Effects of syllable position and vowel context on Japanese /r/: Kinematic and perceptual data

William F. Katz1,*, Sonya Mehta1 and Matthew Wood2

1The University of Texas at Dallas, 800 W. Campbell Road, Richardson, Texas, 75080, USA
2The University of Texas at San Antonio, One UTSA Circle San Antonio, Texas, 78249, USA

(Received 11 April 2017, Accepted for publication 5 October 2017)

Abstract: In order to investigate the articulatory processes involved in producing Japanese /r/, we obtained speech recordings for native talkers of standard Japanese using an electromagnetic articulography (EMA) system. Each talker produced repetitions of /r/ in a carrier phrase designed to contrast syllable (CV and VCV VCV) and vowel (/a/, /i/, /u/, /e/, and /o/) contexts. Kinematic recordings were made using tongue (tip, TT; dorsum, TD; body, TB; left lateral, TLL; and right lateral, TRL) and lower lip/jaw (LL) sensors. We measured TT vertical displacement, TT duration at maximum position, and tongue blade width for the consonant gestures. In a perceptual experiment, American English listeners decided whether these consonants consisted of ‘l,’ ‘r,’ or ‘d.’ The kinematic results indicate Japanese talkers produced CV consonants with greater stricture and longer closures than consonants in intervocalic positions. CV productions also had narrower tongue blade widths than VCV VCV productions, especially in /i/ and /u/ contexts. The data were modeled with Dirichlet regression in order to determine how strongly tongue width and context (syllable and vowel) factors predict listeners’ judgments. The results showed a significant fit for ‘r’ judgments, with the tongue width fit successively increased by the addition of syllable and vowel context information.

Keywords: Speech production, Articulatory movement, Electromagnetic articulography, Speech perception

PACS number: 43.70.–h, 43.70.Mn [doi:10.1250/ast.39.130]

1. INTRODUCTION

The Japanese tap, often referred to as Japanese /r/, is linguistically unspecified for central or lateral features. Rather, this obstruent is broadly considered to be an alveolar or post-alveolar tap (or flap) [1,2] that varies widely, and can include approximant [3,4] and “weak plosive” qualities [2,5,6]. As noted by Magnuson [7], there are several reasons why this obstruent sound is of interest to speech researchers. First, the information is important developmentally, as Japanese infants show characteristic flap substitutions when acquiring /d/ [8]. Second, English learners of Japanese as a Second Language (JSL) frequently report difficulty mastering Japanese /r/ [2,9]. Given that JSL learners may well be hearing different consonantal qualities (‘l,’ ‘r,’ or ‘d’) produced by native Japanese instructors, this should, perhaps, not be very surprising. Third, studying the relationship between Japanese /r/ production and perception can address issues such as whether talker-dependent (e.g., strategic) or listener-dependent (e.g., acoustic/auditory) explanations best account for this type of sub-phonemic variation. For instance, a case study by Magnuson [10] showed that talkers produce more reduced forms (lateral and retroflex approximants) in an unknown speaking condition than in a known condition, suggesting sociophonetic variation.

One might also logically emphasize context-dependent factors and perceptual mechanisms. For instance, a flap would be expected to sound more like a stop than an approximant in a position following a pause than in an intrasyllabic position due to principles of acoustic contrast [11]. That is, following a period of silence, a brief voiced interval would be likely to be perceived as stop-like due to change-detection properties. In acoustic terms, a high co-occurrence of ‘l’ judgments preceding the vowel /o/ might be explained by the fact that the Japanese back, mid vowel /o/ has a low relative $F_2$ (approximately 1,000 Hz [12])

*e-mail: wkatz@utdallas.edu
resembling energy commonly found in General American English /l/ (1,100 or 1,200 Hz [13]).

Context-dependent articulatory factors should also logically correspond with listeners’ judgments. For example, taps produced in an intervocalic environment are expected to be produced more rapidly and with less stricture than those produced pre-pausally and at the start of a syllable. Similarly, perceived differences in consonant quality should relate, at least in principle, to relevant movement patterns and/or shape changes in talkers’ vocal tracts.

Several magnetic resonance imaging (MRI) and electropalatography (EPG) studies have described tongue shape characteristics for liquid approximant production in America English, noting characteristic differences in convexity across anterior/posterior tongue regions [14,15]. Building on this literature, a recent study from our lab estimated tongue width using an electromagnetic articulography (EMA) system and found these data useful for classifying alveolar consonants, including /r/ and /l/ [16].

Although some EPG, x-ray microbeam, and ultrasound data on Japanese /r/ have been obtained [17–20], perceptual and kinematic studies of this speech sound remain limited. As a result, it is unclear exactly which articulatory properties underlie the various perceived qualities of this sound. To address this shortcoming, we conducted two studies aimed at better understanding how Japanese taps are produced and how differences in their productions may reflect different perceptual interpretations by non-native listeners. Our study aims to:

(1) Determine whether the lingual kinematic patterns of Standard Japanese speakers demonstrate syllable, pause, and vowel context patterns, as suggested by the acoustic phonetic literature

(2) Decide whether tongue width and context (syllable and vowel) factors correspond with perceptual impressions of ‘l,’ ‘r,’ or ‘d’ quality, as judged by non-native listeners (speakers of General American English)

We hypothesize that Japanese /r/ will be produced with more stricture and with longer durations in CV than in VCV VCV syllables, and that tongue shape will vary across syllable and vowel context. We also predict that these factors will influence American English listeners’ judgments of Japanese /r/ quality.

2. KINEMATIC ANALYSIS

2.1. Talkers

Eight adult volunteers (ages 34–48 years, mean = 42) participated. All were members of the greater UT Dallas community. Two talkers were male and six were female. All were native speakers of standard Japanese (Kanto dialect). None reported any history of speech, hearing, or language difficulties.

2.2. Stimuli

Talkers produced the Japanese flap consonant in five different vowel environments (/ra/, /ri/, /ru/, /re/, and /ro/) in both /CV/ and /VCV VCV/ non-word utterances. These stimuli were embedded in the carrier phrase: /sore VCV VCV desu # CV/, meaning “this is VCV VCV (pause) CV” (e.g., “sore wa ara ara desu ... ra”). To minimize variability, talkers were instructed to produce the VCV utterances with an H-L (high-low) pitch accent, and a L2 speaker of Japanese (WK) provided a spoken example. Subjects repeated the stimulus list at least 10 times, at a natural volume and reading rate.

2.3. Tongue Motion Tracking Procedure

An AG501 EMA system (Carstens Medezinelektronic, GmbH) tracked both speech movement and a synchronized speech signal. Kinematic data were sampled at 1250 samples/sec, then down-sampled to 250 Hz for further processing. Speech data were recorded at 22.5 kHz (.wav format). Movement data were recorded for tongue sensors (tongue tip, TT; tongue dorsum, TD; tongue body, TB; tongue left lateral, TLL; tongue right lateral, TRL) and the lower lip/jaw (LL) sensor. The TT sensor was placed ~1 cm posterior to the apex. The TB sensor was placed ~4 cm posterior to the TT sensor, and the TD sensor was placed equidistant between TT and TB. The lateral sensors (TLL and TRL) were placed laterally from TD, ~1 cm from the tongue edge (see Fig. 1). Two sensors, attached to a pair of glasses worn by the subject established a head-reference frame. These head-reference data were used to provide head-independent movement information [21]. Before experimental recording, the talkers were engaged in approximately 5–10 minutes of conversation in order to have them adapt to the EMA sensors.
2.4. Kinematic Analysis

The kinematic data were corrected for head movement and filtered at 20 Hz to remove noise. Using the SMASH software program [22], key portions of each utterance were located based on kinematic and acoustic landmarks. The CV\textsuperscript{1} tap onset (e.g., /ara ara/) was defined as the maximum point of vertical tongue tip movement following the TT minimum for “wa.” The CV\textsuperscript{2} flap onset (e.g., /ara ara/) was defined as the next TT vertical prominent peak, and post-pausal CV\textsuperscript{3} (e.g., /ra/) was demarcated as the maximum point of vertical tongue tip excursion following “desu.” For each /r/ gesture, the 3-D (x, y, z) coordinates for the TT, TD, TB, and tongue lateral sensors (TLL and TRL) were obtained using customized MATLAB scripts.

Kinematic correlates of stricture were obtained for each talker by indexing the vertical position of the tongue tip (TT) sensor, yielding a position of maximum tongue tip height against the palate (TT\textsubscript{y\text{max}}). We also estimated the “maximum hold” of TT by calculating successive slope values for the TT trajectories and accepting those slopes that were less than a given threshold (operationally defined as 0.18). This was equivalent to computing the absolute value of the instantaneous velocity and selecting those points where velocity was within ±0.18. This yielded a series of TT\textsubscript{y\text{max}} plateaus, of which the duration was recorded (ms). Examples of these measures are shown in Figs. 2 and 3.

To obtain a tongue shape measure potentially relating to the perception of laterality, we measured the Euclidean distance (across the x, y, and z dimensions) between the lateral sensors (from TLL to TRL), providing us with an estimate of tongue blade width. This measure was motivated by a recent metric (lateral sensor distance to midline) found to be useful for classifying American English alveolar consonants from prior research in our lab [16].

2.5. Results

2.5.1. Stricture

Of the five vowel contexts, stimuli produced in the /a/ environment were selected for the stricture analysis because these productions afforded maximum vertical
tongue movement. Because each participant registered different displacement values as the result of occupying a slightly different position within the EMA magnetic field, the $y_{\text{max}}$ data were normalized before analysis. Each talker’s peak values were subtracted from the mean of their TT$_y$ minima, measured immediately following the first /r/ production (i.e., at the next minima, indicated by T1 in Fig. 1). This yielded a comparable series of TT$_y$ movements (mm) for each talker’s /r/ productions.

Figure 4 shows the TT$_{y_{\text{max}}}$ results as a series of plots for each talker’s CV and VCV productions. There was substantial individual variability between talkers, ranging from relatively large excursions (e.g., talker #2) to individuals showing much less movement (e.g., talker #7). We assume this inter-talker variability is largely due to anatomical differences in vocal tract size and shape. Of key importance is the finding that all talkers showed context-related stricture differences. Namely, stricture appears greater for the post-pausal productions (CV), and weaker in intervocalic positions. This pattern was confirmed statistically in a one-way repeated measures analysis of variance (ANOVA) comparing the mean displacement in mm for the talkers’ VC$_1$V VC$_2$V and CV productions. The results indicate a significant main effect for /r/ context ($F[2,14] = 12.884, p < 0.001$), with pairwise comparisons at the 95% confidence interval showing significant differences between post-pausal CV and the first and second intervocalic flaps (VC$_1$V and VC$_2$V, respectively). There was no significant difference between the two VCV VCV flaps.

2.5.2. Duration of flap closure

Figure 5 shows the “maximum hold” data (i.e., length of time the tongue tip is held against the palate) during /r/ productions first across the three different context conditions, then by talker. Individual talker variability is evident, with some talkers (e.g., #1 and 4) being relatively slow and variable, while others (e.g., #6) were remarkably fast and stable in their repeat productions. Context effects are also observable: post-pausal /ra/ productions are markedly longer for most talkers. This pattern was confirmed statistically in a one-way, repeated measures ANOVA with Greenhouse-Geisser adjustment for sphericity ($F[1.13, 7.93] = 9.56, p < 0.05$). Pairwise comparisons (95% CI) showed significant differences between the
post-pausal CV and the first (VC₁V) intervocalic consonant production. No other comparisons reached significance.

2.5.3. Tongue width

Euclidean distances were computed for the 3-D \( (x, y, z) \) space between each talker’s TLL and TRL sensors measured at \( TT_{\text{ymax}} \) for the consonant in CV and the first consonant in the VCV VCV utterances (VC₁V) across the five vowel contexts. Since the previous analyses indicated no significant difference in stricture or duration between the two intervocalic productions (and because the degree of laterality sounded similar between each talker’s VC₁V and VC₂V productions), VC₁V was arbitrarily selected for analysis.

The results show pronounced effects of vowel contexts (Fig. 6). Overall, VCV VCV contexts produced wider tongue positions, although there was an interaction with vowel context. For the high vowels (/i/ and /u/), the tongue was narrower, whereas the low vowel /a/ corresponded with the largest tongue width values. These patterns were investigated statistically in a two-way (syllable context × vowel context) repeated measures ANOVA. The results indicated a significant main effect of vowel context, \( (F[4, 10.47] = 5.70, p < 0.05) \), as well as a significant syllable context × vowel context interaction \( (F[4, 28] = 4.96, p < 0.01) \). No main effect of syllable context was found \( (F[1, 7] = 3.171, \text{ns}) \).

3. PERCEPTUAL ANALYSIS

3.1. Participants

Fourteen adult volunteers (ages 19–28 years, mean = 22.3 years) participated in the listening experiment. All were members of the greater UT Dallas community. Six were males and eight were females. All were native speakers of General American English. None reported any history of speech, hearing, or language difficulties. None reported familiarity with the Japanese language, nor proficiency in speaking a second language. Students received college research credits for participating.

3.2. Stimuli

Stimuli consisted of excised /CV/ (/ra/, /ri/, /ru/, /re/, and /ro/) and /VCV VCV/ productions (e.g., /ara ara/, /iri iri/, etc.) taken from the Japanese talkers’ sentence level productions in Experiment 1. Three stimuli were randomly selected for each talker, for a total of 240 stimuli (2 syllable conditions × 5 vowels × 8 talkers × 3 stimuli). Prior to playout, all stimuli were peak-amplitude balanced to ensure uniform loudness levels for listeners.

3.3. Procedure

Testing was conducted in a quiet room at the UT Dallas Speech Production lab. The experiment was monitored using DirectRT software (Empirisoft, v2016) on a desktop computer, with subjects listening through sealed earphones (Sennheiser, HD-500). Following Ueno and colleagues [23], each stimulus was played three times, with a 500 ms inter-stimulus interval (ISI), followed by a visual prompt (series of ‘X’ characters on the monitor), at which point the subject was to register a response. Participants first received task instructions, stating they would hear a series of sounds from native speakers of Japanese. Their goal was to determine whether the consonant sounded most like ‘l,’ ‘r,’ or ‘d’ and to indicate this by pressing one of three large, labelled keys on a keyboard. Subjects were encouraged to make a quick and accurate decision, and to “do their best.” Next, a 10-item practice set was given consisting of CV and VCV VCV stimuli spoken by Japanese talkers not included in the main experiment. This allowed the participant to become familiar with the stimuli, equipment, and procedure. After completing the practice set, participants began the main experiment. Participants worked at their own pace, advancing to the next stimulus by pressing the space bar when ready. Listeners received no direct feedback, although some motivational pictures were included at regular intervals (e.g., “Nice work!,” “Almost done”) during the task.

3.4. Results

The listening task took participants approximately 30 minutes to complete. To remove spuriously long reaction times (RT), outliers of greater than 3.0 standard deviations were eliminated. This procedure affected less than 2% of the data.

Figure 7 shows context effects on listeners’ consonant judgments for syllable context (post-pausal v. intervocalic).
Considering first the stimulus variables, phonetic context had marked effects on listeners’ judgments. For consonants in the CV environment, participants heard ‘d’ and ‘l’ more frequently, and heard ‘r’ less frequently. In contrast, in VCV VCV productions participants heard ‘r’ more frequently, and heard ‘l’ less frequently. A chi-square test showed a significant relationship between syllable environment and the frequency of consonant judgments ($\chi^2[2,N = 3,296] = 193.13, p < 0.001$).

Figure 8 shows listener judgments as a function of vowel context. While listeners were relatively consistent in the amount of ‘l’ judgments rendered across the five Japanese vowels heard, the amount of ‘r’ and ‘d’ judgments varied by vowel. For instance, ‘l’ judgments were highest for /o/ and lowest for /i/, yet ‘d’ judgments were highest for /i/ and lowest for /o/. A chi-square test revealed a significant relationship between vowel environment and the frequency of consonant judgments ($\chi^2[8,N = 3,296] = 702.219, p < 0.001$).

Individual participant data are listed in Appendices I and II as bar plots for the eight talkers and 14 listeners, respectively, grouped by response type. Both plots reveal a predominance of ‘d’ judgments (38.4%), followed by ‘r’ (32.6%), and ‘l’ (29%). These figures also show inter-participant variability contributing to the overall judgments.

4. COMPARISON OF PRODUCTION AND PERCEPTION

4.1. Dirichlet Regression

In order to model the data, listeners’ raw responses were converted to proportions and used as compositional data in a Dirichlet regression [24]. Briefly, this statistical procedure permits the analysis of compositional variables (such as proportions) in a GLM-like framework. Using an R-package [25,26], we built four models testing the following sets of predictors: (1) width, (2) width + syllable context, (3) width + vowel context, and (4) width + syllable context + vowel context. In each model, no interaction terms were included and common parameterizations were used.

4.2. Results

The base model included only a main effect of tongue width. The model fit ($AIC = -840.1, df = 6, n = 240$) was only marginally improved by adding syllable context ($AIC = -883.9, df = 9$) but markedly improved with vowel context ($AIC = -950.6, df = 18$). The best-fit model was found by adding both syllable and vowel context ($AIC = -1.028, df = 21$). The base model showed a significant fit to /r/ judgments, and not /d/ or /l/, as reflected by Beta-coefficients for tongue width (estimate = 0.67553, $SE = 0.021813$, $z = 3.076$, $p = 0.01$) and /r/ (estimate = 0.66112, $SE = 0.22578$, $z = 2.898$, $p = 0.01$), whereas for /l/ the tongue width fit was not significant. This suggests that the proportion of /d/ and /r/ responses increased significantly as tongue width increased. Beta-coefficient terms for /d/ judgments showed that the proportion of /d/ responses was significantly greater in /e/, /u/, and /i/ contexts as compared with /a/ contexts. (/e/: estimate = 0.67553, $SE = 0.021813$, $z = 3.076$, $p = 0.01$; /u/: estimate = 0.66112, $SE = 0.22578$, $z = 2.898$, $p = 0.01$; /i/: estimate = 1.64817, $SE = 0.24652$, $z = 6.686$, $p = 0.001$). For /l/ judgments, tongue width was mostly affected by /i/ (estimate = −0.08614, $SE = 0.22316$, $z = −3.613$, $p = 0.001$) and to a lesser extent by /u/ (estimate = −0.5522041, $SE = 0.225434$, $z = −2.316$, $p = 0.05$), showing the proportion of /l/ responses decreased significantly in /i/ contexts as compared with
/a/ contexts. Overall, the findings suggest tongue width relates most to /r/ judgments, while vowel height seems to play an important role, as differences between /i/ and /a/ strongly affect the proportions of listeners’ judgments.

5. GENERAL DISCUSSION

We hypothesized that Japanese /r/ would be produced with greater stricture and longer durations in CV than in VCV CV syllables, and that tongue shape would vary across syllable and vowel context. With few exceptions, these general hypotheses were met. Examining a subset of the recorded data (/ara ara/ and /ra/), it was found that peak vertical displacement of the TT sensor was significantly greater for /r/ in the prevocal and post-pausal environment (/ra/) than in the intervocal (/ara ara/) context. Similarly, a uniform measure designed to estimate the duration of the flap gesture (“maximum hold”), indicated significantly longer closure for /ra/ than /ara ara/ productions. These findings replicate prior reports of a greater degree of tongue contact at the alveolar ridge for post-pausal versus intervocalic contexts [9] and support the notion that rapid speech gestures, such as a tap or flap, will be realized in relative isolation with greater displacement and duration than in contexts that pose concurrent demands for adjacent sound production. Although the amount of stricture and duration did not differ statistically between the two intervocalic peaks (VC1V and VC2V), there were some interesting talker-specific patterns noted: Some speakers had markedly more pronounced (and longer) first VC1V excursions, others emphasized the second peak (VC2V), while still others showed similar values for the two peaks. Future studies might investigate such talker-specific patterns of tap/flap realization as possible sociophonetic or dialectal markers (e.g., [10]).

The current results suggest substantial vowel-dependent variability, widest for /a/ and narrowest for /i/ and /u/. These vowel-dependent patterns also appear to agree with prior findings [9,17,18], although a potential limitation of the current data is that only VC1V was analyzed and it is therefore possible that VC2V may have revealed subphonemic laterality effects not heard by the investigators (‘covert contrasts’). This possibility should be investigated in future studies. In addition, two talkers found it difficult to produce /iri iri/ under EMA conditions, even after the 10 minute warm-up period. This suggests that production of the Japanese flap may be especially vulnerable to sensor interference for some subjects during EMA recording. Future study of this issue using other imaging techniques (e.g., ultrasound, MRI) may be helpful.

We also propose that while the VCV VCV tongue shape differences are strongly affected by vowel-context requirements (e.g., narrowing in order to extend towards the alveolar ridge for /i/, flattening for /a/, backing for /o/) the resultant tongue shapes may also have consequences for consonant perception, biasing English listeners towards /r/ or /l/ variants.

Some support for this notion may be found in the results of the Dirichlet regression. Although only a small amount of the total variance was accounted for in the model, the results nevertheless show that tongue width relates most to ‘r’ judgments (particularly in the intervocalic context), while vowel context also interacts in rather specific way with listeners’ proportions of /r/ and /d/ judgments, especially for /i/. In other words, it does not simply seem to be tongue height alone predicts American listeners’ attribution of which Japanese tap/flaps are judged as rhotics, weak stops, or laterals. Rather, tongue shape interacts with vowel context in an interesting way: the wider the tongue, the more /d/ and /r/ judgments, and less /l/ judgments. Future studies might use this type of modeling framework to further investigate other articulatory factors that can characterize Japanese /r/, including the direction of tap/flap production (back-to-front or front-to-back), as well as a broader inventory of taps/flaps (including palatalized variants) in order to better understand this topic.

ACKNOWLEDGMENTS

The authors wish to thank Patrick Reidy for his help with statistical analysis, Thomas Campbell for support and organization in providing the Callier Center C-Tech Center where this work was conducted, and the many participants who volunteered their time.

REFERENCES


APPENDIX

Fig. A1 Consonant judgment by listener. Each shaded bar represents a unique listener (1–14). Total responses = 3,295.

Fig. A2 Consonant judgment by talker. Each shaded bar represents a unique talker (1–8). Total responses = 3,295.