Duration and F0 as perceptual cues to Japanese vowel quantity

Keisuke Kinoshita [1,2], Dawn M. Behne [1] and Takayuki Arai [2]

Norwegian University of Science and Technology, Norway [1]
Sophia University, Japan [2]
http://www.splab.ee.sophia.ac.jp

ABSTRACT

Vowel duration and local fundamental frequency changes are investigated as acoustical cues to vowel quantity identification by Japanese listeners. To examine the role of these factors, a perception experiment was carried out. The results indicate that, even though vowel duration serves as a dominant perceptual cue, when vowel quantity cannot be adequately cued by vowel duration alone, the F0 information within the vowel can be used to identify vowel quantity in Japanese.

1. INTRODUCTION

Vowel quantity refers to the phonological distinction of a vowel relative to one or more other vowels of similar timbre in the language.

One language traditionally characterized as having distinctions between long and short vowel quantities is Swedish [1]. Behne, Czigler and Sullivan [2][3] examined the effects of vowel duration and the first two formant frequencies on perceived vowel quantity identification in Swedish. Their results suggests that when the duration of a vowel is relatively long, vowel quantity might not be adequately cued by duration alone and might also make use of the vowel spectra to distinguish vowel quantities.

A similar investigation was carried out by Arai and Behne [4] for Japanese, which has distinct vowel quantities but is unrelated to Swedish. Japanese listeners were found to use vowel duration, but not spectral information as a perceptual cue to vowel 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 [5].

In spontaneous speech the adjustment of vowel duration affects numerous linguistically relevant factors in addition to vowel quantity, among them being speaking rate. The changes in vowel duration that occur in different speaking rates can lead to that a given vowel duration can be produced as a short vowel quantity in a slow speaking rate or a long vowel in a fast speaking rate [6].

When vowel quantity is not sufficiently cued by vowel duration, can listeners identify vowel quantity, and if so, how? One possibility could be fundamental frequency. Nagano-Madsen (1990) reported that pitch change observed within a vowel may adequately provides the distinctive to the perception of vowel quantity [7]. This report strongly motivates further investigation into how listeners use vowel duration and F0 change within the vowel as perceptual cues to vowel quantity.

This study examines these two parameters through a perception experiment using materials which includes systematic adjustments of vowel duration and F0, whereas [7] used two durational contexts as stimuli. The F0 synthesis in the current study is done by changing the F0 contour in target vowel and maintain the original accent pattern of target word. In [7], the F0 was manipulated not only in the target vowel, but also in the following mora and consequently the accent pattern of the stimuli would have differed from the original words, and may be the basis for vowel duration being found insignificant for vowel quantity identification. We hypothesized, based on [4] and [5], that duration is a dominant perceptual cue to identify vowel quantity Japanese, and when vowel quantity cannot be adequately cued by vowel duration alone, the declining F0 within the vowel may consequently be used.

2. METHOD

2.1. Production Experiment

2.1.1. Recordings and Measurements

The speech materials used in this investigation were produced by a male speaker. The speaker was a young adult native speaker of Japanese.

The five Japanese long-short vowel pairs ([a-a: e-e: i-i: o-o: u-u:]) were each used in a target word /zVza/. Like in [8], two mora words having a short vowel with an HL accent pattern and three mora words having a long vowel with an HLL accent pattern are used as target words. For each of the 10 words, the speaker produced 5 randomized repetitions of the sentence "kono pen wa zVza desu" ("This pen is zVza. ") at his natural speaking rate and natural intonation. Care was taken to produce target words containing a long vowel with a HLL accent pattern, modeled after the word *chiizu* “cheese” in Japanese. Target words containing a short vowel with an HL accent pattern were modeled after the word *chizu* "map".

From these 50 recordings (5 vowel pairs x 2 quantities x 5 repetitions), the three points of the target vowel illustrated in Figure 1 were identified: (1) the beginning of periodic waveform, (2) the location where the F0 declined by 2 Hz from the beginning of the target vowel, and (3) the beginning of friction associated with the following /z/. Based on these three points, the 4 measurements schematized in Figure 1 were made within the target word: duration of A, duration of B, the fundamental frequency at the beginning of the target vowel, ($F0_{in}$), and the fundamental frequency at the end of the target vowel ($F0_{out}$). Duration A refers to the interval from (1) to (2) where the F0 contour is relatively high and flat. Duration B refers to the interval from (2) to (3) where the F0 contour is declining from relatively high (H) to relatively low (L). For each of the 10 vowel conditions, the mean value of these four measures for the 5 repetitions was calculated and the utterance which best

corresponded to the mean values was chosen for resynthesis. These most representative items will be referred to as "selected productions."

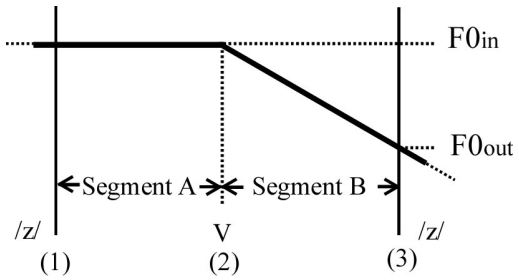

Figure 1: Schematic representation of the F0 contour within the target word. The measurement intervals (Segment A and B) between measurement points (1)-(3) are shown.

2.2. Resynthesis

The selected productions of the 10 target words and their measured values were the basis for resynthesizing the five sets of 100 words. The resynthesis was done using Praat. Since only the isolated word /zVza/ would be presented in the listening test, the carrier sentences, which were used in the production experiment, were cut out from the original recordings. Synthesis was carried out starting from the long vowel quantity of each pair and adjusting the signal toward the measured values of the corresponding selected short vowel.

For each set, the selected productions were used as extreme points of a 10x10 synthesis matrix, with 10 equal-sized steps of duration adjustment. and 10 equal-sized steps of F0 contour adjustment. For each duration step durations A and B were adjusted simultaneously. At each vowel duration step, 10 vowels were synthesized with different degrees of F0 adjustment. As shown in Figure 2, the F0 adjustment was done by changing the slope within segment B from the fully-descending version to a flat version. The values at the "starting F0" and "Target F0" column refer to the starting F0 (F0 at (2) in Figure 2) and target F0 (F0 at (3) in Figure 2), respectively. A fully-descending F0 contour corresponds to a naturally produced long vowel and a flat F0 contour to that of a short vowel quantity. The vowel duration and F0 values for each vowel pair and intermediating steps are summarized in Table 1.

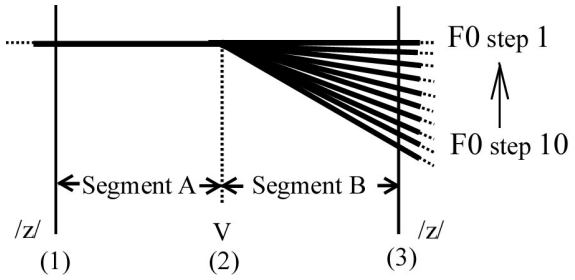

Figure 2: A schematic representation of the F0 adjustments made by changing the F0 slope within segment B.

Table 1: Vowel parameter settings for the five sets of resynthesized materials

Word Pairs		Segment A Duration [ms]	Segment B Duration [ms]	Segment B F0 [Hz] Starting F0	Segment B F0 [Hz] Target F0
/i:/	Step 10	162	77	152	115
↓	step size	11	2	0	4
/i/	Step 1	62	56	152	152
/e:/	Step10	155	90	148	115
↓	step size	9	6	0	4
/e/	Step 1	76	35	148	148
/a:/	Step 10	189	104	143	110
↓	step size	11	8	0	4
/a/	Step 1	90	30	143	143
/o:/	Step 10	182	68	146	115
↓	step size	10	4	0	3
/o/	Step 1	92	29	146	146
/u:/	Step 10	158	104	145	109
↓	step size	5	9	0	4
/u/	Step 1	111	26	145	145

3. PERCEPTION EXPERIMENT

Thirty-one native Japanese listeners between 18 and 26 years old with normal hearing participated in the study. The task was run at Sophia University, Japan.

The subjects were tested individually. A subject was seated wearing headphones at a computer terminal with a monitor and mouse. For each trial, subjects heard a synthesized word over the headphones and two /zVza/ words were presented on the monitor in katakana orthography. Since katakana is most likely to be used to express a nonsense word and also is familiar to Japanese listeners, katakana orthography was chosen. 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 asked to choose which of the two words on the monitor had the same vowel as the one they heard, and to respond by using the mouse to click one of two /zVza/ on the monitor as quickly as possible. In each trial reaction time was measured from the end of the auditory presentation of /zVza/ to the point when a response was given. The few responses which took more than 10 seconds, were disregarded.

Each subject's responses and reaction times were automatically logged in a file using a response collection program prepared for this purpose.

4. RESULTS AND DISCUSSIONS

For each vowel pair, the mean percent of responses for the word having a long vowel quantity were calculated. These are referred to as "percent long responses" in the following discussion.

An analysis of variance was carried out with duration steps and F0-contour steps as factors and percent long responses as a dependent variable. No reliable interactions were observed. The main effects are described below. Although reaction times are not presented here, the pattern of results generally supports those described below for listener responses.

4.1. Vowel duration

In Figure 3 the mean percent long responses are illustrated for the 10 duration steps for each of the five vowel sets. Reliable differences in percent long responses due to vowel duration were found for all five vowel sets: [F=1164.630; p<.0001] for /a/-/a:/, [F=581.818; p<.0001] for /i/-/i:/, [F=167.874; p<.0001] for /u/-/u:/, [F=986.862; p<.0001] for /e/-/e:/, [F=525.242; p<.0001] for /o/-/o:/. As expected, the highest percentage of long responses was obtained for synthesized items which were longest in duration, and a much lower percentage was found for shorter durations.

Differences were observed among the vowel sets. As can be seen in Figure 3, for /a/-/a:/ more long responses were given at relatively short duration steps than for /e/-/e:/, /i/-/i:/ and /o/-/o:/. For /a/-/a:/, the 50% crossover is located between duration step 3 and 4, whereas for /e/-/e:/, /i/-/i:/ and /o/-/o:/ the crossover occurs between duration step 4 and 6. This appears to be an artifact of the selected production of /a:/ having a relatively long duration (293ms) compared to the other /a:/ productions (μ=283ms, median = 287ms). Since the duration steps are equal-sized, the selected /a:/ production with a relatively long duration would likely lead to an increase in the proportion of long responses and the 50% crossover would consequently occur at a shorter duration step. For /u/-/u:/, the percentage of long response barely reaches 40%. The selected production of /u/ had a relatively long duration (138ms) compared to other short quantities (μ=112ms, median = 114ms). These variations taken into consideration, the results show a pattern of vowel duration being used by Japanese listeners to identify vowel quantity, and are consistent with previous research (e.g., Behne and Arai [5]).

4.2. Fundamental frequency contour

Of particular interest in this study is the extend to which Japanese listeners make use of the F0-contour to identify vowel quantity. Figure 3 shows the mean percent long responses for the 10 F0-contour steps for each of the five vowel sets. Compared to the s-curves observed across the duration steps, no reliable difference in the percent long responses is evident across the F0-contour steps. However for vowel sets /i:/-/i/, /u:/-/u/, /e:/-/e/ and /o:/-/o/, the response curve has a slightly positive slope. This suggests that although the F0-contour steps do not show a reliable difference in percent long responses at each duration steps, it might differ at particular duration steps. Of special interest are those duration steps where the 50% crossover occurred for percent long responses (i.e., between duration steps 3 and 4 for /a/-/a:/, between duration steps 4 and 6 for /i:/-/i/, /e:/-/e/ and /o:/-/o/, and between steps 1 and 2 for /u:/-/u/). We hypothesized that the response curve across the 10 F0-contour steps might have a strongly positive slope at duration steps near the 50% crossover point, whereas it might have a flatter slope at other duration steps. To investigate this, the data in Figure 3 was converted to Figure 4.

Three duration slopes were chosen to take a closer look at how listeners' responses are different when the duration information is sufficient to identify vowel quantity (duration step 1 and 10) and insufficient to identify vowel quantity (50% crossover). As was shown in Figure 3, the extreme durations at duration steps 1 and 10 generally provide listeners with adequate cues to identify short or long vowel quantities. At the 50 % crossover point, duration cannot sufficiently cue vowel quantity. Since the 50% crossover point varies from listener to listener, it was defined individually for each listener. Based on the crossover point for each listener, the average percent long response was derived for each vowel set. These values are plotted in Figure 4 for duration step 1, the 50% crossover point and duration step 10.

As is shown in Figure 4, for all vowels, the percent long response for duration steps 1 and 10 change very little across the 10 F0-contour steps, showing that listeners' responses are mostly independent of the F0-contour adjustments for these duration steps. However, the percent long responses at the 50% crossover point for duration generally differs across the 10 F0-contour steps; listeners identify a vowel as having a short vowel quantity when the F0 contour is flat (F0-contour step 1) and as having a long vowel quantity when the F0 contour is descending (F0-contour 10). This general pattern can be seen for /i/-/i:/, /u/-/u:/, /e/-/e:/ and /o/-/o:/ sets. For /a/-/a:/, this is not the case. At the 50% crossover for duration, the percent long response does not change across the F0-contour steps.

When listeners find that the duration information cannot appropriately cue vowel quantity, their responses shows that they are able to use the F0 contour of the vowel to identify vowel quantity. Why /a/-/a:/ differs in this respect is not clear. However, these results confirm that, vowel duration is a primary cue for identifying vowel quantity in Japanese. In addition, lacking duration information, the F0 contour within the vowel can generally serve as cue for vowel quantity identification.

5. CONCLUSIONS

In a previous study, Arai and Behne [3] examined the effects of vowel duration and the first two formant frequencies on perceived vowel quantity identification for Japanese listeners and concluded that their identification was cued by vowel duration. In spontaneous speech vowel duration can be adjusted due to other factors, such as speaking rate, and the use of vowel duration as a cue for vowel quantity may break down [6]. The results presented here show that the F0 contour within the vowel can serve as a supplementary cue in such cases, with the potential also to supplement duration and provide more robust cues to a listener identifying vowel quantity.

In this study, we investigated the role of vowel duration and the F0 contour within a vowel as a perceptual cue to vowel quantity for Japanese listeners. We examined the role of these factors through an identification task using materials which include systematic adjustments of vowel duration and F0 contour. Results indicated that, although vowel duration is a dominant perceptual cue, Japanese listeners generally can and do make use of F0 information when duration in itself is not sufficient enough to identify vowel quantity.

6. REFERENCES

[1] Elert C-C., ``Phonologic studies of quantity in Swedish.", Stockholm: Almqvist & Wiksell, 1964

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

[3] Behne, D. M., Czigler P. and Sullivan, K., ``Perceived vowel quantity in Swedish: Effects of postvocalic voicing..", Proc. Of the 16th International Congress of Acoustics and the 135th meetings of the Acoustical Society of America, 1963-64, 1998.

[4] Arai, T., Behne, D. M., Czigler, P. and Sullivan, K., ``Perceptual cues to vowel quantity evidence from Swedish and Japanese.", Proc. Of the Swedish Phonetics Conference (Fonetik), Vol. 81, pp.8-11, 1999

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

[6] Arai, T., ``A case study of spontaneous speech in Japanese.", Proc. Of the International Congress of Phonetic Sciences (ICPhS), Vol.1, 615-618, San Francisco, 1999.

[7] Nagano-Madsen, Y., ``Perception of mora in the three dialects of Japanese.", Proc of ICSLP, 1, 25-28, 199

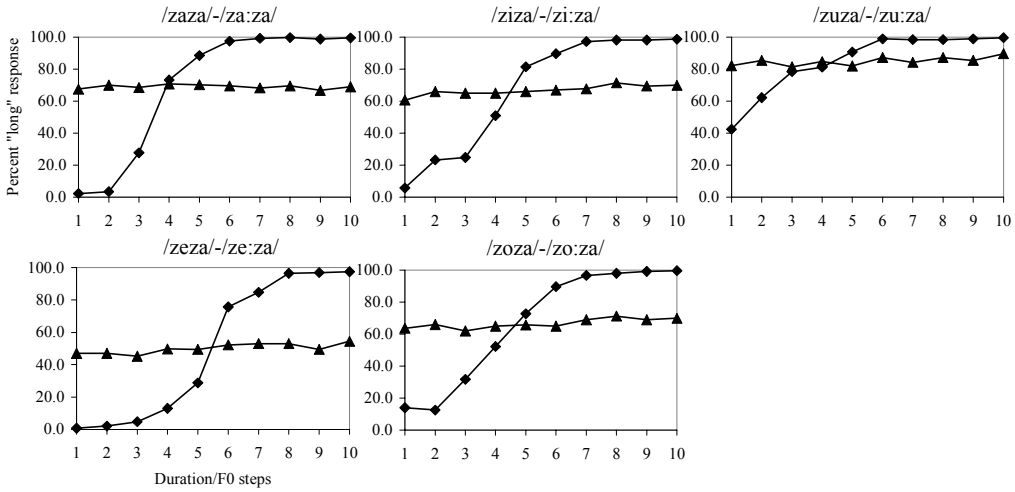

Figure 3: For each vowel set, the mean percent long responses is plotted for the 10 synthesized steps. ♦ and ▲ represent duration steps and F0-contour steps, respectively.

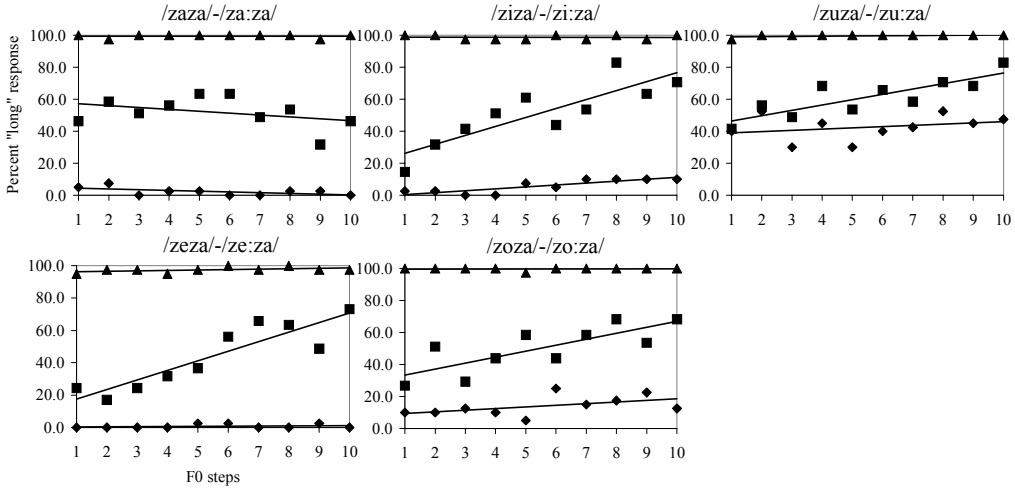

Figure 4: For all vowel sets, the mean percent long responses is plotted for the 10 synthesized F0-contour steps at three duration steps: step 1 (♦), the 50% crossover point (■), and step 10 (▲). The trend line is included for at each step.