Phonological Data and Analysis Volume XX, Article X: X–XX (20XX)
https://doi.org/00.0000/nn.0.00

Received XX January 2018; revised XX May 2018; accepted XX January 2019.

Acoustics and perception of emphatic lengthening in English

Languages can signal two different kinds of meanings by way of phonetic duration. The first kind is a lexical, phonological contrast on vowels and consonants; i.e., short vowels vs. long vowels and singleton consonants vs. geminate consonants. This sort of contrast is usually limited to a binary distinction, and its phonetic properties have been well studied for many different languages. The other use of phonetic duration is to express pragmatic emphasis, as in *Thank you soooo much*. This second use of duration has been very much understudied, especially in languages which do not exploit lexical durational contrasts. Building on previous studies on similar phenomena in Japanese (Kawahara & Braver 2013, 2014), this paper reports experiments on emphatically-lengthened words in English. The results of a production experiment show that, when prompted, at least some speakers can produce six distinct levels of durational differences. One general implication of this ability is thus that speakers have a degree of articulatory control that allows them to make a contrast that goes beyond the standard binary contrast, even when their native language does not possess a purely duration-based contrast. A perception experiment shows, however, that it is hard to pin down precisely which level of emphasis a listener hears—listeners appear to group stimuli into up to three categories. This difficulty in identifying multiple levels of duration distinctions may underlie the fact the lexical duration contrasts are usually binary.

Keywords: Emphatic lengthening; English; duration; length contrasts; contrast dispersion

1 Introduction

Languages make use of phonetic duration to signal two different kinds of meanings. The first kind is to make a lexical contrast: some languages—like Japanese—contrast short vowels with long vowels, and/or short consonants with long consonants (see Kawahara 2015 for a review). For instance, in Japanese [obasan] 'aunt' with a short vowel contrasts with [obaasan] 'grandma' with a long vowel, and [kata] 'frame' with a short [t] contrasts with [katta] 'bought' with a long [t]. The phonetic properties of these lexical contrasts have been studied for many languages, both on consonants and vowels (see Kawahara & Braver 2014 for a recent summary of previous studies on consonantal length contrasts in a variety of languages).

Less well studied is the use of duration to signal pragmatic emphasis. For example, English speakers can say Thank you sooooo much to express an emphasized degree of gratefulness. This use of duration for pragmatic emphasis has been very much understudied in the phonetic literature (though see Fuchs, Savin, Solt, Ebert & Krifka 2019 and Samejon 2019 for recent examples). The bulk of the existing literature on pragmatic emphasis in Japanese is primarily based on impressionistic observations and does not offer substantial quantitative or experimental analysis (Higuchi & Haraguchi 2006; see Nasu 1999, and Kawahara 2013 for phonological analyses.). In two recent studies, however, the phonetic properties of such pragmatic lengthening have been studied for Japanese by Kawahara & Braver (2013, 2014), who found that at least

some Japanese speakers can make six-way durational contrasts, both on vowels and consonants, to express different degrees of emphasis. In these experiments, Japanese speakers were asked to produce utterances with different degrees of emphasis, expressed by way of gemination marks and long vowels, ranging from no emphasis to levels 1 through 5 of emphasis. Some illustrative figures from these previous studies are reproduced here in Figures 1 and 2, which show that unlike the standard binary lexical contrast, Japanese speakers can make up to six-way durational contrasts, both with consonants and vowels.

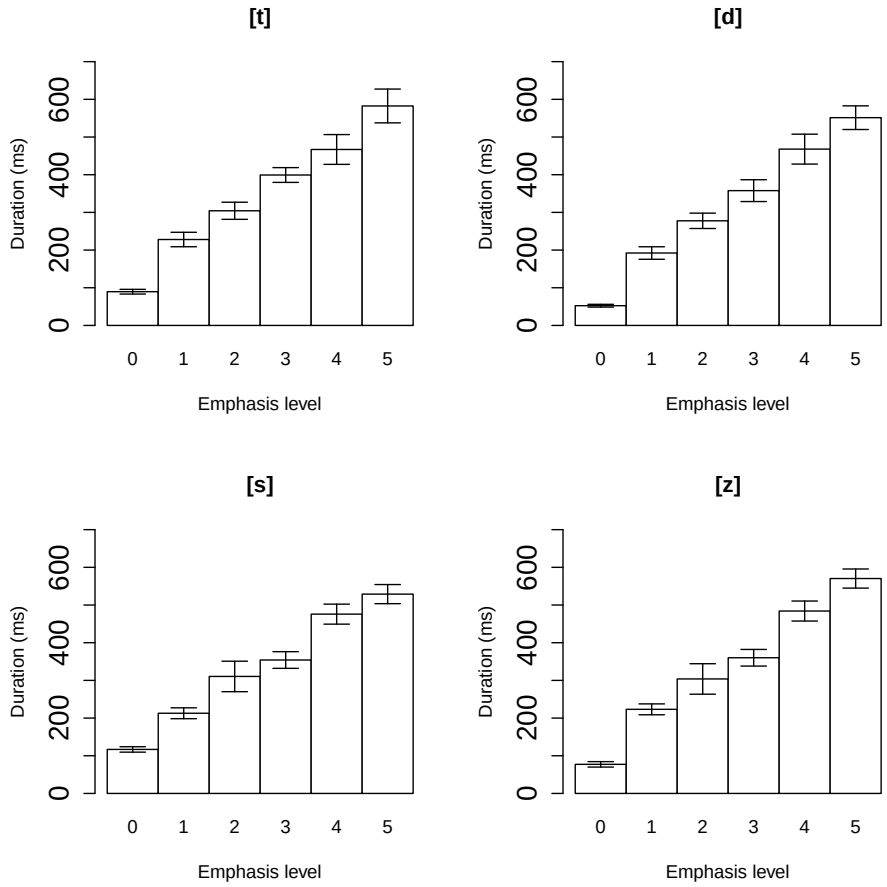


Figure 1: A Japanese speaker making a six-way durational contrast on consonants to express different degrees of emphasis. Reproduced from Figure 3 of Kawahara & Braver (2014).

One question that remains from these studies is whether Japanese speakers are able to make these distinctions because their native language makes use of purely durational contrasts, or whether speakers in general possess the ability to do so, regardless of their native language. To address this question, the current study investigates whether speakers of a language which does not exploit lexical durational differences—English—can make similar durational distinctions. Also unanswered in the previous literature is the question of perception: given six-way durational differences, can listeners perceive these fine-grained distinctions? The current study was designed to address these two questions.

One larger question that lurks behind this general project is the question of why there is such an overwhelming cross-linguistic preference for binary length distinctions. Duration contrasts—be they in consonants or in vowels—tend, cross-linguistically, to be binary. There are a few rare typological exceptions to this claim such as Estonian, in which this contrast may be ternary (Prince 1980), but in general the distribu-

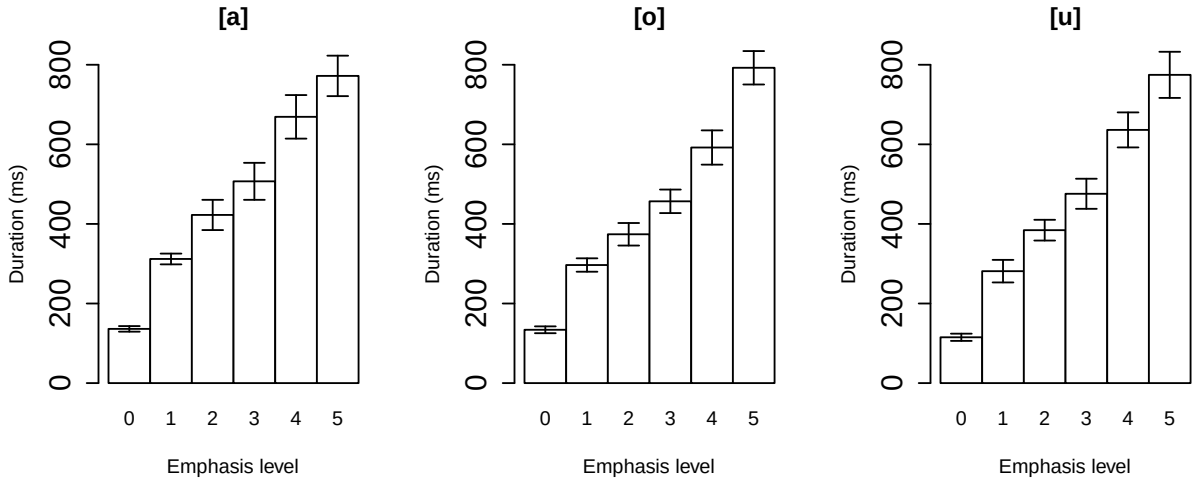


Figure 2: A Japanese speaker making a six-way durational contrast on vowels to express different degrees of emphasis. Reproduced from Figure 2 of Kawahara & Braver (2013).

tion of such contrasts is restricted by various prosodic and morphological factors (Ladefoged & Maddieson 1996; Lehiste 1970; Prince 1980). Additional possible exceptions noted by Ladefoged & Maddieson (1996) include Mixe (Hoogshagen 1959; cf. Jany 2007) and Yavapai (Thomas & Shaterian 1990). In any case, the vast majority of vowel duration contrasts are binary, and in those languages where ternary contrasts exist, such contrasts are prosodically and/or morphologically restricted. We know of no convincing cases of four-way or greater lexical duration contrasts.[1] Why should this be? One potential answer to this question is that three-way or greater durational contrasts may be difficult to produce or perceive. The experiment presented here shows that speakers can, in fact, produce fine-grained durational distinctions, suggesting that restrictions on production are not (solely) responsible for the preference for binary duration contrasts. The perception experiment shows that it is indeed difficult to precisely pin down which level of emphasis was uttered; i.e., it is difficult to categorize so many different levels of durational differences.

2 Experiment 1: Production study

The first experiment examined the extent to which English speakers can produce fine-grained durational distinctions to express pragmatic emphasis. If they are able to do so, it would suggest that the ability to produce these distinctions does not depend on speaking a language with a (binary) lexical length contrast like Japanese.

[1] Four-way durational contrasts may appear to exist in cases where two phonological contrasts interact. For example, vowels tend to be longer before voiced stops than before voiceless stops (Chen 1970), and this lengthening effect may interact with a phonemic vowel length contrast to yield a four-way durational distinction (e.g., VT $<$ VD $<$ VVT $<$ VVD). However, we never observe a single durational lexical contrast that is realized as a four-way durational distinction.

2.1 Method

2.1.1 Stimuli

The target words were seven English adverbs, which are used emphatically in daily, casual speech of (New Jersey) English: mad, very, too, so, way, super, and really.[2] They were embedded in frame sentences, as shown in the leftmost column of Table 1. Each sentence was then modified to create 5 levels of emphasis by orthographically lengthening the target adverb. For example, the target adverb so was placed into the frame sentence That guy is so creepy (no emphasis), and five additional sentences were created, replacing so with soo, sooo, soooo, sooooo, and soooooo to represent emphasis levels 1–5.

No emphasis	Level 1	Level 2	Level 3	Level 4	Level 5
That bag is **mad** expensive.	maad	maaad	maaaad	maaaaad	maaaaaad
That model is **very** tall.	veery	veeery	veeeery	veeeeery	veeeeeeery
That baby is **too** cute.	tooo	toooo	tooooo	toooooo	tooooooo
That guy is **so** creepy.	soo	sooo	soooo	sooooo	soooooo
That band is **way** cool.	waay	waaay	waaaay	waaaaay	waaaaaay
That joke is **super** funny.	suuper	suuuper	suuuuper	suuuuuper	suuuuuuper
That lecture was **really** boring.	reaally	reaaally	reaaaally	reaaaaally	reaaaaaally

Table 1: Stimuli used in Experiment 1. The frame sentences appear on the leftmost column.

2.1.2 Speakers

Eight native speakers of New Jersey English participated in this production study. Participation was limited to female speakers as the use of intensifiers has been argued to be characteristic of feminine speech (Jespersen 1922; Lakoff 1973; Tagliamonte & Roberts 2005, though see Xiao & Tao 2007:242–243 for further discussion), and anecdotally this emphatic lengthening carries a similarly gendered connotation.

2.1.3 Recording

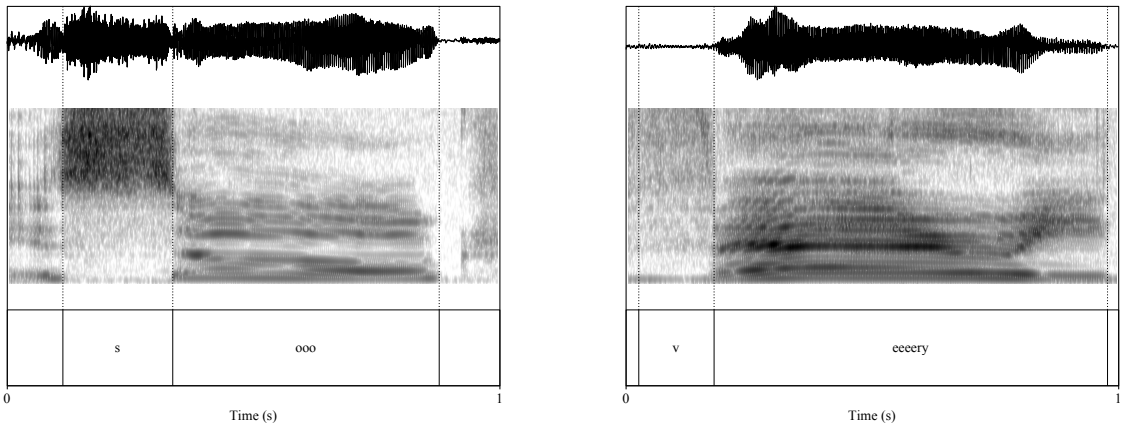
All stimuli (7 adverbs $\times$ 6 levels of emphasis = 42 stimuli) were randomized by Superlab (Cedrus Corporation 2010), and were visually presented to speakers for oral production. Speakers produced all stimuli 10 times, with order of stimuli randomized within each repetition. Recordings were performed in a sound-attenuated booth using an AT44040 cardioid capacitor microphone with a pop filter, amplified through an ART TubeMP microphone pre-amplifier and JVC RX-554V receiver. The speech was digitized as WAV files at a sampling rate of 44.1kHz using Audacity.

2.1.4 Analysis

Acoustic analysis was performed using Praat (Boersma 2001). For adverbs with a monophthong that is emphatically lengthened (e.g. *mad*), the duration of that vowel was measured. For diphthong target vowels (e.g., *way*), the duration of the entire diphthong was measured due to the difficulty of placing a boundary between the main vowel and the following offglide (Turk, Nakai & Sugahara 2006). For the adverbs *very* and *really*, since boundaries between vowels and [ɹ, l] are difficult to determine, duration was measured by

[2] The use of the adverb mad is generally restricted to the greater New York area. All participants in this study were from New Jersey, where this use of mad is common.

calculating the duration of the entire word and subtracting the duration of the initial consonant (e.g., very). Sample waveforms and spectrograms are given in Figures 3a and 3b to illustrate our segmentation procedure.



(a) *so*, level 2 emphasis, Speaker 1

(b) *very*, level 3 emphasis, Speaker 8

Figure 3: Waveforms and spectrograms illustrating our segmentation procedure. Time scales both 1000 ms.

Following previous work on emphatic lengthening (Kawahara & Braver 2013, 2014), we provide the Pearson correlation (r) as a measure of correlation between emphasis level and duration. In this calculation, the non-emphatic condition was excluded because the relationship between this condition and the emphatic conditions is non-linear, as we will discuss below. We also ran a regression analysis on the emphatic conditions to estimate how many milliseconds speakers increase their vowel duration per emphasis level. We also conducted non-paired t-tests for each speaker comparing each successive pair of emphasis levels (e.g., level 0 vs. level 1, level 1 vs. level 2…). In order to avoid Type I error we Bonferroni-adjust our significance level to $\alpha = .01$ (each speaker has 5 comparisons; 0.5/5). A table of these comparisons is available in Appendix A.

2.2 Results

2.2.1 Individual patterns

Since there is non-trivial inter-speaker variation, similar to what was found for Japanese by Kawahara & Braver (2013, 2014), we discuss the behavior of each speaker in turn in this section; summary and comparison of all speakers is provided in section 2.2.2. We start with speakers who show the clearest distinctions among different levels of emphasis.

Three speakers—Speakers 1, 5, and 6—show a clear six-way durational contrast (illustrated in Figures 4, 5, and 6). Beginning with Speaker 1, we observe (a) that there is a fairly substantial increase in duration between the non-emphatic condition and the first level of emphasis (just as with Japanese speakers, see Figures 1 and 2), and (b) there is a steady increase in duration from one emphasis level to the next. The r-value for Speaker 1, assessing correlation between level of emphasis and duration, is .75, which is statistically significant ($p < .001$). The regression analysis shows a best-fitting coefficient of 92 ms, suggesting that for Speaker 1, each additional level of emphasis corresponds to approximately 92 ms of additional duration. Non-paired t-tests show that comparisons between each successive emphasis level are significant at the level of $p < .001$ (see the table of comparisons in Appendix A for details).

Speakers 5 and 6 perform almost as well as Speaker 1, as illustrated in Figures 5 and 6, respectively. Like Speaker 1, they have a large gap between the no-emphasis condition and the first level of emphasis,

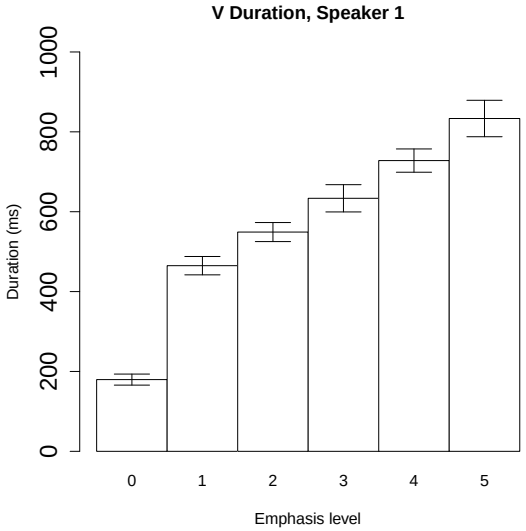


Figure 4: Speaker 1: $r = .75$, regression coefficient = 92 ms. (Error bars here and throughout indicate 95% confidence intervals.)

and duration increases along with emphasis level. The Pearson r-values for the emphatic conditions are 0.73. (Speaker 5) and 0.72 (Speaker 6), which are both significant at the $p < .001$ level. The regression coefficients are 76 ms and 65 ms, respectively. We note that in Figure 6, the error bars for the level 4 and level 5 conditions overlap, suggesting that while Speaker 5 clearly differentiates 6 levels of emphasis, Speaker 6 only produces 5 levels of emphasis. This is reflected in the results of non-paired t-tests between each successive emphasis level—for Speaker 5, every comparison is significant to at least the $p < 0.01$ level, whereas for Speaker 6 the comparison between levels 4 and 5 is not significant (see the table in Appendix A).

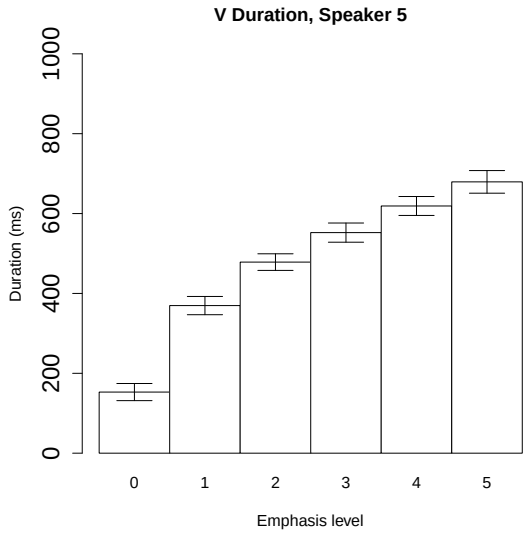


Figure 5: Speaker 5. $r = .73$, coeff. = 76 ms.

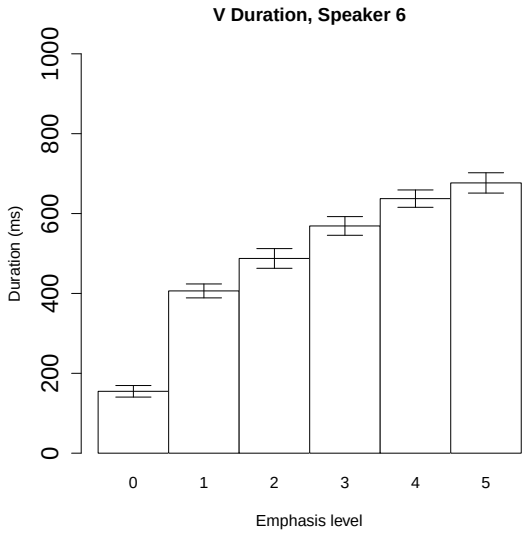


Figure 6: Speaker 6. $r = .72$, coeff. = 69 ms.

Turning now to Speaker 8, illustrated in Figure 7, we again see a steady increase in duration as emphasis

level increases. This speaker's r-value is .63 ($p < .001$), and her regression coefficient is 65 ms. Like the other speakers we have observed, and like the Japanese speakers tested in Kawahara & Braver (2013, 2014), Speaker 8 has a large jump in duration between the no emphasis and level 1 emphasis conditions. Like Speaker 6, Speaker 8 has a relatively small difference in duration between emphasis levels 4 and 5 (in Figure 7, the error bars for these two conditions overlap). This situation is again reflected in the non-paired t-tests (see the table in Appendix A): all comparisons are significant to at least the $p < .01$ level except for the comparison between levels 4 and 5.

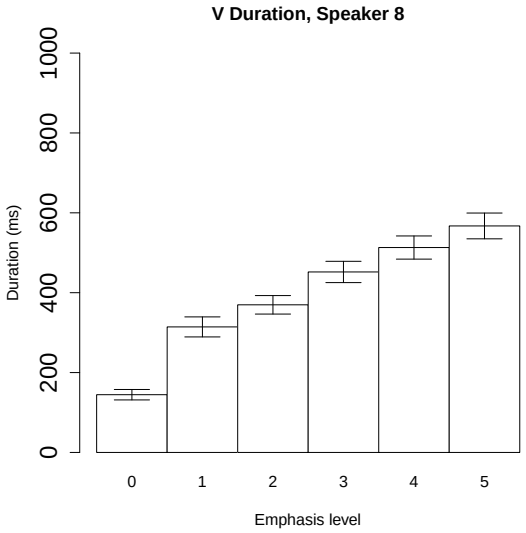


Figure 7: Speaker 8. $r = .63$, coeff. = 65 ms.

Speaker 3 and Speaker 2 have the next highest r-values, at .44 and .42 (both $p < .001$), respectively. The regression analysis indicates that for Speaker 3, each emphasis level results in an additional 60 ms of duration, and 52 ms for Speaker 2. Both of these speakers show relatively little change in duration between emphasis levels 2, 3, and 4 (the error bars overlap in these conditions for Speaker 3 in Figure 8 and for Speaker 2 in Figure 9). The non-paired t-tests included in Appendix A support this observation—for both speakers all comparisons are significant to at least $p < .01$ except for the comparisons between levels 2-3 and 3-4.

Finally, Speakers 4 and 7 both show a positive correlation between emphasis level and duration ($r = .38$ for Speaker 4 and $r = .35$ for Speaker 7; both $p < 0.001$). However, as can be seen in Figures 10 and 11 respectively, the duration differences between each level of emphasis are quite small—error bars overlap between almost every pair of conditions. This is reflected in the relatively small regression coefficients for these speakers: 22 ms and 26 ms, respectively. The biggest duration difference for both speakers is between the no emphasis condition and level 1 emphasis—they clearly differentiate no emphasis from some emphasis, but more fine-grained degrees of emphasis are not clearly reflected in the duration measurements. The non-paired t-tests in Appendix A show a similar pattern: Speaker 4 shows a significant difference between emphasis levels 0 and 1 ($p < .001$), but no other comparisons show significance. For Speaker 7, the comparisons between levels 0 vs. 1, and 1 vs. 2 are significant ($p < .001$ and $p < .01$, respectively), but no other comparisons are significant.

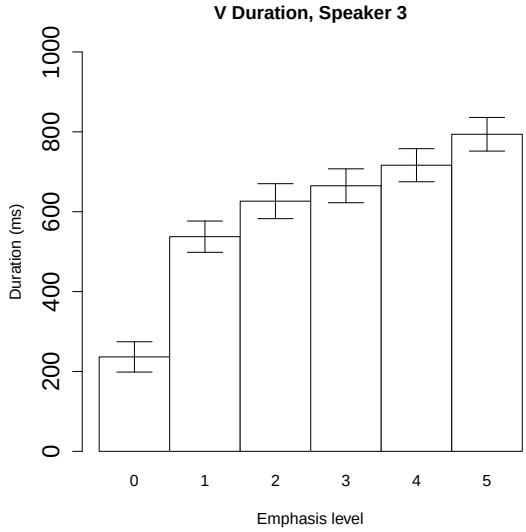


Figure 8: Speaker 3. $r = .44$, coeff. = 60 ms.

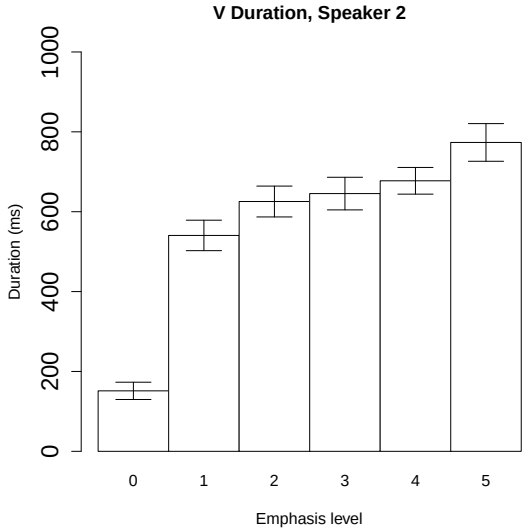


Figure 9: Speaker 2. $r = .42$, coeff. = 52 ms.

2.2.2 Summary

Table 2 gives a summary of each speaker's data. For each speaker, we provide an r-value, regression coefficient, and as a measure of speakers' duration range, the maximum token duration. See Appendix A for a table providing the results of t-tests on comparisons between each pair of consecutive emphasis levels for each speaker.

Speaker	r	Coeff. (ms)	Max dur. (ms)
1	0.75	92	1265
5	0.73	76	1020
6	0.72	69	962
8	0.63	65	876
3	0.44	60	1347
2	0.42	52	1427
4	0.38	22	702
7	0.35	26	803

Table 2: Speakers' r-values, regression coefficients, and maximum vowel/rhyme duration.

The regression analysis shows that all speakers have a significant ($p < .001$) positive correlation between duration and emphasis levels 1–5. Inter-speaker variability is also evident: correlations range from $r = 0.75$ (Speaker 1) to $r = 0.35$ (Speaker 7). All speakers show a large jump in duration from the no emphasis condition to the level 1 emphasis condition, with smaller duration gaps between other emphasis levels. This result points to an interesting cross-linguistic parallel between English and Japanese, as all the speakers tested in Kawahara & Braver (2013, 2014) also showed a bigger jump between no emphasis condition and the level 1 emphasis condition than anywhere else.

Speakers 1, 5, and 6 show a clear six-way durational contrast without much overlap in error bars (see Figures 4, 5, and 6). While other speakers do not show these distinctions quite as clearly, they do show a mostly steady linear increase in duration as emphasis level increases. Overall there are no evident significant

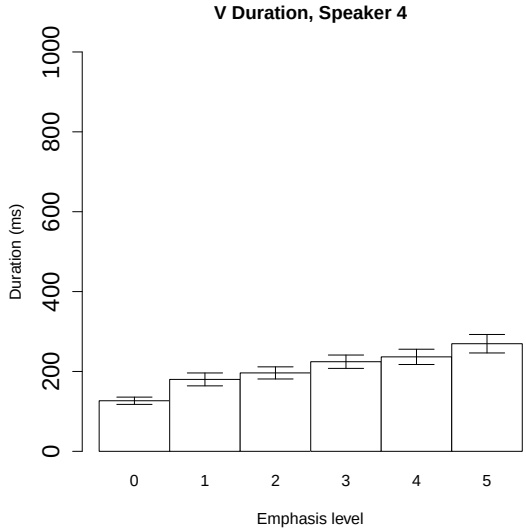


Figure 10: Speaker 4. $r = .38$, coeff. = 22 ms.

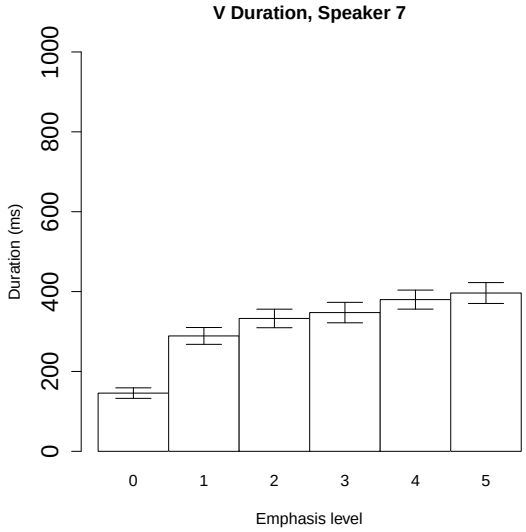


Figure 11: Speaker 7. $= .35$, coeff. = 26 ms

reversals—i.e., no speaker produces shorter durations as emphasis level increases. Speaker 7, and to a lesser extent Speaker 4, show an (almost) binary contrast between the no emphasis condition on the one hand, and all other emphasis conditions on the other.

Table 2 shows that there is a weak correlation between r-value and overall maximum duration: Speakers 1 and 5 have high maximum durations and high r-values, whereas Speakers 4 and 7 have low r-values and maximum durations. This association is not perfect—Speakers 3 and 2 have the highest maximum durations, but relatively low r-values.

2.3 Discussion

The current experiment provides a first experimental description of the emphatic vowel lengthening process in English. In spite of some inter-speaker variability, several speakers produce a six-way durational distinction. Among those speakers who fail to produce such a fine-grained distinction, all except Speakers 4 and 7 produce distinctions more fine-grained than simply binary emphasis/no emphasis. A general implication of this study is that at least some English speakers possess the articulatory control to produce duration distinctions more fine-grained than just short/long.

2.3.1 Effect of a native binary vowel length contrast

A question posed in the introduction is whether native language plays a role in the ability to produce fine-grained duration distinctions. Evidence from emphatic lengthening in Japanese suggests that at least some Japanese speakers are able to make 6-way duration contrasts (Kawahara & Braver 2013, 2014). These speakers have a potential advantage over speakers of English, since Japanese has a lexical short/long contrast. English does not have a contrastive length distinction, yet the speakers in the experiment presented here perform similarly to speakers of Japanese—they are able to produce fine-grained durational distinctions. We conclude that a native binary vowel length contrast, therefore, is not necessary for the production of ternary or greater length distinctions.[3]

[3]Though we admit that this claim should ideally be tested in languages beyond English. Further, we note that tense vowels in English are longer than lax vowels. This distinction is also cued by vowel quality, and indeed vowel duration is not a major cue in tense/lax vowel identification by English listeners (Hillenbrand, Clark & Houde 2000).

2.3.2 Against the "counting hypothesis"

A possible alternative analysis of our results suggests that speakers do not show a linguistic ability to produce fine-grained duration distinctions *per se*, but rather are simply counting the number of vowels in the orthographic representation they were shown. For example, a speaker could count that soooo has four (orthographic) vowels, and could then count four beats during production. We believe that this is not the strategy actually employed by the speakers in this study. All speakers in this experiment (and indeed, all speakers in both Kawahara & Braver 2013, 2014) show a large duration jump from the non-emphatic condition to level 1 emphasis—almost universally a larger duration gap than between any two other conditions. If speakers were simply counting, we should expect a uniformly linear correlation between duration and emphasis level. In fact, we see that speakers treat the distinction between level 0 emphasis and level 1 emphasis differently than they treat other distinctions.

English speakers, like Japanese speakers, overall make a binary distinction between non-emphatic and emphatic vowels, and then within the category of emphatic vowels, speakers vary their productions to express further degrees of emphasis. This categorical difference between emphasis levels 0 and 1 should therefore be expected to be qualitatively different from the more fine-grained distinctions between levels 1 and 5 which do not cross a category boundary.

2.3.3 Effects on consonants

Although the emphasis described here is expressed on vowels on orthography, it is conceivable that we would observe effects of emphasis on preceding consonants as well, especially since Kawahara & Braver (2014) found that emphasis on consonants sometimes manifested on the preceding vowel. To examine this possibility, we conducted a post-hoc examination of the onset consonants for tokens of "mad" and "too" (chosen because of the relative ease of segmenting their onsets from preceding and following material).

Speaker	r (cons)		r (vowel)	comment
Speaker 1	0.08	ns	0.75	localized to vowels
Speaker 2	0.26	$p < 0.05$	0.42	distributed
Speaker 3	0.11	ns	0.44	localized to vowels
Speaker 4	-0.03	ns	0.38	localized to vowels
Speaker 5	0.2	ns	0.73	localized to vowels
Speaker 6	0.22	$p < 0.05$	0.72	distributed
Speaker 7	0.38	$p < 0.001$	0.35	distributed
Speaker 8	0.35	$p < 0.001$	0.63	distributed

Table 3: Correlation between emphasis level and consonant duration (column 2) and emphasis level and vowel duration (column 3).

Table 3 summarizes the degrees of lengthening on consonants and vowels for each listener. Interestingly, some speakers appear to localize their emphasis to vowels, leaving consonants relatively unaffected (Speakers 1, 3, 4, and 5). Other speakers, like 2, 6, 7, and 8 seem to show a more distributed pattern of lengthening, with positive correlations to emphasis level for both consonant and vowel length. There does not appear to be a correlation in which speakers with high correlations for vowels show low correlation for consonants and vice versa. This is exemplified by the fact that speakers with similarly high r values for vowels may be either localizers (like Speaker 1, vowel $r = .75$) or distributers (like Speaker 6, vowel $r = .72$). For these distributers, we assume that the domain of lengthening is the whole word, rather than just the vowel (or rime), suggesting that this phenomenon is not solely and directly tied to the orthographic representation

of the emphasized words, which shows lengthening only on the vowels.

2.3.4 Item effects

The pattern of vowel lengthening was similar across all items, as seen in Figure 12.

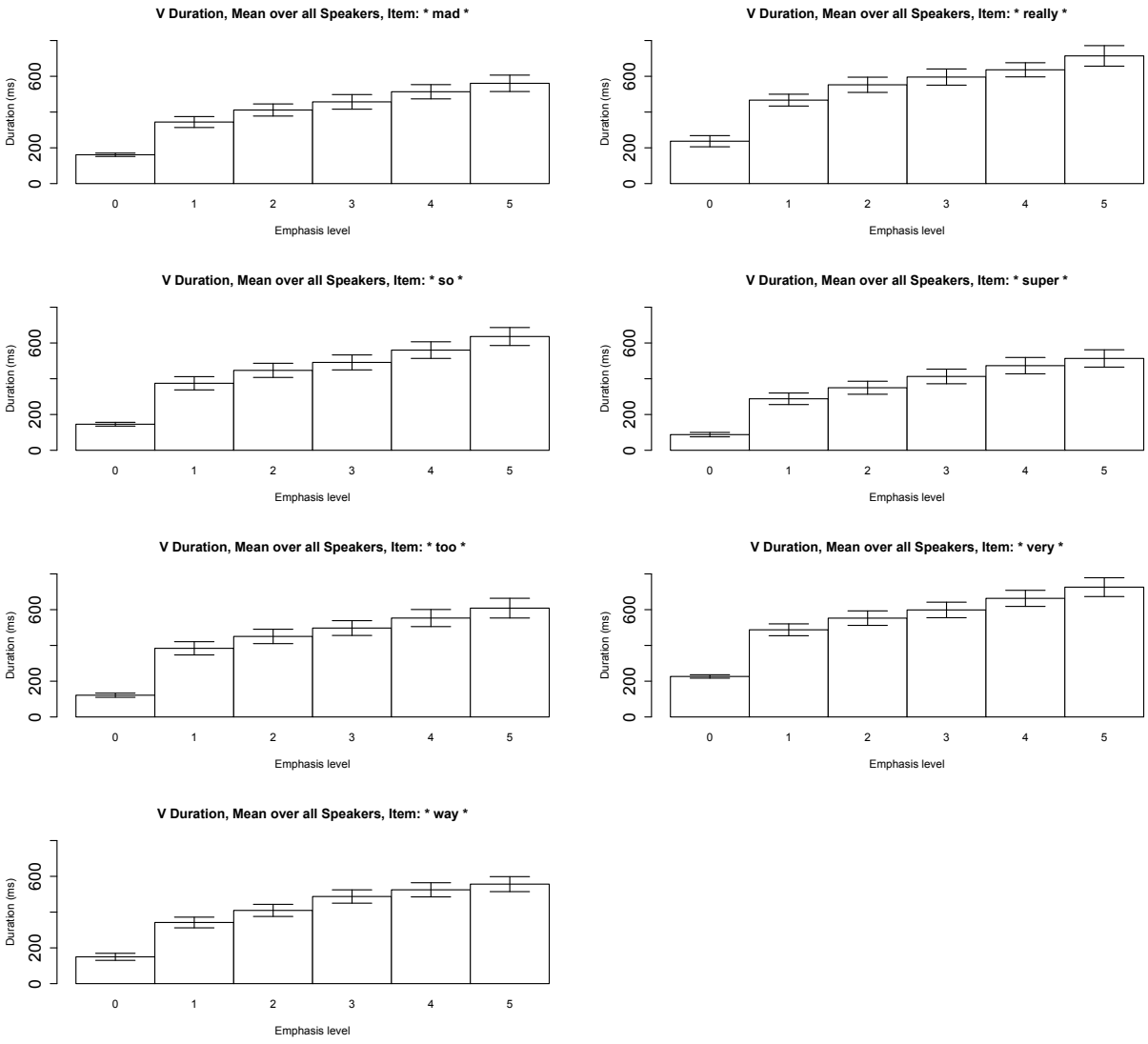


Figure 12: Vowel durations by emphasis level for each of the 5 items.

3 Experiment 2: Perception study

While it is clear that at least some speakers of English, like speakers of Japanese, can produce fine-grained durational distinctions of up to 6 levels, it is still unclear whether they are able to perceive these distinctions. Indeed, we are aware of no studies that directly examine the perceptual properties of emphatic lengthening in English or any other language. The perception study presented here addresses these unanswered question.

3.1 Method

3.1.1 Stimuli

Stimuli were selected from the set of items in the production experiment whose emphasis levels were most clearly distinguished, which include Speaker 5's super, really, so, Speaker 6's really, so, too and Speaker 8's mad, so, super. The current experiment included 10 repetitions for each item. The total number of stimulus items was thus 540: 3 speakers $\times$ 3 items $\times$ 6 levels of emphasis $\times$ 10 repetitions.

3.1.2 Procedure

The participants were 24 native speakers of English, recruited from a pool of undergraduate students. They were first told that English makes use of lengthening to express emphasis and were given an example. They listened to each stimulus and judged which "emphasis level" they heard. The trials were blocked by item and speaker (e.g. *super* from Speaker 5), to maximize the easiness of the task. Superlab was used to present the stimuli and feedback (Cedrus Corporation 2010). The order of the stimuli was randomized. All participants wore high quality headphones (Sennheiser HD 280 Pro), and registered their responses using a Cedrus RB-730 response box. The experiment took place in a sound-attenuated laboratory.

3.2 Overall perception results

Perfect hits—trials on which the participant's response matched exactly the stimulus emphasis level—consisted of 29.5% of trials over all speakers, which is significantly greater than chance (1/6 possible responses, $16.\overline{6}\%$; $t(23) = 25.12, p < 0.001$). A more fine-grained look at the results of the perception study are provided in the confusion matrix in Table 4, where the emphasis level of stimuli is labeled across the top and participants' response levels along the side; for example, upon hearing a stimulus of level 2, 35.11% of responses were of emphasis level 3 (a mismatch). Bold numbers indicate the most common response for each stimulus emphasis level (e.g., for stimulus emphasis level 1, the most common response was level 3, with 33.19% of responses). The shaded cells along the diagonal represent an exact match between stimulus and response emphasis levels ('perfect hits')—if participants mostly gave correct responses, we would expect the most common response for each stimulus emphasis level to fall along this diagonal.

		Stimulus emphasis level					
		0	1	2	3	4	5
Response emphasis level	0	**42.75**	1.39	0.14	0.14	0.25	0.24
	1	35.69	10.06	5.80	2.92	1.18	1.21
	2	12.21	28.50	20.98	11.17	6.81	4.41
	3	5.44	**33.19**	**35.11**	32.93	26.57	21.37
	4	2.94	19.97	26.10	**34.42**	**36.62**	**38.23**
	5	0.98	6.85	11.83	18.41	28.38	34.16

Table 4: Confusion matrix for perception experiment: percent of response emphasis level per stimulus emphasis level. Bold numbers indicate the most common response per stimulus emphasis level. Shaded cells along the diagonal indicate matched stimulus/response emphasis levels.

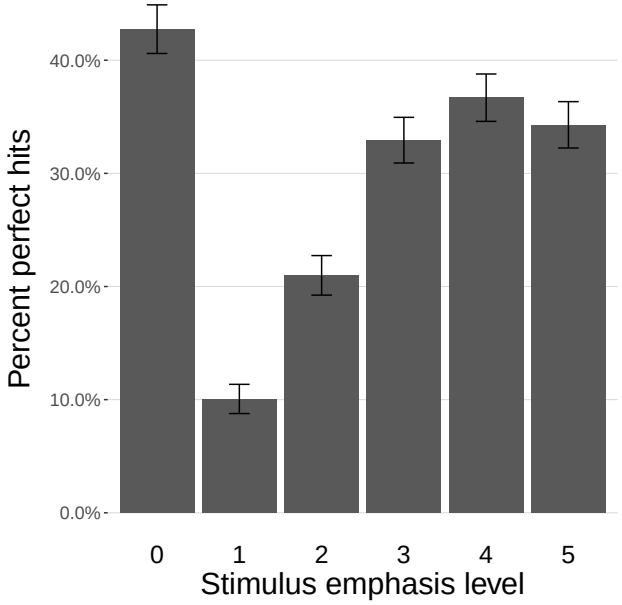


Figure 13: Percent of trials on which participants' responses matched exactly the stimulus emphasis level, by stimulus emphasis level.

As can be seen in the confusion matrix, the most common response for a given stimulus level matched exactly only for levels 0 and 4; the most common response for all other stimulus levels was something other than the matching level. This suggests a relatively poor ability of participants to distinguish accurately between levels for emphasis levels other than 0 and 4. This is reflected as well in Figure 13. which shows the percent of perfect hits for each stimulus emphasis level. It is clear that participants were correct on emphasis level 0 more than any other level, suggesting a strong ability to discriminate at least between 'no emphasis' and 'at least some emphasis'.

A more charitable interpretation of these results is that participants broadly categorized stimuli into three categories, based on the most frequent response for each emphasis level: not emphasized (stimulus emphasis level 0, categorized by respondents mostly as level 0), emphasized (stimulus emphasis levels 1–2, with responses centering around level 3), and super-emphasized (stimulus emphasis levels 3–5, with responses centering around level 4). Participants' ability to categorize emphasis into three levels more closely accords with the sorts of lexical vowel duration contrasts we see in natural language—such contrasts are usually binary, but there are reports of systems with three levels (e.g., Estonian, Prince 1980).

3.3 Perception results by participant

Participants' perfect hit accuracy ranged from a low of 17.0% (participant cm) to a high of 39.3% (participants jn and xm). The distribution of participants' perfect hit accuracy scores is shown in the histogram in the left panel of Figure 14, with the majority of participants' accuracy scores above the chance level of $16.\overline{6}\%$.

If a linguistic contrast is to be robust, however, we should perhaps expect an accuracy rate nearing ceiling, rather than simply beating chance. To consider whether a contrast with fewer levels might meet this higher standard, we consider here 'near hits'—trials on which participants' responses were either perfect matches or within 1 level of the stimulus emphasis level. On this metric, participants ranged from a low of 57.3% (participant sp) to a high of 82.8% (participants jn and bp) with a mean of 70.4%. This measure is shown in the right panel of Figure 14.

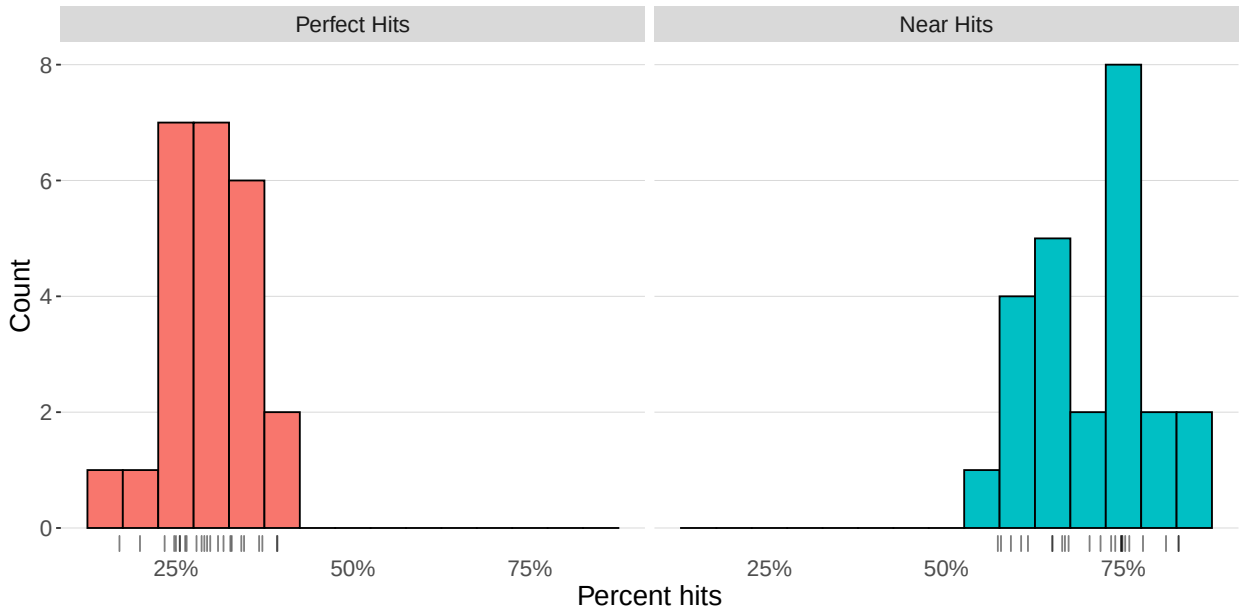


Figure 14: Histogram of participants' accuracy in the perception task. Perfect hits (left panel) include trials on which participants responded with exactly the level of the stimulus. Near hits (right panel) include responses off by 1 emphasis level.

It is also worthwhile to examine the distribution of responses within each participant in order to further consider how participants perfect once we take near hits into consideration. Figure 15 shows, for each speaker, the proportion of trials on which their response was correct (perfect hits), over or under by 1 (near hits), or over or under by 2 or more ('misses'), ordered from least to most perfect hits.

As discussed previously in this section, participants' perfect hit rates ranged from a low of 17% to a high of 39%. Comparing these extremes to chance (16.$\overline{6}$%), we have some participants (like cm and rz) who are performing right about at chance, while others (speakers qw through xm) who are performing better than two times chance.

We note that all participants except xm and rr overestimated emphasis levels rather than underestimating them (in Figure 15, the red portions of each bar are greater than the blue portions). We speculate this tendency may be related to the design of the experiment: participants were asked to identify the 'emphasis level' of each trial, and may have tended to assume that all stimuli were emphasized.

4 Discussion and conclusions

The two experiments presented in this paper examine speakers' ability to produce and perceive fine-grained duration distinctions in emphatic contexts. While previous work has shown that speakers of a language with lexical duration contrasts can produce these distinctions (Kawahara & Braver 2013, 2014), the current production study showed that this ability extends to speakers of English—a language which lacks lexical duration contrasts, though there is a significant degree of inter-speaker variation. This suggests that being a speaker of a language with a binary duration contrast is not a requirement for being able to produce even more fine-grained distinctions.

The perception study tested English speakers' ability to perceive similarly fine-grained distinctions. The results suggest that such speakers are not able to reliably distinguish between the 6 levels of emphasis/no emphasis presented to them. Speakers did seem, however, to group stimuli into three broad categories: not emphasized, emphasized, and super-emphasized. This ability stands in contrast to the fact that English does

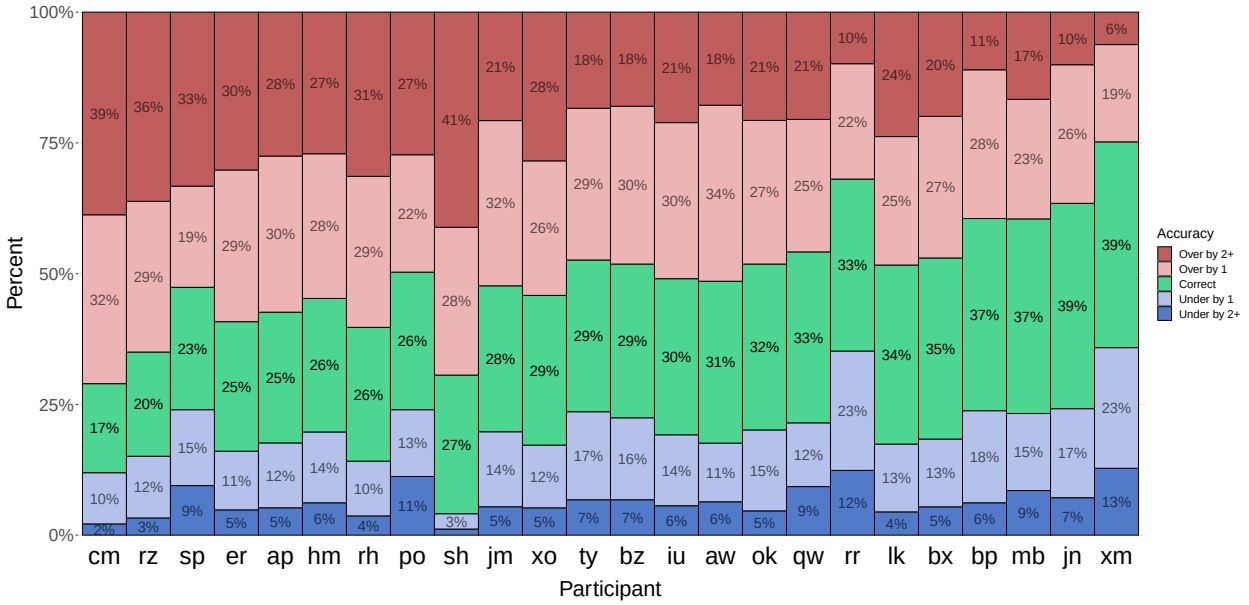


Figure 15: Participants' accuracy in the perception task. Each participants' percent of correct responses are shown in green; responses too low/too high by 1 are shown in pale blue/pale red respectively; responses too low/too high by 2 or more are shown in dark blue/red. Participants are ordered from least correct responses (cm) to most correct responses (jn and xm).

not maintain a lexical duration contrast.

In light of the results of this experiment, a question that arises is why natural languages are generally restricted to two-way length contrasts. We noted the possibility in the introduction that binary contrasts are preferred due to restrictions on speakers' ability to produce or perceive more fine-grained distinctions. Given that both English speakers (in this experiment), and Japanese speakers (in Kawahara & Braver 2013, 2014) can produce up to six-way duration distinctions, we argue that the restriction does not lie in production. Further, speakers are generally unable to perceive the distinction between six levels of emphasis, though they may more reliably categorize such stimuli into a three-way contrast.

We suggest that three-way or greater durational contrasts are difficult to perceive without ambiguity, much as vowel spaces are constrained by requirements to be sufficiently dispersed for ease of perception. (see, e.g., Flemming 2001; Liljencrants & Lindblom 1972; Lindblom 1986; and especially Engstrand & Krull 1994; Podesva 2000; Kawahara & Pangilinan 2017 for the grammatical imperatives on perceptual dispersion in durational contrasts). Languages, therefore, prefer binary (or maximally ternary) duration contrasts since they are more easily perceived than more fine-grained duration distinctions.

Sources, Acknowledgements, Abbreviations

5 Appendix A: Comparison of consecutive emphasis levels by speaker

Speaker	Comparison	mean diff. (ms.)	t(df)	p
	level 0 vs. level 1	285.27	$t(87.32) = 21.32$	$p < .001$
	level 1 vs. level 2	84.17	$t(107.96) = 5.09$	$p < .001$
1	level 2 vs. level 3	84.53	$t(86.69) = 4.08$	$p < .001$
	level 3 vs. level 4	94.46	$t(96.22) = 4.23$	$p < .001$

	level 4 vs. level 5	105.35	$t(91.80) = 3.90$	$p < .001$
2	level 0 vs. level 1	389.16	$t(96.52) = 17.75$	$p < .001$
	level 1 vs. level 2	84.92	$t(123.00) = 3.13$	$p < .01$
	level 2 vs. level 3	19.88	$t(123.63) = 0.71$	$n.s.$
	level 3 vs. level 4	32.00	$t(117.45) = 1.21$	$n.s.$
	level 4 vs. level 5	96.04	$t(110.51) = 3.32$	$p < .01$
3	level 0 vs. level 1	301.08	$t(135.96) = 11.02$	$p < .001$
	level 1 vs. level 2	88.91	$t(135.13) = 3.02$	$p < .01$
	level 2 vs. level 3	38.54	$t(136.80) = 1.26$	$n.s.$
	level 3 vs. level 4	51.45	$t(137.92) = 1.73$	$n.s.$
	level 4 vs. level 5	77.55	$t(137.96) = 2.62$	$p < .01$
4	level 0 vs. level 1	53.44	$t(107.48) = 5.75$	$p < .001$
	level 1 vs. level 2	16.26	$t(136.36) = 1.46$	$n.s.$
	level 2 vs. level 3	28.07	$t(134.37) = 2.48$	$n.s.$
	level 3 vs. level 4	12.13	$t(126.63) = 0.95$	$n.s.$
	level 4 vs. level 5	32.93	$t(128.08) = 2.19$	$n.s.$
5	level 0 vs. level 1	216.56	$t(133.69) = 13.79$	$p < .001$
	level 1 vs. level 2	108.88	$t(135.62) = 7.03$	$p < .001$
	level 2 vs. level 3	73.84	$t(135.21) = 4.63$	$p < .001$
	level 3 vs. level 4	66.69	$t(137.97) = 3.94$	$p < .001$
	level 4 vs. level 5	60.34	$t(132.56) = 3.26$	$p < .01$
6	level 0 vs. level 1	251.38	$t(133.48) = 22.13$	$p < .001$
	level 1 vs. level 2	81.30	$t(124.15) = 5.37$	$p < .001$
	level 2 vs. level 3	81.37	$t(137.60) = 4.77$	$p < .001$
	level 3 vs. level 4	68.42	$t(137.26) = 4.28$	$p < .001$
	level 4 vs. level 5	39.34	$t(134.72) = 2.35$	$n.s.$
7	level 0 vs. level 1	143.19	$t(115.54) = 11.5$	$p < .001$
	level 1 vs. level 2	43.76	$t(136.67) = 2.78$	$p < .01$
	level 2 vs. level 3	14.70	$t(136.75) = 0.85$	$n.s.$
	level 3 vs. level 4	32.47	$t(137.32) = 1.85$	$n.s.$
	level 4 vs. level 5	16.49	$t(136.85) = 0.93$	$n.s.$
8	level 0 vs. level 1	169.74	$t(103.80) = 11.98$	$p < .001$
	level 1 vs. level 2	55.32	$t(137.14) = 3.23$	$p < .01$
	level 2 vs. level 3	82.30	$t(135.34) = 4.65$	$p < .001$
	level 3 vs. level 4	61.13	$t(137.03) = 3.10$	$p < .01$
	level 4 vs. level 5	54.10	$t(136.46) = 2.49$	$n.s.$

Non-paired t-tests between each emphasis level for all speakers, showing the effect of emphasis level on duration. $\alpha = .01$ after Bonferroni adjustment (each speaker has five comparisons; 0.05/5).

References

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10). 341–345.
Cedrus Corporation. 2010. Superlab v. 4.0. Software.

Chen, Matthew. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22. 129–159. doi:10.1159/000259312.

Engstrand, Olle & Diana Krull. 1994. Durational correlates of quantity in Swedish, Finnish and Estonian: Cross-language evidence for a theory of Adaptive Dispersion. *Phonetica* 51. 80–91. doi:10.1159/000261960.

Flemming, Edward. 2001. *Auditory representations in phonology*. New York: Garland Press.

Fuchs, Susanne, Egor Savin, Stephanie Solt, Cornelia Ebert & Manfred Krifka. 2019. Antonym adjective pairs and prosodic iconicity: Evidence from letter replications in an English blogger corpus. *Linguistics Vanguard* 5(1). doi:10.1515/lingvan-2018-0017.

Higuchi, Marii & Shosuke Haraguchi. 2006. Final lengthening in Japanese. *On-in Kenkyu [Phonological Studies]* 9. 9–16.

Hillenbrand, J ames M., Michael J. Clark & Robert A. Houde. 2000. Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America* 108(6). 3013–3022. doi:10.1121/1.1323463.

Hoogshagen, Searle. 1959. Three contrastive vowel lengths in Mixe. *Zeitschrift für Phonetik und allgemeine Sprachwissenschafe* 12. 111–115. doi:10.1524/stuf.1959.12.14.111.

Jany, Carmen. 2007. Phonemic versus phonetic correlates of vowel length in Chuxnabán Mixe. In *Proceedings of Berkeley Linguistics Society 33s: Languages of mexico and central america*, Berkeley: Berkeley Linguistics Society. doi:10.3765/bls.v33i2.3502.

Jesperson, Otto. 1922. *Language: Its nature, development and origin*. New York: Norton & Co.

Kawahara, Shigeto. 2013. Emphatic gemination in Japanese mimetic words: A wug-test with auditory stimuli. *Language Sciences* 40. 24–35. doi:10.1016/j.langsci.2013.02.002.

Kawahara, Shigeto. 2015. The phonetics of sokuon, or obstruent geminates. In Haruo Kubozono (ed.), *The handbook of Japanese language and linguistics: Phonetics and phonology*, 43–73. Berlin: Mouton.

Kawahara, Shigeto & Aaron Braver. 2013. The phonetics of emphatic vowel lengthening in Japanese. *Open Journal of Modern Linguistics* 3(2). 141–148. doi:10.4236/ojml.2013.32019.

Kawahara, Shigeto & Aaron Braver. 2014. Durational properties of emphatically-lengthened consonants in Japanese. *Journal of International Phonetic Association* 44(3). 237–260. doi:10.1017/S0025100314000085.

Kawahara, Shigeto & Melanie Pangilinan. 2017. Spectral continuity, amplitude changes, and perception of length contrasts. In Haruo Kubozono (ed.), *Aspects of geminate consonants*, 13–33. Oxford: Oxford University Press. doi:10.1093/oso/9780198754930.003.0002.

Ladefoged, Peter & Ian Maddieson. 1996. *The sounds of the world's languages: 2nd edition*. Oxford: Blackwell Publishers.

Lakoff, Robin. 1973. Language and woman's place. *Language in society* 2(1). 45–79. doi:10.1017/S0047404500000051.

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge: MIT Press.

Liljencrants, Johan & Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48. 839–862. doi:10.2307/411991.

Lindblom, Björn. 1986. Phonetic universals in vowel systems. In John Ohala & Jeri Jaeger (eds.), *Experimental phonology*, 13–44. Orlando: Academic Press.

Nasu, Akio. 1999. Chouhukukei onomatope no kyouchou keitai to yuuhyousei [Emphatic forms of reduplicative mimetics and markedness]. *Nihongo/Nihon Bunka Kenkyuu [Japan/Japanese Culture]* 9. 13–25.

Podesva, Robert. 2000. Constraints on geminates in Burmese and Selayarese. In Roger Bilerey-Mosier & Brook Danielle Lillehaugen (eds.), *Proceedings of West Coast Conference on Formal Linguistics 19*, 343–356. Somerville: Cascadilla Press.

Prince, Alan. 1980. A metrical theory for Estonian quantity. *Linguistic Inquiry* 11. 511–562.

Samejon, Kevin. 2019. Emphatic vowel length in cebuano. *Philippine Journal of Linguistics* 50. 18–32.

Tagliamonte, Sali & Chris Roberts. 2005. So weird; so cool; so innovative: The use of intensifiers in the

television series Friends. *American speech* 80(3). 280–300. doi:10.1215/00031283-80-3-280.

Thomas, Kimberly D. & Alan Shaterian. 1990. Vowel length and pitch in Yavapai. In M.D. Langdon (ed.), *Papers from the 1990 hokan-penutian languages workshop*, 144–153. Department of Linguistics, University of Southern Illinois. doi:https://doi.org/10.1121/1.2029048.

Turk, Alice, Satsuki Nakai & Mariko Sugahara. 2006. Acoustic segment durations in prosodic research: A practical guide. In Stefan Sudhoff, Denisa Lenertova, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schliesser (eds.), *Methods in empirical prosody research*, 1–27. Berlin: Walter de Gruyter.

Xiao, Richard & Hongyin Tao. 2007. A corpus-based sociolinguistic study of amplifiers in british english. *Sociolinguistic studies* 1(2). 241–273. doi:10.1558/sols.v1i2.241.