A Study of Japanese Speech Timing from the Syllable Perspective

Nick Campbell ATR-ITL

Abstract

This paper examines the durational characteristics of a corpus of Japanese speech from the point-of-view of the syllable-timing hypothesis. Japanese is clearly a mora-timed language, but there is evidence of syllable- level timing control that is comparable to that found for English. In particular, vowels stretch to fill the syllable frame and show accommodation with the special mora that occupy the same syllable. We show that although syllable duration is affected to some extent by changes in the phonemic composition, there is a strong top-down effect that governs individual segment durations. However, unlike English, we show that consonants in Japanese are much more robust against durational change from the syllable level, and that most of the accommodation is absorbed by the vowels in Japanese speech.

本稿では、アコモデーション理論の立場から日本語連続発話における音韻継続時間長の特徴について述べる。アコモデーション理論は、英語の発話タイミングを説明するためのもので、その日本語への適用を試みる。日本語は主にモーラ言語であるが、英語に見られる音節レベルからのタイミング制御の機能も認められ、子音も母音も音節枠内で伸縮しあう。本論文では特殊拍によるタイミングの差を基に、モーラ枠と音節枠との相違を検討した。その結果、両言語ともに音節枠からの影響が見え、日本語の場合、英語に比べ子音の固定性は高く、母音はその子音と調整しながら、音節枠内での継続時間を決定する。つまり発話内のタイミング制御は階層的構造で機能し、高次層では言語情報、発話速度、強調などにより音節枠が決定され、また低次層では調音的（物理的）要因によって音節枠内で音素間の伸縮が行なわれると考えられる。

Keywords: speech timing, accommodation theory, mora vs syllable, higher-level timing control

1 Introduction - Higher-level Timing Control

The phonetic and pragmatic contexts in which speech sounds are realised can have considerable influence on segmental timing. Part of the variation can be accounted for by the mechanical processes of speech production and by taking into account the movements of the articulators involved. For example, the jaw is more massy and slower in its movements than the tongue tip, so we can expect that sounds which depend on jaw movement, such as open vowels, might tend on average to be longer in duration than those which require less jaw movement, such as close vowels [1, 2]. Similarly, sounds which depend primarily on rapid articulator movements, such as flaps and dental taps, might tend on average to be shorter in duration than those such as fricatives which require more prolonged alignment of the articulators. We can thus account for a large part of the variation in speech timing processes by phonetic factorisation according to place, manner, and context of articulation. However, it is not possible to account for all the timing variation by phonetic considerations alone; much of the expression of meaning in speech is controlled by prosodic variation, and higher-level timing control has been shown to play a large part in such distinctions [3, 5, 6].

A model has been proposed to account for speech timing in English according to the joint principles of elasticity and accommodation [7]. It shows that variation in the durations of individual speech sounds can be accounted for by a multi-level process of accommodation in which syllable durations are first calculated from higher-level knowledge about the discourse context and the syntactic and semantic composition of the utterance, and then their component segments are accommodated into this syllable-level durational framework according to the principle of elasticity [6]. Under the elasticity principle, individual segments are assumed to stretch or shrink uniformly in duration, so that they conform to the higher-level timing framework.

In other words, phoneme durations are derived as a secondary process from the syllable durations, rather than syllable durations simply being the sum of their independently-determined component phoneme durations. This process is not inflexible though, and has been shown to include some two-way influences, such that syllables composed of inherently longer phonemes tend to be slightly longer than predicted, and those composed of inherently shorter phonemes may be shorter. However, the overall rhythm determined by the syllable-level timing control predominates. for text-to-speech synthesis, this model produces better timing control as it prevents prediction

errors from accumulating to disturb the rhythms of the speech, and isolates any errors within a given level.

In this paper, we will not be concerned with the prediction of higher-level timing patterns, since these require access to information about the structure and intended meaning of each utterance, but will instead examine a database of segmented Japanese speech and attempt to account for its durational variation under the syllable-level accommodation framework. That is, we will take the syllable durations as given, and examine the evidence for accommodation on the basis of local durational variation alone. Although the syllable-level timing model was developed for English, we will show that similar accommodation can be seen in the data from the Japanese speech.

Japanese and English have very different durational characteristics, but they have many similarities at the segmental level, and we will start by examining these. Section 2 presents a summary of English phoneme dirstibutions and shows that similarly articulated sounds in Japanese have similar 'inherent' durations. Section 3 examines the syllable structure of Japanese speech, with reference to the predominant moraic description, and shows that similar processes of accommodation can be seen. Section 4 shows the extent to which these durational characteristics can be prdicted by statistical methods. The paper finishes with discussion of syllable influence on the timing of Japanese speech.

2 Segmental Duration Characteristics

To the casual observer the timing patterns of Japanese and English sound very different: we have the impression of fast, regular, almost clipped speech in the former and of 'waves of prominence' in the latter, and the two languages have been cited as polar examples of quite different stress-based and syllable-based timing structures [9]. However, direct comparison of the segmental durations of the two languages fails to reveal much about these differences, and it is in their syntagmatic structuring, rather than in any inherent or paradigmatic characteristics that the difference can be found.

Figure 1 (from [8]) shows the durational distributions of four male and four female Japanese speakers reading 503 phrase and sentence-length excerpts from Japanese newspapers and magazines, and compares them with equivalent durations measured from two male and two female speakers of British English (RP) reading 200 similarly constructed sentences. There appears to be surprisingly little difference at this level between the two languages.

Figure 3 about here

In order to visualise the more abstract 'timing' effects separately from the physical 'durations', which are dependent on phone type and local phonetic context, we can employ a procedure of normalisation borrowed from the social sciences. The 'z-score' transform allows us to factor out predictable contextual differences by expressing the durations on a scale from 'long' to 'short', with the 'average' being the centre-point. This measure of *lengthening* is independent of any phoneme-specific features of *length*. By calculating the averages separately for each class of context we factor out predictable effects such as the longer duration of an /a/ compared to that of an an /u/, and consider instead the 'lengthening' they undergo in common as a result of higher-level contextual effects.

To determine the z-score for each segment, means and standard deviations are first calculated for each phoneme type, then the means subtracted from the individual durations and the differences expressed in terms of the standard deviations. By expressing lengthening in terms of position within a distribution, or as z-scores, the raw millisecond durations are transformed into unitless values, and comparison can be made not only within each type but also between phones of different types. A positive value of z represents lengthening, and a negative value shortening, relative to the mean duration observed for all tokens of the same type in the database. With normally distributed (Gaussian) data, 68% of the tokens can be expected to fall within $\pm$ 1 SD, and 99% between $\pm$ 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. The skew observed in most distributions of raw segmental durations is reduced towards Gaussian by taking the log of each duration instead of its raw value.

Figure 2 (also from [8]) shows durations and equivalent z-scores for the 5 vowels and 10 consonants of Japanese. The boxes in the plot 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 each sample, and the notches indicate significance at the 5% level in the difference of the distributions if they show no overlap. Differences in the raw durations are all significant (except perhaps between /e/ and /o/), but the z-score normalisation removes such phone-specific dependencies.

Figure 2 about here

For this study, we examined the segmental durations of one speaker, a female professional announcer FTK (not included among those of Figure 1), reading the same list of 503 phonemically-balanced sentences [10] extracted from a large corpus of newspaper and magazine articles. The corpus has been prosodically labelled according to the J-ToBI conventions [11, 12]. Although we do not have access to morphological, semantic, or syntactic information about the utterances, the prosodically-relevant boundaries are available from the break index labels, so phrasing and and accenting distinctions are available. Table 1 lists the phonemic labels that were used and indicates minimum and maximum durations observed, with quantile durations and the number of tokens for each label type.

Table 1 about here

3 Syllable-based speech timing

The syllable-timing hypothesis was proposed by Campbell & Isard [6] and expanded further in Campbell 1992 [?] to account for the interaction between higher and lower levels of timing control. It shows the function of the syllable as an intermediate level of timing control and offers a way to map the effects of linguistic and semantic contexts onto the physiological constraints of speech sound production. It shows how all segments in a syllable can share the available space equally as a function of their elasticity and that when lengthening is viewed in variance rather than percentage terms both consonants and vowels have similar lengthening characteristics.

The phonology of Japanese classifies speech sounds in terms of morae [13, 14, 15], where each mora consists of a vowel and an optional consonantal onset, with special-morae being additionally related. These include the nasal mora (/N/) which follows the vowel, doubling or compounding of vowels (/V2/), and gemination of consonants (/Q/). Velarisation of consonantal onsets is also encountered but is not considered to affect the mora count (we use the notation /CyV/ to indicate velarisation). The sound sequence /kya/ is considered to be a single mora, /N/, /be/, and /ru/ likewise. The name 'Campbell' is therefore pronounced as a sequence of 4 morae in Japanese. 'Nick' is composed of 3 morae; /ni/, the doubling geminate /Q/, and /ku/ (transcribed as nikku).

Under the syllable view, these names would be considered as tri- and bi-syllabic in their Japanese pronunciation. If we use the symbols V(vowel) C(consonant) N(moraic nasal) and y(velarisation) to denote the syllable components, then the names are composed of CyVN+CV+CV and CV+QCV respectively. Table 2 lists the syllable types in the above notation, as found in the 503-sentence corpus, sorted in order of frequency of occurrence.

Table 2 about here

There is considerable evidence for the moraic structure of speech timing in Japanese [16, 17, 18], and the moraic structure closely corresponds to the kana alpahabets which Japanese children first learn to read from. In young childrens' reading we can clearly hear a one-to-one character-to-timing relationship. However, as they mature, their articulation becomes habitual, more sophisticated and more co-articulated, and there is evidence of the influence of syllabic structuring on the speech articulation and timing. Below, we will compare the influence of these two underlying structures, the mora and the syllable, as expressed in the durational characteristics of adult speech.

3.1 The moraic nasal

In normal Japanese speech, the moraic nasal can be phonetically difficult to distinguish or segregate from the vowel it is supposed to 'follow'. It is therefore as feasible to consider the syllabic nucleus as being nasalised as it is to consider the bi-moraic syllable as being constructed of two sounds in sequence. A simple way to distinguish between these two competing views is to examine the durations of the segments; if the timing of the sequence is equivalent to the sum of both the vowel duration and the nasal duration then we can accept the concatenative view; if less, then we should consider the overlaying view. Since other influences on the durations can never be completely factored out, the solution in practice is not so simple, but the results are clear: As we can see from Table 2, the median length of a V syllable is 90 ms, and a VN syllable is 223 ms, A CV syllable has median duration of 142 ms and a CVN syllable 251 ms. The moraic nasal is therefore confirmed as an independent moraic unit.

Table 3 about here

For comparison, the dental nasal /n/ which is arguably phonetically related but always co-exists in a moraic frame with an associated following vowel has observed durations which are approximately half those of the moraic nasal (see Table 3 bottom). The relative shortening observed in the duration of the /N/ in the CVN syllables is not significant $t_{760} = 1.18$.

However, this simple view is complicated somewhat by the lengthening of the preceeding vowel, and by what appears to be a compensatory adjustment in the duration of the following nasal mora. Tables 3 and 4 show that vowels are significantly longer when followed by a nasal mora (or a /N/ in the same syllable). Furthermore, the nasal mora itself undergoes significant lengthening when the previous vowel is short. There is a negative correlation of 0.46 between the two sets of V and N durations. This is consistant with the view that they both occupy a space within the same higher-level framework, accommodating to each other to optimally fill the frame. Thus, the syllable-level view also appears justified.

Table 4 about here

3.2 Gemination of consonants

In a similar way, we can test the same-syllable influence of another 'special mora', often transcribed by /Q/, which is marked in our corpus by doubling or gemination of consonants. The name transcribed as /ni+kku/ above is expressed in Japanese as a three-mora sequence (/ni+Q+ku/). We can see from Table 5 that the Q-mora (realised as a longer closure of the plosive or extended frication of the /s/) occupies a vowel-length timing slot, with an average of 80 ms difference between the simple and geminate forms of each consonant.

Table 5 about here

Again, these data reveal cross-mora influences, and if we examine the durations of the vowels coming before the geminate, we find a significant lengthening (Table 6) $t_{1796} = 19.98p < 0.001$, This might be attributable to segmentation differences in the labelling of the waveform, but the same criteria of formant energies are used, and the following phonetic context is quite similar in each case, so other explanations should be sought.

Table 6 about here

The lengthening of the vowel before the longer geminates runs counter to the accommodation principle, which would predict a shortening of vowels sharing the same syllable with a longer consonant. However, the duration of vowels *following* the geminated onsets (see Table 7.) confirms the expected shortening for vowels occupying the same higher-level frame as the lengthened consonant i.e., the vowel following the longer consonant is typically shortened. This timing shift is difficult to account for in terms of a sequence of individual morae, but if we suppose that the previous syllable is closed by the /Q/ and the following syllable is opened by the lengthened /C/ (which is consistent with the principle of maximal onset of syllables) then these duration characteristics can be explained. Figure 3 shows that the preceeding vowel lengthens slightly into the /Q/ and the geminated consonant exerts a shortening force on the following vowel which is in the same mora (or syllable) frame.

Table 7 about here

Figure 3 about here

3.3 Lengthening and plosives

In English, when a vowel is followed by a voiced plosive (/b/,/d/, or /g/) in the same syllable it is generally longer than an equivalent vowel followed by an unvoiced plosive of the same place of articulation (/p/, /k/ or /t/) [4]. In Japanese, as can be seen from Figure 4, there is also a very clear separation of effects on vowel length according to the manner of voicing of the following plosive. By grouping together the z-scores for the vowels (which all by definition have zero mean) and displaying only values for that subset which is followed or preceeded by a plosive, further factored in the figure according to the type of plosive, we can ignore any differences in vowel type and see instead the general effects of the environment. The left-hand side of the figure shows the plosive durations in millioseconds. The right-side plot shows the distribution of z-scores for all vowels in pre-plosive contexts. There are significant differences in the lengthening according to the type of following plosive but in this case the effect known for English appears to be reversed and although separation can be seen according to the voicing class of the following plosive, the vowels appear to be *shorter* relative to their mean duration when the following plosive is voiced. It is only when viewed in the context of the higher-level framework that this discrepancy is resolved. All becomes clear if we look at the distribution of lengthening of vowels when they *follow* a plosive. In Japanese, vowels show significant lengthening when they follow a voiced plosive and shortening after an unvoiced one.

Figure 4 shows the difference in lengthening clearly for the two types of voicing. If we think just in terms of sequential order effects then the two languages appear to be in contradiction. The difference, however, is in the higher-level structuring of the languages rather than in the linear segmental organisation. In Japanese, a plosive following a vowel is in a different mora, while one that precedes the vowel is more closely affiliated in terms of moraic (CV) organisation. Seen in terms of the higher-level unit, the lengthening characteristics of vowels in both languages are consistent and, as in English, the vowel simply stretches to share the available space with the other segment(s) in the syllable. The amount of space it can have is determined by higher-level factors and the elasticity of the segments.

Figure 4 about here

4 Syllable normalisation

Syllable durations can be similarly normalised. We assume as a working hypothesis that the main influence on syllable duration is its phonemic composition. Thus a syllable of type QCyVN can be expected to be longer than one of type CV, which in turn can be expected to be longer than one consisting of a vowel alone (V). Since phonemic durations are also type-specific, we can expect a single-vowel syllable of type /a/ to be longer than one of type /u/. Table 8 shows the mean durations for each syllable type and confirms that in general this is indeed the case.

Table 8 about here

By applying z-score normalisation to the syllable durations, calculated from the sum of their component phone durations, according to syllable type (QCyV2N, but ignoring phonetic distnctions) we can factor out structural complexity differences and simplify the analysis of durational lengthening which arises from context differences.

Linear regression has been shown to be effective for the prediction of both segmental and syllable durations [7, 19, 20]. In this section we employ a simple least-squares method [21] to determine the contribution of various factors to the observed timing characteristics. Predicting syllable duration from the QCyV2N syllable-type classes (a rough indicator of syllabic complexity) yields a Multiple R-Squared value of 0.69, accounting for almost 70% of the variation and justifying our initial assumption for the normalisation. This model predicts durations with a median error of 5 ms and an interquartile error of approximately 20 ms, which is less than 10% of the average syllable duration.

Predicting a syllable's duration from the number of phones it includes is less effective. Coding a V syllable as length 1, a CV or a V2 sylllable as length 2, CVN, QCV, or CV2 as length 3, etc., allows us to predict only 51% of the variance. However, this is an interesting way to confirm the intuition that velarisation does not affect mora count: we can code CyV as either three phones or two, depending on whether we classify the /y/ as an independent consonant or as a feature of articulation on the consonant-vowel combination. The former (incrementing by one the count of all syllable types that include velarisation) yields an r-squared value of 0.473, which indicates a significantly worse prediction of syllable duration when the velarisation is given independent consonantal status.

The amount of stretch undergone by a syllable as a result of differences in phonemic composition can be calculated from the multiple r-squared value produced by an analysis of variance of the regression results. Syllable z-scores (syl-z) fitted by syllable type (QCyV2N) yield a multiple r-square of zero (6.39e-27), confirming that no relationship can be found between the composition of the syllable and its stretch. This is because we originally factored according to composition and simply verifies the accuracy of the normalisation procedure. However, fitting vowel-type to syl-z yields an r-squared value of 0.12 (a correlation of about 0.34 between predicted and observed) which suggests that 12% of the variance in the stretch of the syllable lengthening can be accounted for by the nature of the syllabic peak, confirming that in Japanese longer vowels such as /a/ produce longer syllables, on average, than those composed of shorter vowels, such as /u/) (see also Table 5).

The influence of consonantal composition on the syllable lengthening can be similarly computed. Fitting syl-z by two factors, nature of the vowel and the type of onset consonant, accounts for 18% of the syllable stretch. The remaining variation is left to be explained by factors at levels higher than the phonetic, such as phrasal position, overall speaking rate, semantic type, focus, pragmatic force, etc., which are beyond the scope of this paper (but see for example [22]).

Conversely, we can model the phoneme durations from consideration of the syllable composition and the stretch it undergoes in a given context. Predicting onset consonant durations, we can account for 78% of the variation by consonant type alone, 80% by consonant type and vowel type together, and 86% if we include syl-z as a factor. Predicting vowel durations however, we can only account for 21% of the variance from vowel type alone, 28% if we include consonant type, and 30% if we also include coda information (whether it is followed by a nasal mora), but this increases to 74% if we include syl-z.

5 Conclusion

This paper has presented an analysis of segmental durations of Japanese under the framework of sylllable-level accommodation, and has shown that this view is not unfounded. There is clear evidence for moraic structuring from the durational characteristics of the speech, but there is also evidence of interactions that can better be explained within the syllable framework.

For the prediction of segmental and syllablic durations in a text-to-speech synthesis system, the use of relatively simple linear regression techniques is not optimal and more sophisticated statistical methods such as artificial neural-networks or tree-based classification are to be recommended, but the newer statistical models are less amenable to post-hoc analysis and for the purposes of this paper we prefer the well-established analysis of variance methods for distinguishing the contributions of individual factors,

We found that only 18% of syllable stretch can be accounted for by segmental variation, and that linear models predict 86% of the variance for consonants and 74% for vowels if syllable stretch is included.. This is in accordance with the elasticity hypothesis described in Section 1, under which phoneme durations accommodate to the syllable frame, but also confirms the difference that was clear in Figure 1; although the vowels and consonants of Japanese and English appear to occupy the same space in terms of duration and variability, the significant difference between the two languages lies in the flexibility of the consonants. It appears that Japanese consonants are much less susceptible to influence from the syllable stretch than are their English counterparts, and that Japanese vowels accommodate as much to the surrounding context as to the syllable lengthening.

References

[1] V. A. Kozhevnikov and L. A. Christovich. *Speech Articulation and Perception.*, volume 30-543. Joint Publications Research Service, U.S. Department of Commerce, 1965.

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

[3] D. H. Klatt. "Vowel lengthening is syntactically determined in a connected discourse." *Journal of Phonetics*, 3:129–140, 1975.

[4] Klatt, D. H. "Linguistic uses of segment duration in English." *Journal of the Acoustical Society of America*, 59:1208–1221., 1976.

[5] Huggins, A. W. F. "The perception of timing in natural speech: compensation within the syllable." *Language and Speech*, 11:1–11, 1968.

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

[7] Campbell, W. N. **Multi-level Timing in Speech**, PhD Thesis, Sussex University (UK) 1992.

[8] Campbell, W. N. "Timing in Speech: a multi-level process", in **Text and Language Technology** Festschrift for Gosta Bruce, Ed. Merle Horne, Kluwer Publishers (in Press), 1999.

[9] Dauer, R., "Stress timing and syllable timing reanalysed." *Journal of Phonetics*, 11:561–62, 1983.

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

[11] Pierrehumbert, J., & Beckman, M. **Japanese Tone Structure.** Cambridge, MA: MIT Press. 1988.

[12] Venditti, J. J., "Japanese ToBI Labelling Guidelines", Technical Report, Ohio-State University, U.S.A. 1995.

[13] Kaneda, I. 金田一春彦：「日本語音韻の研究」 東京堂出版, 東京, 1967.

[14] Hattori, S. 服部四郎：「言語学の方法」 岩波書店, 東京, 1960.

[15] Sugito, M. 杉藤美代子：「日本語アクセントの研究」 (三省堂, 東京, 1982.

[16] Higuchi, N. 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. 1981.

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

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

[19] Sagisaka et.al 匂坂芳典・海木延佳・岩橋直人・三村克彦："ATR ν - Talk 音声合成システム", 情報処理学会「音声言語情報処理と音声入出力装置」研究グループ、電子情報通信学会「音声認識の実用化をめざす新手法」時限研究専門委員会研究会資料 1992.

[20] Sagisaka, Y. コーパス ベース音声合成 Journal of the Signal Processing Society, pp.407-414, 1998.

[21] McCullagh, P. and Nelder, J. A. **Generalized Linear Models**. London: Chapman and Hall. 1983.

[22] ニック キャンベル「韻律解釈における基本単位」 in 音声と文法 II、音声文法研究会、(Spoken Language Working Group) くろしお出版 1999

Table 1: Labels used in the FTK 503-sentence corpus, showing minimum, maximum, quartile and median durations in milliseconds. (I,U: devoiced vowels, N: the moraic nasal). The last column shows the count of tokens for each type in the corpus.

	min	1Q	med	3Q	max	count
o	35	80	90	107	220	3654
a	40	90	105	120	240	3514
i	25	65	80	100	197	2438
u	20	60	75	90	204	2085
e	35	80	95	110	225	1952
k	20	50	62	75	190	1304
n	15	45	50	60	110	1252
r	10	20	25	35	99	1129
I	22	45	55	68	110	315
U	19	39	49	61	104	305
t	15	40	55	71	150	916
N	27	65	84	104	180	891
m	20	55	60	70	113	804
g	20	45	55	65	108	691
s	23	67	80	94	165	660
d	16	40	45	54	80	592
sh	22	74	91	115	197	541
h	20	48	60	75	125	428
w	20	35	40	45	82	413
y	23	50	60	68	105	408
j	27	54	65	77	148	354
b	25	45	55	65	105	354
ch	31	65	78	100	168	250
ts	30	66	80	100	170	248
tt	60	130	162	210	335	228
z	33	51	60	70	125	214
f	20	40	53	61	110	150
ky	50	84	105	129	180	98
p	21	44	55	70	100	66
pp	70	120	135	151	190	60
kk	85	125	155	165	260	53
gy	51	87	99	105	128	33
ny	58	96	100	120	145	29
hy	66	104	111	120	157	28
ry	40	66	76	90	136	41
ssh	66	145	158	183	232	33
by	55	93	110	117	148	12
cch	80	128	151	170	209	12
ss	97	131	146	159	198	10
kky	145	165	170	178	198	6
my	70	75	100	120	135	6
ppy	130	155	159	197	200	5
dd	91	103	115	126	137	3
py	83	86	90	94	98	2
tts	158	167	176	185	195	2
dy	75	75	75	75	75	1

Table 2: Syllable-level labels from the 503-sentence corpus, showing minimum, maximum and quartile and median durations in milliseconds. (C: consonant, Q: gemination, V: vowel, 2: vowel doubling, Y: velarisation, N: nasal mora). The last column shows the count of tokens for each type in the corpus.

	min	1Q	med	3Q	max	count
CV:	40	125	142	160	345	7592
CV2:	191	289	312	340	498	1385
V:	33	75	90	110	191	775
CyV:	60	147	170	195	362	765
CVN:	163	230	251	276	365	727
QCV:	60	130	166	244	468	610
QCV2:	109	169	199	245	480	207
V2:	122	163	185	212	278	106
QCVN:	205	296	390	488	833	48
CyVN:	225	267	284	314	391	37
VN:	140	180	203	219	242	35
CV2N:	279	314	324	350	399	22
QCyV:	165	206	223	255	275	19
V2N:	138	157	158	162	224	13
QCyVN:	290	341	372	391	493	8
QCyV2:	221	230	254	272	335	6
QCV2N:	346	346	346	346	346	1

Table 3: Vowel durations when followed by a nasal mora (The dental nasal /n/ is included for comparison).

	vowel						nasal				
	min	1Q	med	3Q	max	count	min	1Q	med	3Q	max
V	33	75	90	110	191	775	-	-	-	-	-
VN	50	95	108	124	144	35	48	79	86	100	150
CV	19	75	90	107	240	7592	-	-	-	-	-
CVN	52	100	110	120	178	727	27	65	84	104	180
/n/	-	-	-	-	-	1281	15	45	50	57	110

Table 4: The difference in the duration of the nasal mora is significant in all cases.
(after /a/ vs. after /o/: $t_{348} = 10.62$ $p < 0.001$, after /u/ vs. after /o/: $t_{228} = 7.56$ $p < 0.001$)

vowel	V	V(N)	N
a	103	115	83
i	80	100	83
u	73	100	90
e	92	110	83
o	90	109	87

Table 5: Duration statistics for single and geminate consonants.

	min	1Q	med	3Q	max	n
k	20	50	62	75	190	1304
kk	85	125	155	165	260	53
p	21	44	55	70	100	66
pp	70	120	135	151	190	60
t	15	40	55	71	150	916
tt	60	130	162	210	335	228
s	23	67	80	94	165	660
ss	97	131	146	159	198	10
sh	22	74	91	115	197	541
ssh	66	145	158	183	232	33

Table 6: Vowel durations before geminates:

	min	1Q	med	3Q	max	n
V-(CV)	19	70	85	105	235	7592
V-(QCV)	27	90	105	120	205	658
V-(CV2)	23	75	95	110	235	1385
V-(QCV2)	27	94	105	120	201	207

Table 7: Vowel durations after geminates:

	min	1Q	med	3Q	max	n
(C)V	19	75	90	107	240	7592
(CC)V	20	50	70	95	206	658
(C)V(V)	34	80	90	108	188	1385
(CC)V(V)	35	67	79	95	160	307

Table 8: Syllable durations factored by vowel type

vowel type	I	U	a	e	i	o	u
mean syldur	117.5	111.6	186.3	179.6	147.6	165.3	156.8

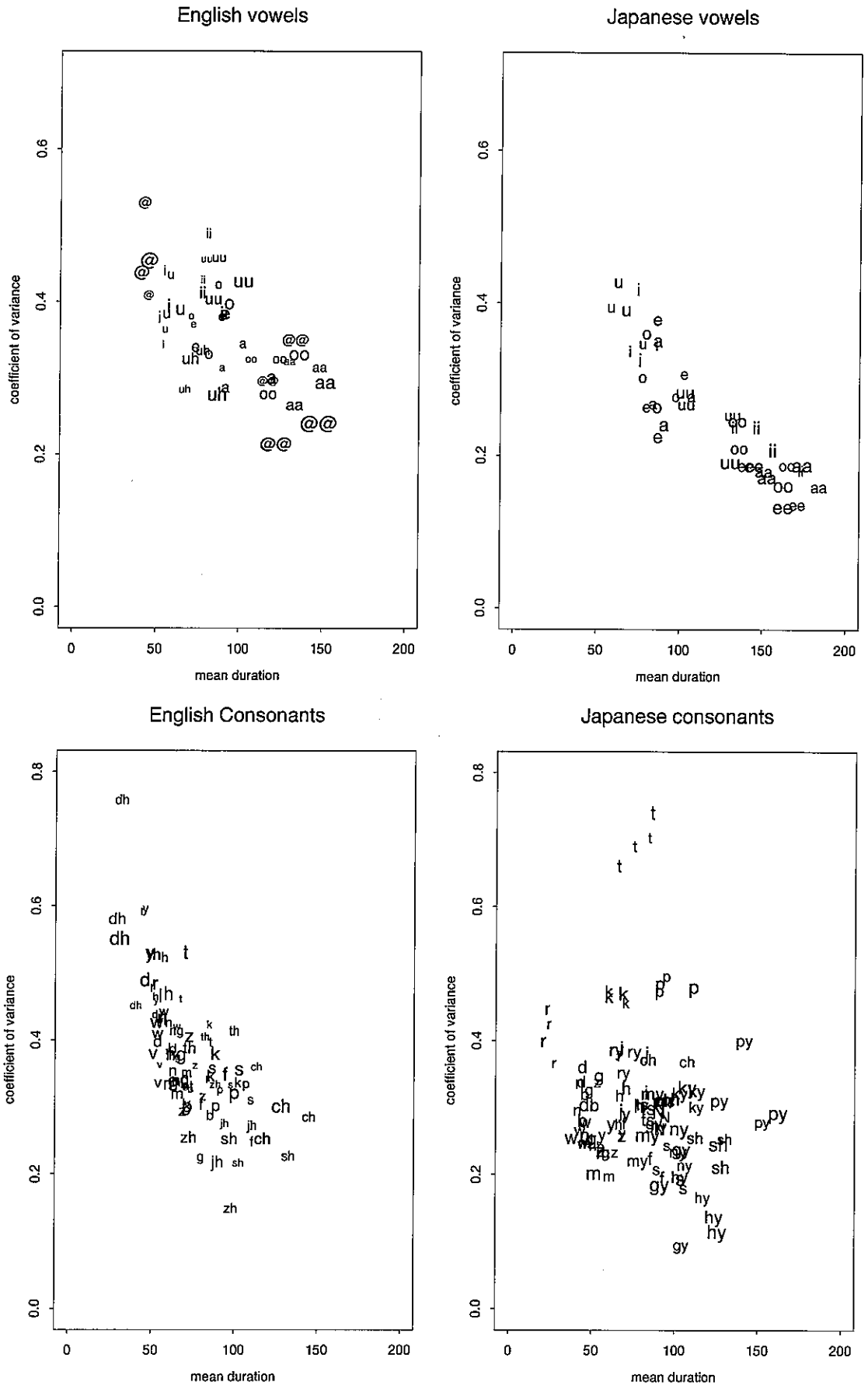

Figure 1: Japanese and English segmental durations from multiple speakers' data, showing mean durations for each phone (msec) plotted against the coefficient of variance. Machine-readable phonemic symbols are shown in the plot to identify the phones.

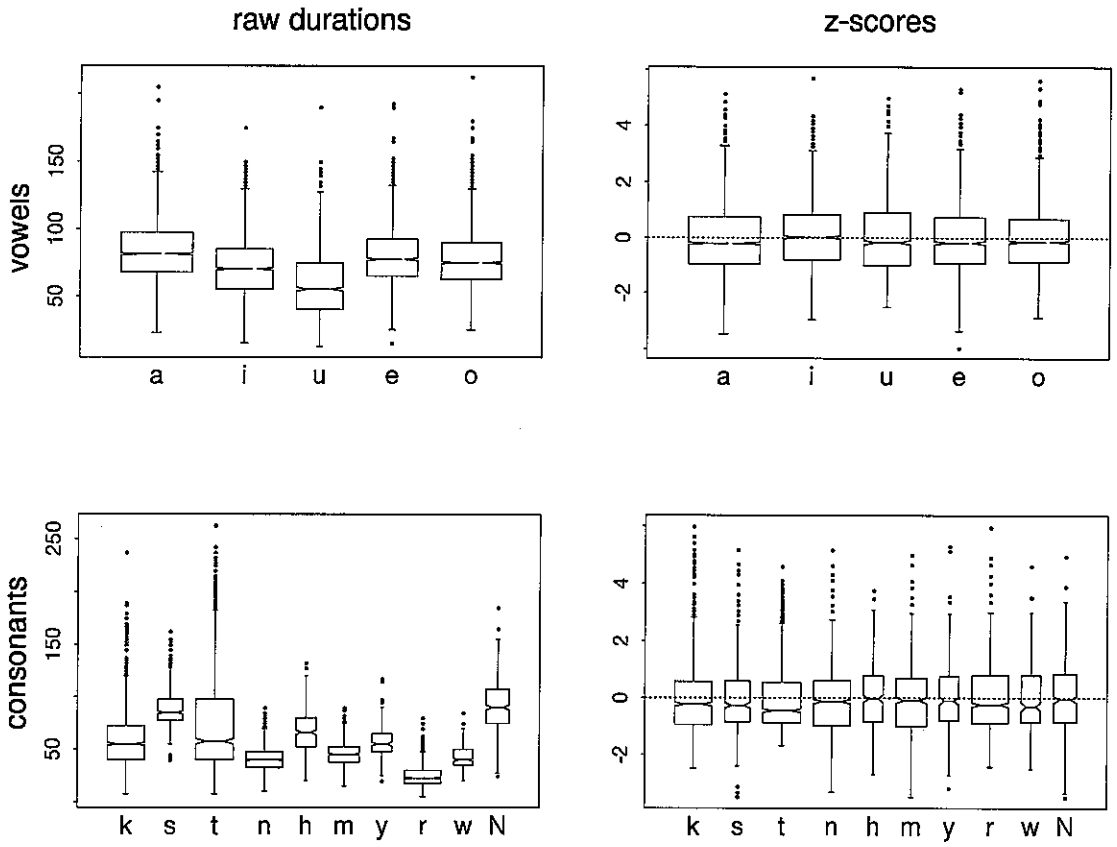

Figure 2: Segmental durations and equivalent z-scores for Japanese.

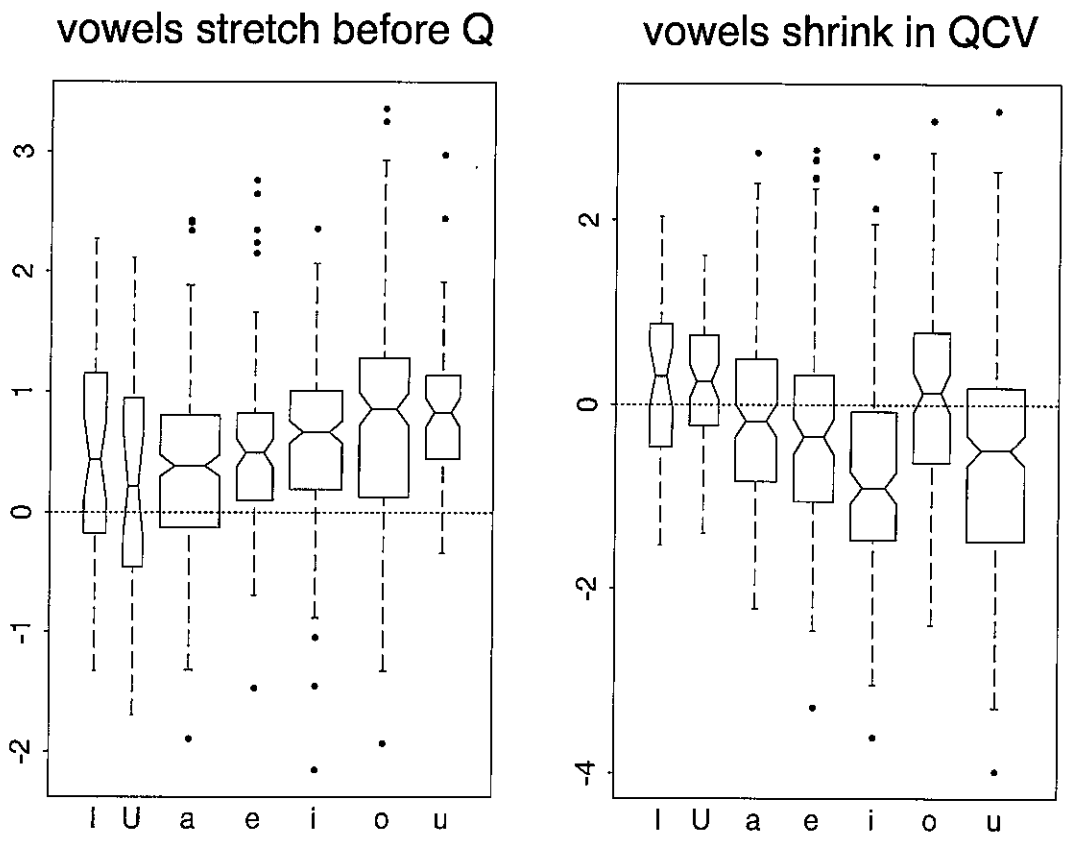

Figure 3: Vowel durations before and after a geminate

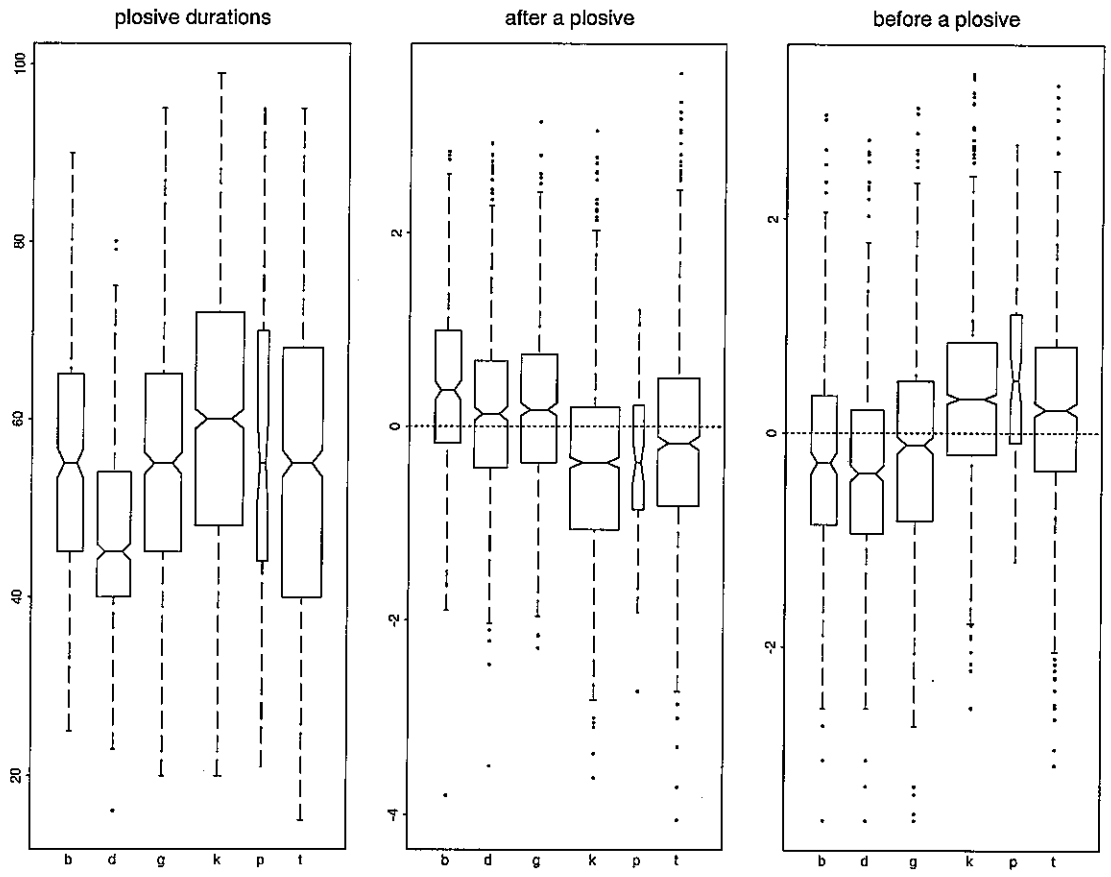

Figure 4: Vowel lengthening before and after a plosive