The phonetics of emphatic vowel lengthening in Japanese

Shigeto Kawahara and Aaron Braver
Rutgers University

Abstract

Many languages exploit a short vs. long lexical contrast in vowels. In most, if not all of these languages, the contrast is binary. In Japanese, however, speakers can lengthen vowels to express emphasis, and multiple degrees of lengthening can be used to express different degrees of emphasis. This paper offers the first experimental documentation of this emphatic vowel lengthening phenomenon. The current results demonstrate that, among the seven speakers recorded, at least a few speakers show six-levels of distinction in duration, and all but one speaker showed a steady linear correlation between duration and level of emphasis. We conclude that Japanese speakers have articulatory control that allows them to make very fine-grained durational distinctions, which go beyond mere binary short vs. long distinctions.

1 Introduction

Many languages distinguish short vowels from long vowels to make lexical contrasts, but these duration-based length contrasts are usually binary; e.g. [h**a**to] 'dove' vs. [h**aa**to] 'heart' and [ob**a**saɴ] 'aunt' vs. [ob**aa**saɴ] 'grandmother' in Japanese. While there is the rare typological exception such as Estonian, in which this contrast can be ternery (**?**), the distribution of superlong vowels is constrained by various prosodic and morphological factors (see **???** for discussion). **?**, p. 320 state that Mixe (**?**) is the only language that they know of that has a purely lexical duration-based three-way contrast (cf. **??**), although they also mention Yavapai (**?**) as another possible candidate. At any rate, three-way vowel length contrasts are rare at best cross-linguistically, and in the languages where they do exist, the ternary contrast is prosodically and/or morphologically restricted. As far as we know, there are no convincing cases of languages that make use of a purely lexical four-way (or greater) duration-based length contrast in vowels.[1]

[1] When two phonological contrasts interact, it is possible to have a four-way durational difference. For example, vowels are usually longer before voiced stops than before voiceless stops (e.g. **?????**). This lengthening effect may interact with a phonemic vowel length contrast to yield a four-way durational difference; e.g. VT < VD < VVT < VVD. What we do not observe, however, is one lexical contrast that is realized as a four-way durational differ-

In Japanese, however, speakers can use vowel lengthening to express emphasis. This process is commonly found in colloquial Japanese; a quick Google search (http://www.google.co.jp) with examples like [suɡ**oo**-i] (すごーい) ‘great’ and [çid**oo**-i] (ひどーい) ‘awful’ with lengthened stem-final vowels yields many hits. In addition, this pattern can manifest as multiple levels of emphasis (and therefore lengthening), extending beyond the familiar short/long binary distinction.[2]

This study offers the first experimental documentation of the vowel lengthening pattern.[3] One theoretical contribution of this paper is to investigate exactly how many levels of durational distinction Japanese speakers can make in expressing different degrees of emphasis—especially given that lexical vowel length contrasts are usually limited to a binary distinction in many languages, including Japanese.

Durational properties of Japanese short vowels and long vowels have been studied rather extensively in the previous literature both in terms of their production and perception (**??????????**). These studies have shown that duration is the major acoustic and perceptual correlate of short vs. long contrasts in Japanese, although there may be slight differences in formant characteristics as well, in such a way that long vowels are more dispersed in F1 and F2 dimensions than short vowels (**?**).

Although the phonetics of Japanese short and long vowels has been well studied in the past, to the best of our knowledge, there has not been experimental documentation of the emphatic lengthening pattern, which makes use of multiple levels of durational distinctions. One relevant study is **?** which tested the production of (heteromorphemic) sequences of the same vowels across morphemes in Japanese (e.g. *Matsu***e e e***jiten-wo okutta* ‘(I) sent a picture dictionary to Matsue’), and showed that Japanese speakers do make a distinction among 2 consecutive [e]s, 3 consecutive [e]s, 4 consecutive [e]s, and 6 consecutive [e]s in their production. Drawing on this study, our study below investigates vowel lengthening patterns with multiple levels, and shows that Japanese speakers can make similar fine-detailed durational distinctions even within single morphemes, and that this fine distinction can hold across a wider range of vowels in Japanese.

ence.

[2] Japanese speakers can also lengthen consonants to express emphasis (**????**). For a phonetic study testing different degrees of lengthening of Japanese consonants, see **?**. For a previous phonetic study investigating various acoustic properties of “paralinguistic focus”, which may be similar to what the current project examines, see **?**.

We also note, as we will discuss in section 5, that English has a similar process, as in *Thank you sooooooo much* and *She’s so cuuuuuuute*. See a post on Language Log by Mark Liberman (http://languagelog.ldc.upenn.edu/nll/?p=2006) for related observations. It is beyond the scope of the current study to conduct a cross-linguistic comparison, but a cross-linguistic study of this sort of lengthening phenomena is certainly hoped for.

[3] The current study focuses on the durational properties of the vowel lengthening pattern. See section 5 for discussion of other acoustic correlates that may possibly accompany the lengthening pattern.

2 Method

2.1 Stimuli

This study used emphasis of stem-final vowels in adjectives which are commonly observed in Japanese casual speech. The stimuli were grouped according to their final vowels, [a, o, u], which commonly appear stem-finally in Japanese adjectives.[4] For each vowel, two adjectives were chosen. The adjectives used in this experiment are listed in Table 1, where [-i] is an adjectival ending (present/non-past tense). All the stimuli were disyllabic and had a lexical pitch accent on the second syllable (i.e. the second syllable had an HL falling pitch contour). A subject noun was added to each adjective to make a complete sentence: e.g. [çiza-ɡa ita-i] '(I have) a knee pain'.[5]

Table 1: The list of the stimuli

[a]	[o]	[u]
[kata-i] 'hard'	[suɡo-i] 'great'	[nemu-i] 'sleepy'
[ita-i] 'aching'	[çido-i] 'awful'	[samu-i] 'cold'

In Japanese orthography, vowel length can be expressed with "ー" following the target vowel.[6] In this experiment, in addition to the non-lengthened rendition, five different degrees of emphasis were included as stimuli, as illustrated in Table 2.

Table 2: An illustration of one stimulus set in Japanese orthography

Japanese orthography	Transcription	Condition	G
a. いたい	[itai]	no emphasis	'p
b. いたーい	[itaai]	level 1 emphasis	'p
c. いたーーい	[itaaai]	level 2 emphasis	'p
d いたーーーい	[itaaaai]	level 3 emphasis	'p
e. いたーーーーい	[itaaaaai]	level 4 emphasis	'p
f. いたーーーーーい	[itaaaaaai]	level 5 emphasis	'p

There were a total of 36 stimuli (3 vowels * 2 adjectives * 6 emphasis levels). A random number was assigned to each stimulus item so that transcribers could later track which item had been produced.

[4] Most stem-final vowels in Japanese adjectives are back, although there are some exceptions (e.g. [samiɕi-i] 'lonely'.

[5] The target words were placed sentence-finally, and as a result some of them showed some creakiness and/or weakening (see e.g. **??**). Although this property of the stimuli did not cause a particular problem for the present acoustic analysis, a follow-up study which places the target stimuli in sentence internal positions may be worthwhile.

[6] A long vowel is written as a sequence of two letters in the case of the hiragana orthography. For example, [kaasaN] 'mother' is written in hiragana as 'かあさん' (k**a+a**+sa+N). Long [e:] and [o:] can also be orthographically expressed as *ei* and *oi* in some contexts. For example, *ou* [oo] 'king' would be written in hiragana as 'おう' (o + u) and *eiga* [eega] 'movie' as 'えいが' (e + i + ga). In loanwords as well as in this expressive emphasis pattern, however, the length mark (ー) is used to express vowel length. See **?** for a recent explanation (in English) of the Japanese orthographic system.

2.2 Participants

The participants were seven native speakers of Japanese (anonymously coded as Speakers TF, TN, TX, TW, TT, SX, TV). They were all undergraduate students at International Christian University (Tokyo, Japan). They were paid 500 Japanese yen for their time. They all signed a consent form before participating in the experiment.

2.3 Procedure

The recording sessions took place in a sound-attenuated room at International Christian University. The stimuli and all instructions were presented in Japanese orthography using Superlab ver. 4.0 (**?**). In the instructions, speakers were told that the experiment was about multiple levels of emphasis in Japanese, and that they were going to read sentences with vowels of differing length. They were instructed to read the whole frame sentence, not just the target words, for each stimulus.

Each block contained one token of every stimulus item. The speakers were allowed to take a short break after each block. The order of the stimuli within each block was randomized by Superlab. The speakers went through ten blocks, which resulted in a total of 360 tokens (36 stimuli*10 repetitions). Each speaker was assigned 30 minutes for the experiment.

Before the main session, as practice, each speaker read all the stimuli once to familiarize themselves with the stimuli and the task. After the practice phase, the experimenter (the first author) clarified any questions that they had. Speakers were recorded directly via a portable recorder (TASCAM DR-40) with a 44k sampling rate and a 16 bit quantization level. The first author sat with each speaker throughout the experiment to monitor the progress of the recording.

2.4 Acoustic analysis

The duration of each stem-final vowel plus the adjectival suffix [i] was measured. We did not attempt to put a boundary between the stem final vowels and the suffixal [i], because the transitions from the stem vowels into the suffixal [i] were blurry (a vowel-to-vowel transition is generally blurry and hard to unambiguously locate in an acoustic analysis: **?**). However, since only the stem-final vowels were emphasized, and not the suffixal vowel (see Table 2), the duration of [i] should be more or less constant across all conditions. Vowel onset and offset were determined by inspecting both waveforms and spectrograms, and the boundaries were placed where F2 and F3 (dis-)appear. Sample spectrograms are shown in Figure 1. After the segmental boundaries were placed, the durations of the target intervals were automatically extracted. Acoustic measurements were done using Praat (**?**).

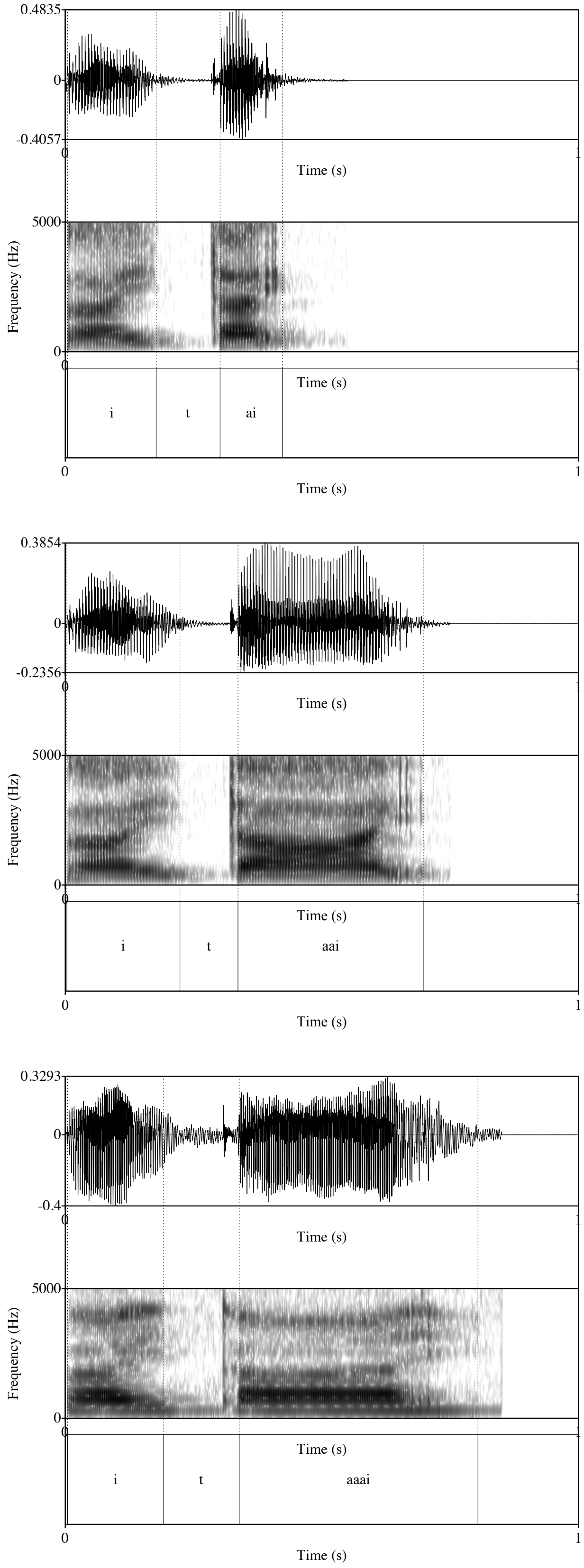

Figure 1: Sample spectrograms: no-emphasis, emphasis level 1, emphasis level 2. The ti

2.5 Statistics

Since there are many comparisons (6 levels of emphasis * 3 types of vowels * 7 speakers), no pair-wise comparisons at each emphasis level were conducted, in order to avoid Type I error (i.e. to avoid finding some significant effects by chance). However, error bars, which represent 95% confidence intervals, are provided in the result figures. They were calculated over 20 repetitions of each vowel (2 adjectives * 10 repetitions), except when speakers mispronounced some relevant token. A post-hoc inspection of the data showed that a linear regression analysis would be useful, so they are reported in the results section. All statistical analyses were performed using R (**?**). R was also used to generate result figures.

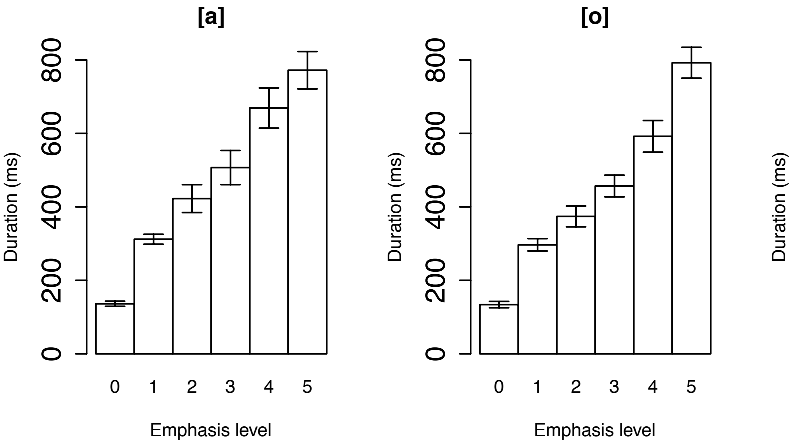

Figure 2: The average durations of each emphasis level with 95% confidence intervals: Speaker TF.

3 Results

Since different speakers showed different patterns, we report the results of individual speakers separately, and present a summary in section 4 after reporting the results of individual speakers. We start first by discussing those speakers who showed the clearest distinctions among the different emphasis levels. First, as shown in Figure 2, Speaker TF seems to make a perfect six-way distinction; i.e., the vowel durations for each level of emphasis are different from those of every other level of emphasis for this speaker, and error bars do not overlap.

There are large jumps in duration from the non-emphatic level to the first level of emphasis; with each additional degree of emphasis, there is a shorter, but steady, increase in duration.

To assess the correlation between emphasis level and duration, a linear regression analysis was run with vowel duration as the dependent variable, and emphasis level as the independent variable. Since the increase from non-emphatic vowels to the first level of emphasis is non-linear, they were excluded from this regression analysis. The coefficient estimate of the regression analysis is 120 ms ($t(247) = 30.8, p < .001$). This correlation estimate represents an average durational increase per emphasis level for this speaker. In other words, it estimates that for each level of emphasis, vowel duration should increase by 120 ms. The correlation between duration and emphasis level is very high (r=.89), showing that the linear relationship between durational increase and emphasis level is very strong.

As shown in Figure 3, like Speaker TF,

Speaker TX shows a six-level distinction among emphatic vowels. The average duration for each condition differs, and error bars barely overlap. In the regression analysis, the coefficient estimate is 105ms ($t(245) = 20.2, p < .001$), and the correlation estimate r is .79. As with Speaker TF, there are large durational jumps from non-emphatic to emphatic vowels. The emphatic vowels show steady, linear increases in duration, except for exceptionally large differences between emphasis level 4 and emphasis level 5. These large, non-linear jumps may be responsible for the lower r-value compared to that of Speaker TF. Presumably, for this speaker, the most emphatic vowel has a special status, so it receives extra lengthening.

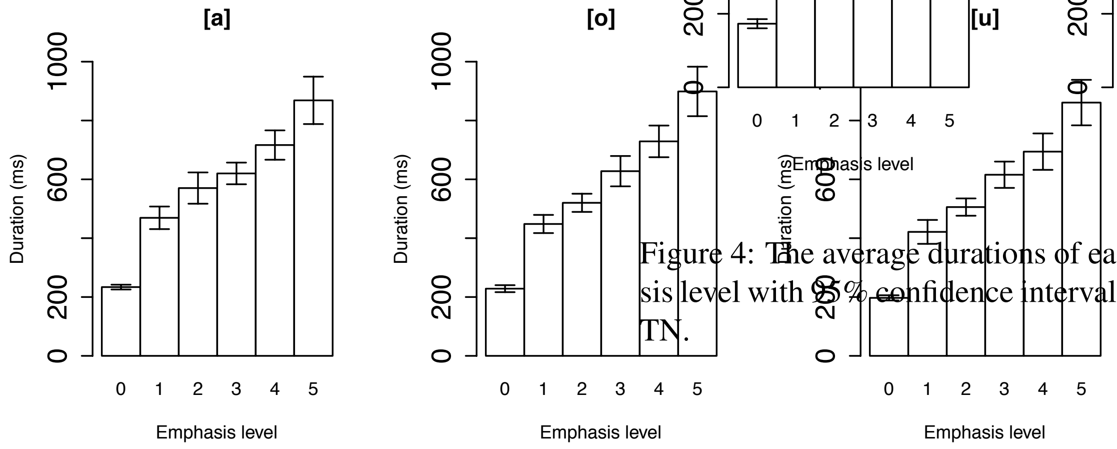

Figure 3: The average durations of each emphasis level with 95% confidence intervals: Speaker TX.

Speaker TN, as shown in Figure 4, showed the next clearest increase in duration as the emphasis levels go up. Although the speaker generally does not show a clear difference between level 1 and level 2 emphasis for any of the three vowels, the speaker nevertheless seems to make a difference between the other emphasis levels. This speaker also makes an exceptionally large increase from level 4 to level 5 for [a]. In the regression analysis, the coefficient estimate is 78ms ($t(230) = 20.7, p < .001$), and the correlation coefficient r is .81.

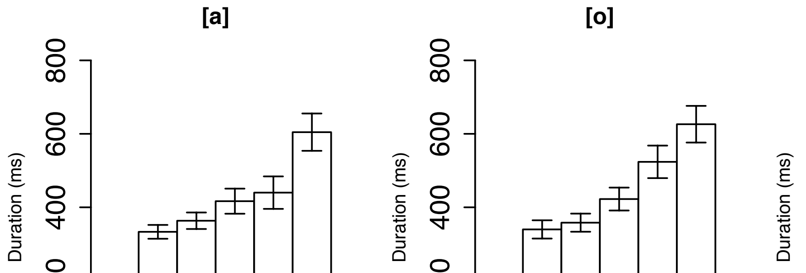

Figure 4: The average durations of each emphasis level with 95% confidence intervals: Speaker TN.

Speaker TW did not show differences between level 1 and level 2 (or level 3 for [u]), as illustrated in Figure 5. It is as though this speaker was treating these levels of emphasis as one category of emphasis. However, the speaker did make a distinction between other levels of emphasis. The correlation between emphasis level and duration is therefore still high (r = .76). In the regression analysis, the coefficient estimate is 51ms ($t(240)$ = 17.9, p < .001). The smaller estimate is also reflected in this speaker's duration range; in Figure 5, the duration range is about 600ms, whereas for the previous speakers, the duration ranges are between approximately 800ms and 1000ms (Figures 2-4).

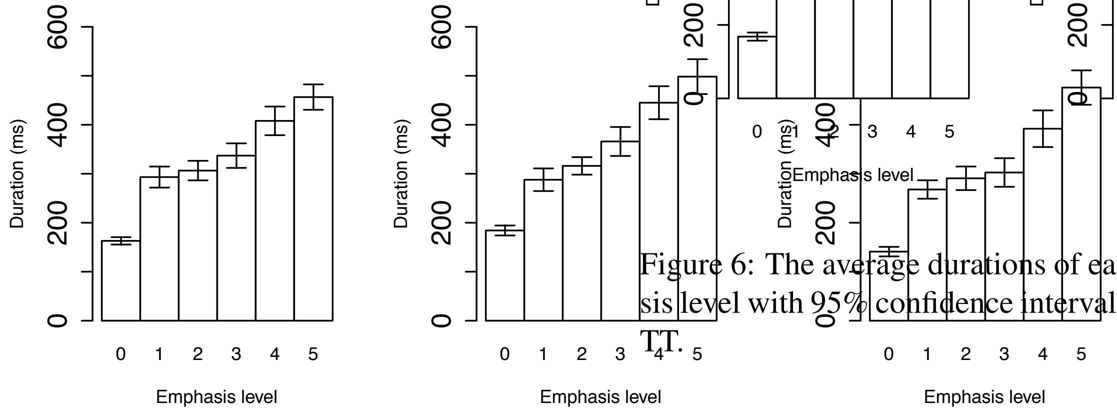

Figure 5: The average durations of each emphasis level with 95% confidence intervals: Speaker TW.

As shown in Figure 6, Speaker TT shows the next highest correlation between emphasis level and duration (r = .66). This speaker shows large variability in several conditions (as represented in the size of the error bars for these conditions); e.g. emphasis level 5 for [a], and at all emphasis levels for [u]. This speaker also does not show a difference between level 1 and level 2 for [u]. These behaviors may be responsible for the lower r-value of this speaker compared to those discussed above. In the regression analysis, the coefficient estimate is 75ms ($t(242)$ = 13.8, p < .001).

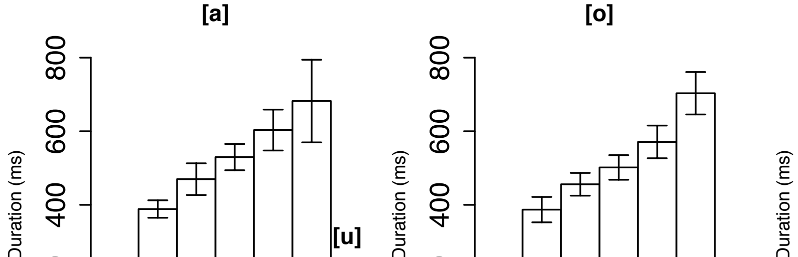

Figure 6: The average durations of each emphasis level with 95% confidence intervals: Speaker TT.

As shown in Figure 7, Speaker SX does not show differences between several of the conditions: between level 2 and level 3 for [a], between level 4 and level 5 for [o] and between level 3 and level 4 for [u]. The lack of differences in these conditions resulted in an *r*-value that is lower than previous speakers ($r = .61$); however, this linear correlation is still high. In the regression analysis, the coefficient estimate is 27ms ($t(248) = 12.1, p < .001$).

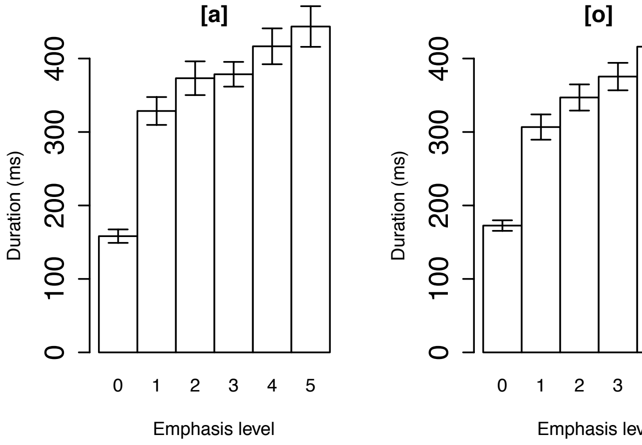

Figure 7: The average durations of each emphasis level with 95% confidence intervals: Speaker SX.

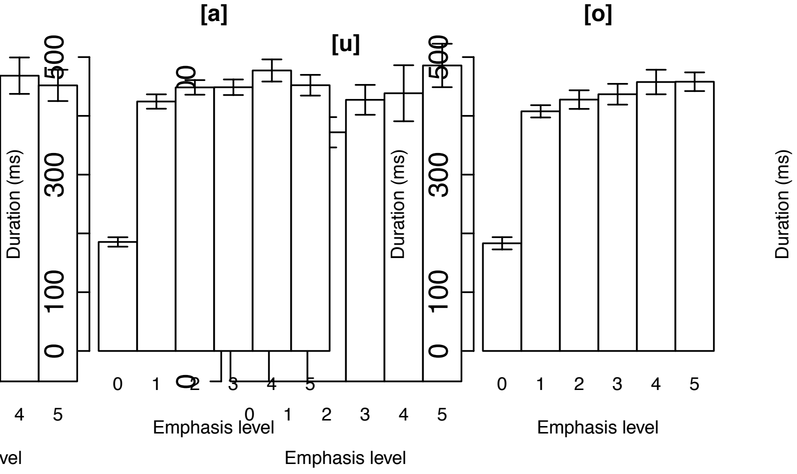

Figure 8: The average durations of each emphasis level with 95% confidence intervals: Speaker TV.

Finally, as shown in Figure 8, Speaker TV shows a more or less binary distinction—i.e. non-emphatic vs. emphatic—although we do observe a slight increase in duration as emphasis levels go higher ($r = .41$). Indeed the regression analysis reveals that the coefficient estimate is as low as 12ms, although it did reach statistical significance ($t(245) = 7.2, p < .001$).

4 Summary

Table 3 provides a summary of each speaker's data. It provides a regression function, an *r* value as a measure of the strength of the linear correlation between emphasis levels and duration, and maximum duration (token-wise) as a measure of their duration range—the range each speaker is willing to use for the emphatic vowels.

Table 3: The summary of each speaker's behavior

Speaker	Regression function	*r*	Max duration (ms)
Speaker TF	$y = 152 + 120x$	.89	975
Speaker TN	$y = 217 + 78x$	.81	782
Speaker TX	$y = 320 + 105x$	.79	1301
Speaker TW	$y = 209 + 51x$	.76	670
Speaker TT	$y = 299 + 75x$	.66	1055
Speaker SX	$y = 291 + 27x$	.61	603
Speaker TV	$y = 395 + 12x$	.41	533

In spite of some inter-speaker variability, all speakers showed a positive, steady correlation between level of emphasis and vowel duration. Speakers TF and TX showed a perfect six-way durational distinction, without much overlap in error bars. While other speakers did not show all these distinctions quite as clearly, they showed a (mostly) steady linear increase in duration as emphasis levels increased. Furthermore, all speakers except Speaker TV made an at least 5-way distinction: they either had all levels distinguished, or did not show a difference between two (but not more than two) adjacent levels (with the potential exception of [u] for Speaker TW). On the other hand, Speaker TV appeared to make an (almost) binary distinction between emphasized and non-emphasized vowels. Overall, there were no evident significant reversals, where higher emphasis levels would have shown shorter durations (perhaps except for Speaker TV's [a], level 4 and level 5).

In Table 3, we can observe that there is some association between the strength of correlation (*r*) and the maximum duration a speaker used; for example, Speaker TF, who showed the highest correlation, used a large duration range, whereas Speaker TV, who showed the lowest correlation, used the smallest duration range. The correlation is not perfect, however, since Speaker TT showed the second-largest duration range, yet this speaker has the third-lowest *r*-value.

5 General discussion

5.1 Summary

The current study, to the best of our knowledge, has provided the first experimental description of the emphatic vowel lengthening pattern in Japanese. Although there is some inter-speaker variability, several speakers were able to make durational distinctions as fine-grained as six-ways. Other speakers showed a positive correlation between durations and emphasis levels, all to a statistically significant degree. These patterns are in line with the conclusions drawn from a companion study on emphatic consonant lengthening in Japanese (**?**), which used a method that is similar to the current experiment to measure the duration of Japanese consonants with multiple degrees of emphasis. (The current speakers and those who participated in **?** do not overlap.)

Taken together, one general implication of our current study, beyond providing an experimental description of the Japanese lengthening pattern, is that the current results show that Japanese speakers have articulatory controls which enable them to potentially make six-way durational distinctions.

5.2 Further questions

One question that arises, given that speakers can make such fine-grained durational distinctions, is why natural languages generally deploy only a two-way distinction for lexical contrasts (as discussed in the introduction). One possible answer to this question is that a three way durational contrast may be difficult to unambiguously perceive in real communicative situations—in other words, perceptual distinctiveness restricts a range of possible contrasts that the grammar can deploy (see e.g. **?????????**; see especially **?**, **?**, and **?** for the grammatical imperatives on perceptual dispersion in durational contrasts). For a discussion of an alternative, more formally-based explanation, see the companion paper (**?**).

The current study also raises many questions which should be addressed in future studies. For example, would Japanese listeners be able to track these different degrees of emphasis? The current experiment used only up to 5 levels of emphasis, but given how well some speakers performed, what would the real limit for Japanese speakers be? Although the current paper focused on vowels only, it is possible to lengthen consonants (**?**), and it is also possible to lengthen both vowels and consonants: e.g. [sugg**oo**-i] (すっごーい). How vowel lengthening and consonant lengthening interact is an interesting question. Also, it is possible to lengthen stem-initial vowels [s**uu**go-i] (すーごい) instead of stem-final vowels [sug**oo**-i] (すごーい). Whether position of emphasis affects durational manifestations of vowels is question worth pursuing. Additionally, the differences, if any, between lengthened vowels and sequences of (heteromorphemic) vowel sequences (**?**), merits investigation. Toshio Matsuura (p.c.) offers an example paradigm in Table 4 to address this last question.

Table 4: An illustration of one stimulus set in Japanese orthography

Japanese orthography	Transcription	Condition	Gloss
a. 甥と言った	[**oi**-to-itta]	1 [o]	"I said
b. 遠いと言った	[**too-i**-to-itta]	2 [o]s	"I said
c. 子を置いた	[k**o-o-oi**ta]	3 [o]s	"I place
d 甲を置いた	[k**oo-o-oi**ta]	4 [o]s	"I place
e. 憎悪を置いた	[z**oo-o-o-oi**ta]	5 [o]s	"I set a
f. 法王を置いた	[h**oo-oo-o-oi**ta]	6 [o]s	"I set a

Comparing a paradigm like the one in Table 4, which contains heteromorphemic strings of up to six [o]s, with our current results, which show six-way contrasts within a single adjective, may reveal interesting effects of morphological boundaries on phonetics (see e.g. **????**).

Moving beyond Japanese, would we expect speakers of other languages be able to produce similar durational differences (and would they make as many levels of distinction)? Would other languages draw the boundaries between each durational level at the same place? Would there be a difference between languages that exploit duration-based contrasts (as in Japanese) and those that do not? In English, for example, we observe examples like: *Thank you sooooooo much*, *I looooooove you* and *She's so cuuuuuute*. Given these stimuli, would English speakers make distinctions similar to those of the Japanese speakers tested in this experiment?

Finally, as an anonymous reviewer points out, semantic focus can be realized in acoustic dimensions other than duration; e.g. stronger intensity and pitch range expansion (see **????** among many others). It remains to be investigated how Japanese speakers (and speakers of other languages, for that matter) make use of these acoustic dimensions to express the sort of emphasis investigated in this paper. Furthermore, **?** show that long vowels are more dispersed in their F1 and F2 dimensions than short vowels in Japanese. Thus, the effects of emphatic vowel lengthening on formant displacement should be explored in future studies. All of these are interesting questions, which are, however, beyond the scope of the current study.

5.3 A final remark

We would like to close with a remark about the distinction between non-emphatic vowels and emphatic vowels. Recall that all the speakers produced the emphatic vowels as longer than the non-emphatic vowels, despite the fact that not all speakers realized differences among all different levels of emphasis. Moreover, as observed in all the figures, all speakers showed a very large increase in duration from non-emphatic vowels to emphatic vowels, and this increase is larger than the observed differences between the various levels of emphatic vowels. We therefore suggest that Japanese speakers overall make a binary distinction between emphatic and non-emphatic durations, and within the emphatic durations, speakers differ in how to acoustically realize the degrees of emphasis. This conclusion may imply that, semantically speaking, the difference between non-emphatic and emphatic is more important than different degrees of emphasis. Further, Japanese speakers attempt to reflect this difference in semantic importance in their production of emphatic and non-emphatic vowels. Again, we find the same patterning in the companion study on consonant lengthening (**?**), which reinforces this conclusion.

Acknowledgments

The recording for this experiment was conducted while the first author was a visiting scholar at International Christian University, which was made possible by a fellowship offered by the Japan ICU Foundation (JICUF).

We also thank Shin-ichiroo Sano and Tomo Yoshida for their support in arranging the recording sessions. Research assistants at the Rutgers Phonetics Laboratory—Nat Dresher, Chris Kish, Sarah Korostoff, Michelle Marron, Jessica Trombetta, and especially, Melanie Pangilinan—offered indispensable help with the acoustic analysis. We thank John Kingston for comments on this project, as well as Kyle Gorman, Manami Hirayama, Chris Kish, Hope McManus, Toshio Matsuura, Atsushi Oho, Akiko Takemura, and anonymous reviewers for comments on earlier versions of this paper. The publication of this paper was supported by startup funds to the first author from Rutgers University. Remaining errors are ours.