Japanese.

7 Discussion

Raw millisecond observations of length tell us only a limited amount about the lengthening undergone by segments in a given situation. By normalising for the segment-specific length characteristics, and examining lengthening instead, we are able to understand more about the complex timing interactions in speech. Laboratory experiments with specially constructed sentences allow precise comparisons to be made between minimally distinctive phenomena, but at the expense of naturalness; sentences spoken in a laboratory environment for the purpose of eliciting speech data are rarely truly representative of the prosody of natural speech. In this paper, the effects of moraically motivated grouping on the lengthening of segments was examined in a large corpus of spoken material. Any loss of individual controls can therefore be made up for by the statistical significance of large numbers of samples.

Evidence has been presented that suggests, but does not prove, that compensation is taking place in Japanese timing, that results in a tendency for more regularity in the onset of vowels than would be accounted for by consideration of the mean durations of the individual segments alone. Variants to the more common CV pattern were examined, and in each case a tendency towards more uniform timing was noted.

That this regularity may be expressed in terms of moraic units rather than syllables was shown by the case of /ɴ/ which, although appearing in what would in syllable-based languages be coda position, closing the syllable after the vocalic peak, clearly takes up a mora-sized slot in the timing framework. Compensatory lengthening of neighbouring segments may be assumed to make up for the lack of a vowel in the mora.

If these are strong effects, then modelling them at the higher level of the mora may reduce the need for the complex modelling of the interactions that is required at the segment level. Such a two-level model has been described by Campbell & Isard [1] for English, and tests are currently being carried out, using z-score prediction to determine whether similar modelling for Japanese would allow better control of such continuously variable parameters as speaking rate, that are not well modelled by current methods.

Acknowledgement

Particular thanks are expressed to the labellers at ATR who painstakingly created this database by hand, and to the management ATR for enabling this research to be performed.

References

[1] Campbell, W. N. & Isard, S. D. (1991) *Segment durations in a syllable frame* Journal of Phonetics #19.

[2] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., & Shikano, K. (1990) *ATR Japanese Speech Database as a tool of Speech Recogntion and Synthesis* Speech Communication 9, 357-363.

[3] Higuchi, N. (1981) 日本語連続音声における単音の持続時間に関する研究, PhD Dissertation, Tokyo University.

[4] Sagisaka, Y. (1985) 音声合成のための韻律制御の研究, PhD Dissertation, Waseda University.

[5] Sato, H. (1987) 規則による音声合成の研究, PhD Dissertation, Hokkaido University.

[6] Vatikosis-Bateson, E. (1988) *Linguistic Structure and Articulatory Dynamics* Indiana University Linguistics Club.