VOWEL DURATION AND SPECTRA AS PERCEPTUAL CUES TO VOWEL QUANTITY: A COMPARISON OF JAPANESE AND SWEDISH

Dawn Behne[*], Takayuki Arai[†], Peter Czigler[‡], and Kirk Sullivan[‡]
[*]*Norwegian University of Science and Technology, Trondheim, Norway,*
[†]*Sophia University, Tokyo, Japan, and* [‡]*Umeå University,Umeå, Sweden*

ABSTRACT

A distinction in vowel quantity is typically realized acoustically by vowel duration. Research on the perception of Swedish vowel quantity supports this and further suggests that when the duration of a vowel is relatively long (due, e.g., to inherent duration), vowel quantity may not be adequately cued by duration alone and may also make use of the vowel spectra to distinguish vowel quantities. If this account of the Swedish findings is correct, other languages which have vowel quantity distinctions would be expected to show a similar pattern.

The current project investigates the perceptual cues used to distinguish vowels quantities in Japanese. Of particular interest is whether Japanese listeners use spectral cues to identify the quantity of vowels which have a relatively long inherent duration. Results are compared with the findings for Swedish and discussed in terms of the perceptual role of vowel duration and spectra as cues for vowel quantity.

1. INTRODUCTION

1.1 Background

The vowel systems of some languages are described as having contrastive vowel quantities. Vowel quantity refers to the phonological distinction of a vowel relative to one or more other vowels of similar timbre in the language. Contrasts in vowel quantity are often acoustically realized by the duration of vowels, with a long vowel quantity having a duration which extends over more time than a short vowel quantity. The greater amount of time associated with a long vowel quantity also allows the possibility for a more extreme articulation than a corresponding short vowel quantity. Consequently, the vowel spectrum, in particular the first and second formant frequencies, and perceived timbre [1] may also be affected by vowel quantity.

One language traditionally characterized as having distinctions between long and short vowel quantities is Swedish [2]. In a classic study Hadding-Koch and Abramson [3] investigatigated whether vowel duration or spectral attributes of a vowel have a more dominant perceptual role in distinguishing vowel quantities in Swedish, and concluded that although vowel duration was a primary perceptual cue to Swedish vowel quantity, the role of the vowel spectra could not be excluded.

Recent studies [4, 5] have reexamined the effects of vowel duration and the first two formant frequencies on perceived vowel quantity identification in Swedish. Based on natural Swedish productions, three sets of 100 /kVC/ words were synthesized, one set for each of three "long-short" vowel pairs. Within each set 10 stepwise adjustments of vowel duration and 10 stepwise adjustments of F1 and F2 were made from the long vowel quantity to the short one, resulting in a total of 100 synthesized words for each of the three vowel quantity pairs. This was done for /kVt/ [4] and /kVd/ [5] words. Subjects´ identification responses for the /kVd/ words are presented again here in Figure 1. The results for /kVt/ words illustrated that Swedish listeners use vowel duration more than spectral information to identify the quantity of the vowel pairs /i:/-/ɪ/ and /o:/-/ɔ/, whereas for the vowel pair /ɑ:/-/a/, listeners use both vowel duration and spectral attributes of the vowel. In the case of the /kVd/ items, listeners were even more likely to use both vowel duration and spectra to identify vowel quantity.

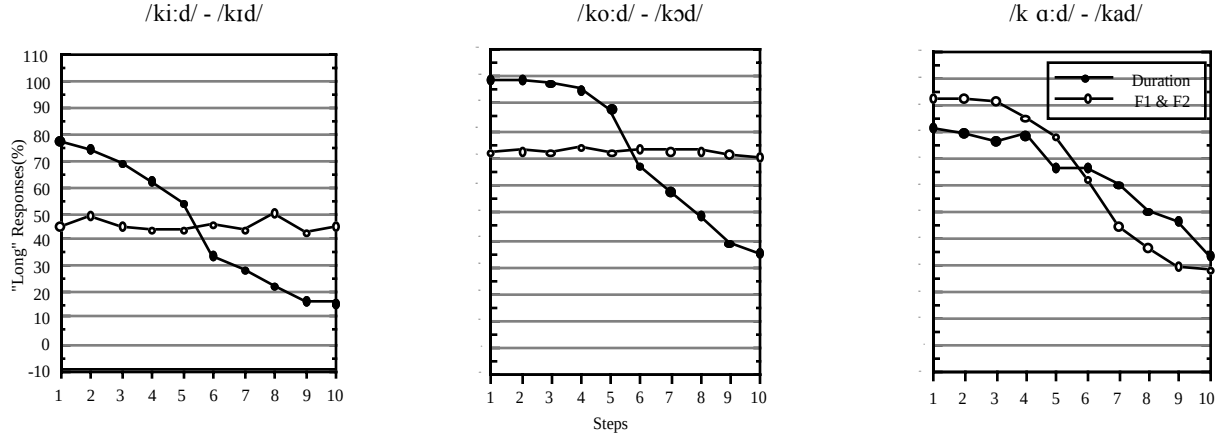

Figure 1: Percent "long" responses from [5] are plotted for 10 duration (-•-) steps and 10 spectral (-o-) steps. Step 1 corresponds to the "long" vowel quantity and step 10 corresponds to the "short" vowel quantity for /ki:d/-/kɪd/, /ko:d/-/kɔd/, and /kɑ:d/-/kad/.

1.2. Hypothesis

The production of the vowels /ɑ:/ and /a/ involves moving the relatively heavy jaw, resulting in a longer inherent vowel duration than the other vowels studied [6]. Postvocalic voicing is also known to cause the duration of a preceding vowel to be longer in Swedish and other languages (e.g., [7]). As a possible explanation for our results, we considered that it may simply be difficult to lengthen relatively long vowels indefinitely (see also, [8]) without disturbing the natural rhythm of speech, and hypothesized that when the duration of a vowel is relatively long (due, for example to inherent duration or postvocalic voicing)-- pushing what might be the extent of what is vaguely permissible for a vowel as a rhythmic unit within a word--vowel quantity may not be adequately cued by vowel duration alone and may consequently make use of the vowel spectra too. If so, listeners would in those cases be more dependent on secondary acoustic cues, such as the vowel spectrum. If this motivation for the Swedish findings is well founded, we would expect to find the same general pattern of results in a language such as Japanese, which distinguishes vowel quantities but is also unrelated to Swedish.

1.3. Current study

As an initial investigation of this hypothesis, the current study was carried out to examine how Japanese listeners use vowel duration and the first two vowel formant frequencies when identifying Japanese vowel quantity, and to compare it with the Swedish listeners´ responses from the earlier studies [4, 5]. Of particular interest is whether Japanese listeners are more likely to use spectral cues to identify the quantity of the inherently long vowels, /a:/ and /a/, compared to other vowel pairs.

2. METHOD

The recordings, measurements, synthesis and experimental procedures used here are closely matched with those in [4, 5]. For consistency, the measurements and synthesis (sections 2.12. and 2.13.) were carried out at Umeå University where the Swedish materials had been prepared. The experiment running program used for the identification task (section 2.2) was based on the program originally used for the Swedish study.

2.1. Materials

2.1.1. Recordings. A set of six /kV/ real words containing the vowels /i, i:, o:, o, a:, a/ were used as targets. Since the vowel in /ki/ is commonly devoiced, the speaker produced voiced and devoiced variants.

A adult native phonetically-trained male speaker of the Tokyo dialect of Japanese was recorded producing 10 random repetitions of the six target words in the sentence “Mou ichido ___ to iu tango-wo itte kudasai.” (“Please say the word ___ again”). The productions were made at his natural speaking rate and using a natural intonation. Care was taken to avoid a falling pitch with long vowel quantities.

2.1.2. Measurements. From the 10 productions of each target word, ESPS/*waves*+™ was used to measure the vowel duration and the first three formant frequencies of the vowel, measured at the center of the vowel’s most evident steady state.

For each repetition of the six target words, the mean value of these measures was calculated and the production which best corresponded to the mean values was chosen to be used as the basis for resynthesis. These most representative items will be referred to as "selected productions".

Among the productions of /ki/ which were expected to have a devoiced vowel, voicing was obversed in two of the 10 cases. The vowel duration and formant frequencies of these 2 items were comparable to the 10 productions of /ki/ with the voiced vowel. Consequently, representative items were selected from among variants of /ki/ with the voiced vowel for resynthesis.

2.1.3. Synthesis. Using the Kay Elemetrics LPC Parameter Manipulation/Synthesis program, the selected productions of the six target words and their measured values were the basis for resynthesizing three sets of 100 words. Each set was based on the measurements from the selected productions of a pair of long-short vowel quantities: /i/-/i:/, /o/-/o:/, and /a/-/a:/.

For each set, the mesurements of the selected productions were used as extreme points of a 10x10 synthesis matrix, having ten degrees of vowel duration and ten degrees of simultaneous first and second formant frequency adjustment. Starting from the selected production of /ki:/, /ko:/ and /ka:/, the closure duration of the postvocalic consonant was adjusted in equal-sized steps toward the measured postvocalic closure duration of the selected productions of /ki/, ko/ and /ka/ respectively. In each series third formant frequency of the vowel was kept the same as it had been in the long vowel quantity.

Word pairs		Spectrum			Vowel
		F1 (Hz)	F2 (Hz)	F3 (Hz)	duration
/ki:/	Step 1	205	2424	3410	147
↓	Step size	2	-28	0	-10
/ki/	Step 10	223	2172	3410	54
/ko:/	Step 1	290	731	2609	190
↓	Step size	-1	28	0	14
/ko/	Step 10	280	980	2609	67
/ka:/	Step 1	308	1146	2494	199
↓	Step size	-5	12	0	-13
/ka/	Step 10	262	1258	2494	80

Table 1. Parameter settings of the vowel for the three sets of resynthesized materials.

2.2. Identification task

Twenty-four native speakers of the Tokyo dialect of Japanese participated in the study. The participants were an equal mix of young adult males and females and at the time of the experiment, were affiliated with Sophia University in Tokyo where the perception experiment was run.

Subjects were seated wearing headphones at a computer terminal with a monitor and mouse. For each trial, subjects heard a synthesized word and two /kV/ words were presented on the monitor in katakana. The two words on the monitor differed in vowel quantity and had the same vowel quality as the target words which the synthesized items in that series were based on.

Subjects were instructed to use the mouse to click on the visually presented word which they heard. They were asked to

respond as quickly as possible and were allowed up to 10 seconds to respond before the beginning of the next trial, although subjects rarely encountered this upper limit.

Subjects heard 5 randomized repetitions of each synthesized word, a total of 1500 items (3 vowel qualities x 100 items x 5 repetitions). Before starting the experiment, subects had three practice trials, and after each set of 50 trials, subjects had the opportunity to take a short break.

Subjects´ responses and reaction times for each trial were analyzed to determine the extent to which vowel duration and spectral characteristics lead to the perception of a phonologically long and short vowel.

3. RESULTS

For each word pair, the mean percent of responses for the word having a long vowel quantity were calculated and are referred to as “long" responses in the following discussion. In Figure 2 the percent long responses and reaction times for the 10 duration steps and spectral steps are presented for each of the three word pairs investigated.

Subjects´ responses clearly indicate that Japanese listeners find vowel duration to be a more dominant perceptual cue for vowel quantity than spectral information. This is the case for all three pairs of vowels, /i:/-/i/, /o:/-/o/ and /a:/-/a/, as can be seen from the the three sharp s-curves of the plotted responses across the 10 steps of duration steps. This is supported by the increased variability of subjects responses crossing the category boundary from short to long in the region of duration steps 4 through 7. Reaction times across vowel duration steps show the same pattern, with increased reaction times reflecting increased cognitive effort in the region of the category boundary and coming to a peak at duration step 7 in all cases. Notably, both the percent long responses and reaction times across the 10 steps of vowel duration show a a slight perceptual skew in favor of long responses. This is likely a result of preparing the resynthesised word pairs starting from the word containing a long vowel quantity.

Subjects´s responses and reaction times across the 10 spectral steps of the three word pairs show a different pattern than was observed across the duration steps. Responses based on simultaneous adjustments of F1 and F2 show no sign of categorization for any of the three word pairs, as can be seen from the flat curves across the spectral steps for each pair. The high variance of responses suggests that listeners had a difficult time responding consistently based on spectral information along. The similarity of listeners´ reaction times across the spectral steps further reflects that the effort used to identify a vowel´s quantity was unaffected by adjustments of F1 and F2.

.

4. DISCUSSION AND CONCLUSIONS

Listeners´ responses and reaction times across the 10 vowel and spectral steps demonstrate a clear pattern. For all three word pairs, the results show listeners using vowel duration, but not spectral information, to identify a vowel´s quantity.

These findings are consistent with general observations of how vowel quantity is acoustically realized across languages and earlier investigations of vowel duration as a perceptual cue to vowel quantity in Japanese (e.g., [9]). Based on results from Swedish [4, 5], it was also expected that vowel duration would offer a primary perceptual cue for the vowel quantity identification in Japanese, but had further hypothesized that listeners would tend to make use of spectral information in cases when the vowel was relatively long due to other linguistic factors. In particular, with the materials used in this study, we were interested in whether listeners would be more likely to use the vowel spectra to identify the quantity of the inherently longer low Japanese vowels, /a:/ and /a/. The results showed no indication of this pattern.

Notably, the materials used in the present study may be considered a strict context for observing the pattern earlier found for Swedish. In the studies on Swedish, /kVt/ and /kVd/ words were used. To match this phonetic environment as closely as possible within the (C)V(N) syllable structure of Japanese, /kV/ words were used immediately preceding a /t/ in the carrier sentence. The closed syllable used in the Swedish study would put greater restrictions on the possible vowel duration of the syllable than the open syllable used in the Japanese study; this may have resulted in a phonetic context where the vowel quantity could not be adequately cued by vowel duration alone in Swedish and that corresponding syllable-internal limitations on vowel duration were present in the Japanese study.

The Japanese results are, however, consistent with previous research, including the findings for Swedish. These findings suggest a complex, but systematic, role of vowel duration and spectra in distinguishing vowel quantity and set us in the direction of a follow-up study to further examine effects of syllable structure on the use of vowel duration and spectral information as perceptual cues to vowel quantity.

ACKNOWLEDGMENTS

The authors thank Yuko Mimura, Kenji Okada, Ola Andersson, and Thierry Deschamps for their programming and technical assistance, and Natasha Warner for discussion at early stages of the project. We also thank the many listeners from Umeå University and Sophia University for their patient participation in the studies described here.

REFERENCES

[1] Stevens K. and House A. 1955. Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America* 27, 484-493.

[2] Elert C-C. 1964. *Phonologic studies of quantity in Swedish*. Stockholm: Almqvist & Wiksell.

[3] Hadding-Koch K. and Abramson A. 1964. Duration versus spectrum in Swedish vowels: some perceptual experiments. *Studia Linguistica* 2, 94-107.

[4] Behne D., Czigler P. and Sullivan K. 1996. Acoustic characteristics of perceived quantity and quality in Swedish vowels. *Speech Science and Technology '96*, 6, Adelaide, 49-54.

[5] Behne, D. M., Czigler, P. E, and Sullivan, K. P. H. 1998. Perceived vowel quantity in Swedish: Effects of postvocalic voicing. *Proceedings of the 16th International Congress of Acoustics and the 135th Meeting of the Acoustical Society of America*; 2963-64, 1998.

[6] Peterson G and Lehiste I. 1960. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32, 693-703.

[7] House A., and Fairbanks G. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America* 25, 105-113.

[8] Behne, D., Moxness, B., and Czigler, P., 1995. Syllable and rhyme-internal timing: Evidence from English, Norwegian, and Swedish. *Proceedings of the XVth Scandinavian Conference of Linguistics*, Oslo, Norway, 50-61.

[9] Fujisaki, H., Nakamura, K., and Imoto, T. 1975. Auditory perception of duration of speech and non-speech stimuli. In Fant, G. and Tatham, M. (eds.) *Auditory analysis and perception of speech*. London: Academic Press, 197-219.

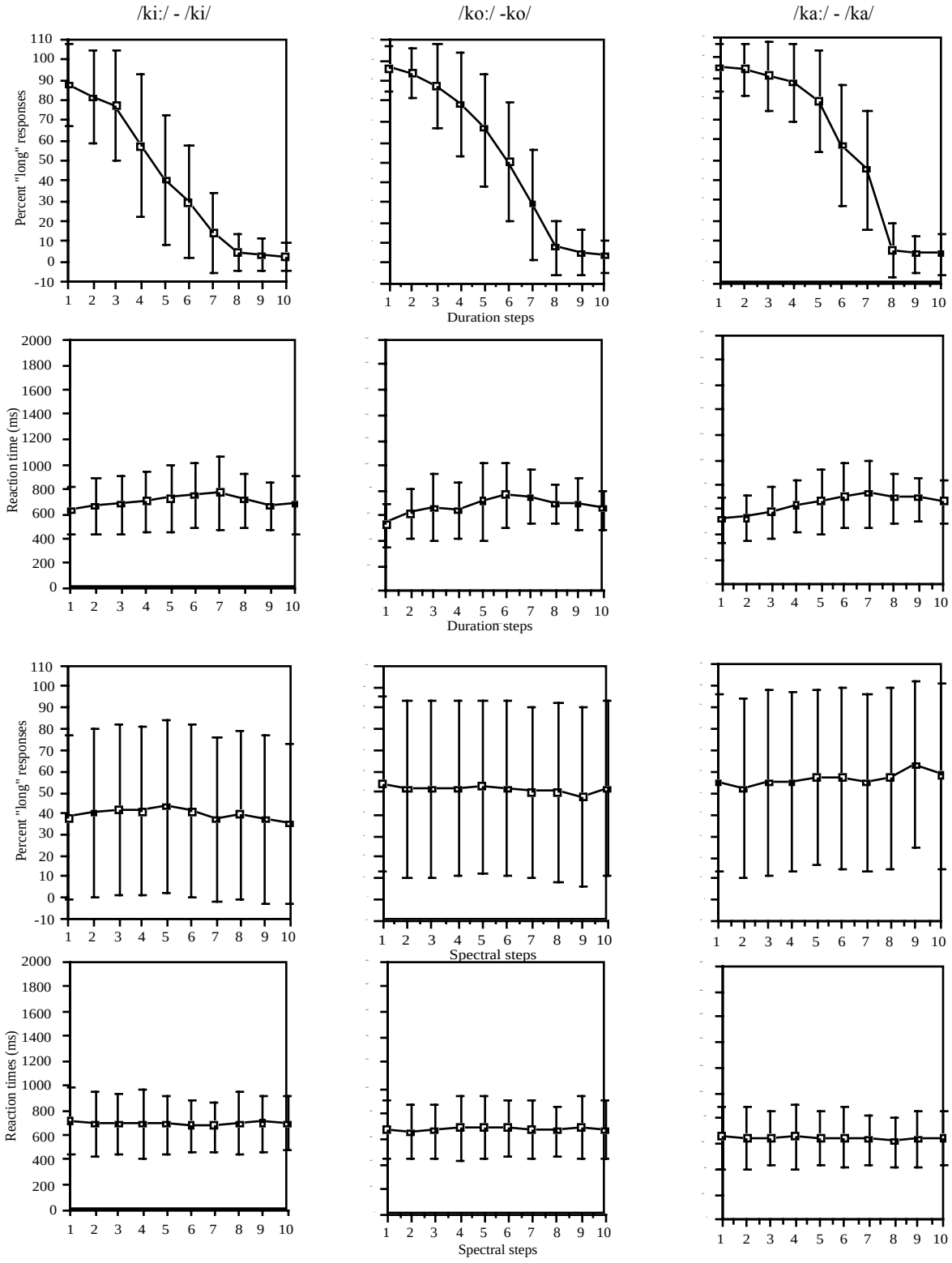

Figure 1. Mean percent long responses and reaction times are plotted for the 10 synthesized duration steps and spectral steps for the three pairs: /ki:/ - /ki/, /ko:/ - /ko/ and /ka:/ - /ka/.