Language Learning & Technology
http://llt.msu.edu/issues/june2016/okunohardison.pdf
June 2016, Volume 20, Number 2
pp. 61–80

PERCEPTION-PRODUCTION LINK IN L2 JAPANESE VOWEL DURATION: TRAINING WITH TECHNOLOGY

Tomoko Okuno, **University of Michigan**
Debra M. Hardison, **Michigan State University**

This study examined factors affecting perception training of vowel duration in L2 Japanese with transfer to production. In a pre-test, training, post-test design, 48 L1 English speakers were assigned to one of three groups: auditory-visual (AV) training using waveform displays, auditory-only (A-only), or no training. Within-group variables were vowel, preceding consonant, pitch pattern, and training talker's voice. Perception pre- and post-tests measured identification accuracy and response time (RT). Training involved eight sessions with feedback, including waveforms for the AV group. Results indicated significant improvement for the AV and A-only groups with generalization to novel stimuli and a new voice as well as transfer to production; the AV group showed a greater rate of improvement. Participants found waveform displays very helpful. Vowel type, preceding consonant, and pitch pattern significantly affected perception in testing and training as did the training talker's voice. The easiest pitch pattern was Low–High in the first syllable, perhaps reflecting English prosodic preference, and High–High in the second, which may be more salient. Perception was facilitated by talkers demonstrating greater pitch movement. Accuracy and RTs increased after training; participants reported spending more time evaluating post-training input. Results support the perception-production link, and the role of variable talker- and context-dependent perceptual categories.

Language Learned in this Study: Japanese

Keywords: Computer-assisted Language Learning, Listening, Pronunciation, Research Methods, Second Language Acquisition

APA Citation: Okuno, T., & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology*, *20*(2), 61–80. Retrieved from http://llt.msu.edu/issues/june2016/okunohardison.pdf

Received: March 16, 2015; **Accepted:** September 12, 2015; **Published:** June 1, 2016

INTRODUCTION

Studies involving the training of second language (L2) learners to improve their perception of nonnative sounds date back several decades. After years of mixed results, in the early 1990s, a series of auditory training studies reported significant improvement in the perceptual identification accuracy of American English (AE) /r/ and /l/ by learners whose first language (L1) was Japanese (e.g., Lively, Logan, & Pisoni, 1993). These studies demonstrated a benefit to training with natural stimuli produced by multiple talkers, and generalization of improved performance to perception of novel stimuli and those produced by a new voice. Subsequent research found a transfer of the benefits of perception training to production improvement and retention of improved skills even in the absence of continued L2 input (e.g., Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). From these accomplishments, the following hallmarks of successful auditory perception training emerged: (a) multiple talkers producing sufficient exemplars in natural speech to represent the variability of the target sounds, (b) feedback during training, and (c) the use of an identification (vs. a discrimination) task. Other researchers then implemented this high variability paradigm to explore the contribution of visual speech cues from a talker's face to learners'

identification accuracy (e.g., Hardison, 2003; Hirata & Kelly, 2010) and earlier word identification (Hardison, 2005b). From a theoretical perspective, the above findings suggested that L2 learners may rely on multimodal context- and talker-dependent exemplars as representations in memory to which input can be matched for identification (e.g., Hardison, 2003, 2012; Lively et al. 1993) in contrast to prototypical representations (i.e., phonemes) resulting from the abstraction of talker and contextual information from the speech signal.

LITERATURE REVIEW

Computer-based Visual Feedback for Training

Visual cues have produced beneficial outcomes for learners of various target languages through the use of computerized displays. Displays offered by speech analysis software include fundamental frequency (pitch), spectrograms, and waveforms. The current study explored the efficacy of computer-based training to facilitate the acquisition of duration contrasts in L2 Japanese vowels.

Several early reports advocated the use of pitch displays to improve learners' intonation (e.g., de Bot, 1983; Leather, 1990; Pennington & Esling, 1996). They are user-friendly and informative by making visually salient some of the challenging features in L2 speech (e.g., Chun, 1998; Chun, Hardison, & Pennington, 2008). In a study on the acquisition of L2 French prosody, learners (L1 AE) showed significant improvement in both prosody and segmental accuracy in sentence productions following training using real-time displays along with feedback from displays of native speaker (NS) productions of the same sentences (Hardison, 2004). Participants reported that they (a) became more aware of the pitch differences between native and learner speech because they could see them and (b) found visual feedback in training very helpful.

Following the discussions by Chun (2002) and Levis and Pickering (2004) on the value of pitch displays at the discourse level of speech, Hardison (2005a) explored the training potential of visualizing pitch synchronized with the video of the oral presentations of advanced L2 speakers of English (L1 Chinese). Anvil (Kipp, 2001), a web-based video annotation tool, was used to integrate segments of the learners' videorecorded oral presentations with the associated pitch tracking. Participants showed improvement over several weeks; the presence of synchronized video was particularly helpful with discourse-level input. An alternative type of intonation feedback was developed by Hincks and Edlund (2009) consisting of flashing lights to show learners how much pitch variation they had produced. This feedback significantly improved the pitch variation made by L1 Chinese learners of English in their oral presentations.

Spectrograms were also helpful for lower-proficiency learners of L2 Spanish in raising awareness to the spectral differences (i.e., those related to the component frequencies that make up a sound) between stop consonants /b, d, g/ and their intervocalic realizations as approximants [β, ð, ɣ] in Spanish (Olson, 2014). Only brief tutorials on the use of Praat (Boersma & Weenink, 2014) were required for the participants to make use of this tool.

For durational contrasts, waveforms, requiring only minimal instruction, were used to compare auditory-visual (AV) and auditory-only (A-only) web-based training for beginning-level learners of L2 Japanese (L1 AE) in the perception of geminate consonants (Motohashi-Saigo & Hardison, 2009). Geminates have a longer duration than their singleton counterparts. Segmental duration is a contrastive feature in Japanese and important for communication. Stimuli were singleton and geminate /t/, /k/, and /s/ followed by /a/ or /u/ (produced as a high back unrounded vowel with lip compression). Identification accuracy improved significantly, especially for the AV group who saw waveform displays as feedback during training. Production of geminates also improved significantly, especially for the AV group. In post-study interviews, learners commented that taking web-based training outside of regular classes was convenient and that they enjoyed the immediate feedback.

In terms of vowels, Wang and Munro (2004) used synthesized speech to train L2 English speakers (L1 Chinese) to ignore duration, and focus on the vowel quality differences between the members of the pairs [i]-[ɪ], [u]-[ʊ], and [ɛ]-[æ]. Learners used a self-paced procedure and showed significant improvement on all three vowel contrasts after two months of training, and retained performance levels 3 months later. Using the software program Sona-Match (KayPentax), Carey (2004) found that Korean speakers were able to improve their production of the vowel [æ] in citation form, but not in continuous speech. Sona-Match uses the first and second formant frequencies of a speaker's vowel production to plot it in a vowel space on the computer screen.

Finally, with reference to vowel length contrasts in Japanese, the focus of the current study, Hirata and Kelly (2010) found that L1 English speakers with no prior exposure to Japanese showed greater improvement in identifying vowel duration after four sessions of perception training if they were able to see the speaker's mouth on the computer screen.

Phonological Features of Japanese Vowels

One of the challenges in acquiring durational contrasts in L2 Japanese is the role of the mora, which is a unit of timing in the language that is important in both perception and production (e.g., Kubozono, 1999). Special morae include the second half of a long vowel; for example, *kiite* "listening" has three morae /ki-i-te/ which form two syllables /kii.te/. In addition, Japanese has a pitch-accent system. As the current study concerns Tokyo Japanese, we focus on that dialect's patterns in which accent is realized as a high (H) pitch followed by a low (L) pitch (Haraguchi, 1999). The location of the accent corresponds to the mora before the pitch drop. In the case of a long vowel, the first mora carries the H pitch so that an HL pitch contour occurs within the long vowel.

Thus, accented long vowels have two phonologically distinctive features—duration and pitch-accent—which can differentiate meaning between words. Figure 1 shows the pitch patterns selected for this study. They occur in *kyootsuu-go*, the common variety of Japanese widely used in the Tokyo dialect (Shibatani, 1990). The examples are labeled according to syllable structure type and pitch pattern within these categories: unaccented, initial-, second-, and third-mora accented (Haraguchi, 1999).

L2 learners of Japanese appear to be sensitive to pitch pattern differences. Minagawa (1997) investigated whether the patterns HH, LL, HL, and LH affected perception of vowel duration for learners whose L1s were Korean, Chinese, English, Spanish, and Thai. Results indicated (a) greater perception accuracy of long vowels with the HH pattern and (b) a tendency to perceive a long vowel as a short vowel when the word had a LL pattern. Koguma (2000) found that long vowels in word-final position appeared to be more difficult for learners to perceive accurately compared to those in word-initial position.

In the current study, we investigated the effects of both pitch pattern and syllable location of the vowel (i.e., in the first or second syllable) in a word on response accuracy and latency in the perception of vowel duration in L2 Japanese by L1 AE speakers. A pre- and post-test design with controls was used to address the following research questions: (a) What are the effects of training with and without waveform displays on the perception accuracy and response latency of vowel duration by L2 learners of Japanese? (b) Does pitch pattern and syllable location of the long vowel affect perception? (c) Are there talker effects in perception accuracy and latency? (d) Does perception training transfer to production improvement in the absence of explicit instruction? (e) Does training generalize to perception of novel stimuli and those produced by a new talker's voice? Based on previous research (e.g., Motohashi-Saigo & Hardison, 2009), we hypothesized that training overall would be successful but that variability would also be found in response to talker and stimulus variables such as pitch pattern and syllable location of the vowel.

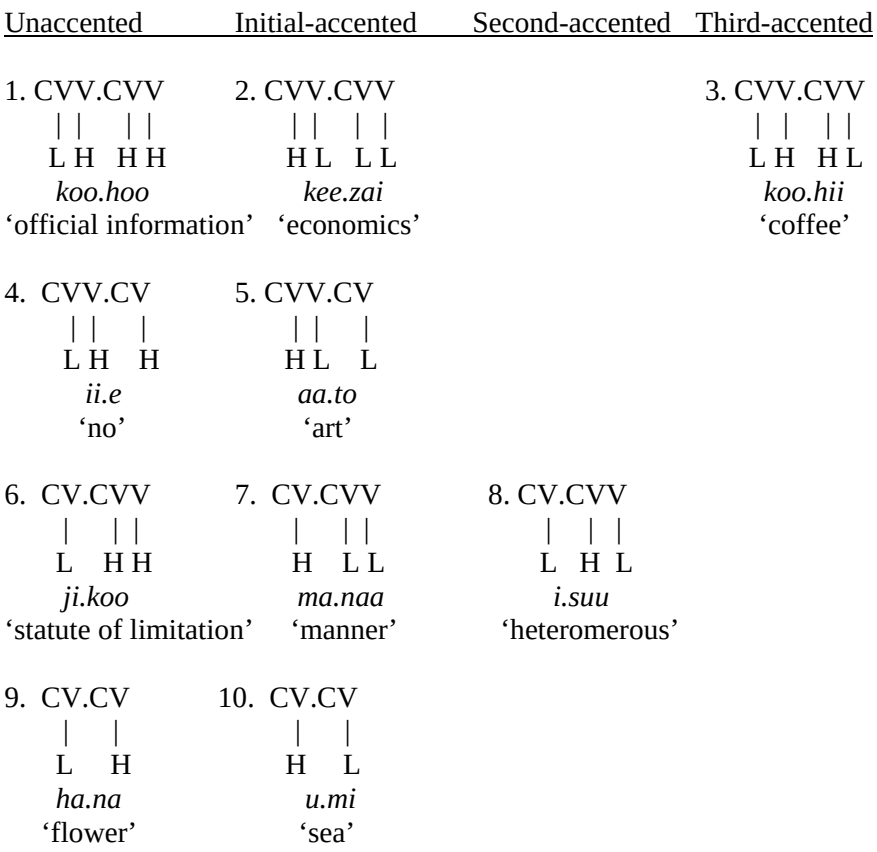

Figure 1. Pitch patterns in the Tokyo dialect used in the study. The period denotes a syllable boundary.

METHOD

Participants

A total of 64 learners of Japanese as a foreign language (aged 18–22; 36 female, 28 male; L1 AE) volunteered to participate in this study at a large Midwestern university in the US. There were no heritage speakers of Japanese or students who had studied abroad in Japan. All reported normal hearing and vision. They were enrolled in Japanese courses as follows: first year of study (n = 24), second year (n = 17), third year (n = 16), and fourth year (n = 7). The course instructors corrected inaccurate pronunciation during oral drills and communicative activities, but did not focus on segmental duration or pitch-accent.

Materials

Production Testing

Materials for production testing included 4 practice tokens (e.g., *noono)* and 16 bisyllabic tokens contrasting long and short vowels in a range of syllable structure types (see Appendix A). The consonants /k/ and /s/ were combined with vowels /a/ and /u/ to construct pseudowords to avoid the effects of word frequency and learner knowledge (e.g., Bundgaard-Nielsen, Best, & Tyler, 2011). The stop and fricative were selected based on the findings of Hardison and Motohashi-Saigo (2010), suggesting a role in learner perception for consonant-vowel sonority differences[1]. The vowels /a/ and /u/ represent the longest and shortest vowels respectively in the Tokyo dialect (Yoshida, 2006). The choice of materials was guided by

classroom observations of learner difficulty and the results of pilot testing with a peer group (N = 40), which indicated that tokens with a long vowel in the first syllable (CVV.CV) or in both syllables (CVV.CVV) were produced more accurately than those with a long vowel in the second syllable (CV.CVV) or a short vowel in each syllable (CV.CV).

Perception Testing and Training

Perception materials included two sets of four practice stimuli each (one set for testing; one for training); neither involved the consonants (/k/ and /s/) or vowels (/a/ and /u/) under test. Selection of syllable types and pitch patterns for perception testing (Appendix B) and training (Appendix C) was guided by the results of the aforementioned peer group pilot testing with a large number of tokens from all patterns in Figure 1. Findings revealed that learners had more difficulty identifying vowel duration in the second syllable, especially with an HL or LL (vs. HH) pattern. In the current study, most items were pseudowords; five in the training set were real words but occurred infrequently in the participants' instructional input.

Testing stimuli were produced by a female NS of Japanese from Tokyo and digitally recorded. Training stimuli were produced by four NSs of Japanese (2 female, 2 male) from Tokyo. With the goal of providing natural speech, the vowel /u/ was allowed to devoice as this can occur in the Tokyo dialect when it follows a voiceless consonant and either precedes a voiceless consonant or occurs in word-final position, unless the vowel is in a position to receive an accent (Tsujimura, 2007). Speech rate was consistent throughout the recordings. There was no significant difference in duration between accented and unaccented vowels (e.g., for long /a/, $U = 9.00$, $p = .412$). All stimuli were accurately identified by NSs prior to their use.

For the AV training condition, waveforms generated in Praat served as visual input. Examples from the study are shown in Figure 2. The vertical lines drawn on each side of /k/ represent the closure period for the stop. The examples in the top row have a long vowel (i.e., *kaaka* and *suusu*); those in the bottom row have a short vowel (i.e., *kaka* and *susu*).

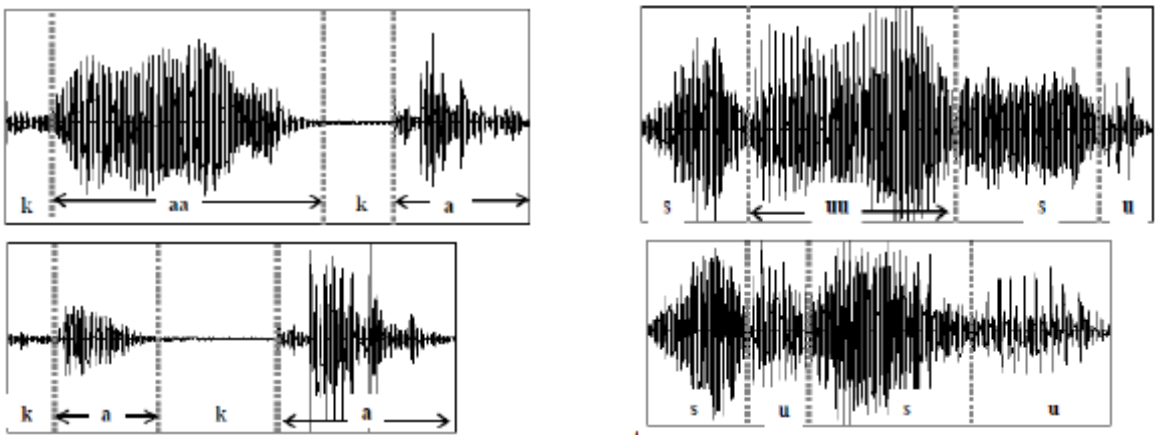

Figure 2. Examples of waveform displays used in the study.

Tests of Generalization

Tests of generalization (TGs) investigated the generalizability of training to the perception of unfamiliar stimuli produced by a familiar female voice from training (TG1), and familiar stimuli produced by an unfamiliar female voice (TG2). The unfamiliar stimuli for TG1 involved the consonant /t/ and vowel /e/ (see Appendix D), which had not been used in the pre-test, post-test, or training.

Procedure

Production Testing

A computer-based production test was created using E-Prime and administered to participants individually in a quiet room prior to the perception test to avoid exposing them to auditory input involving the target tokens. Practice tokens familiarized the participants with the task. All were written in *roomaji* (the Roman alphabet representation of Japanese sounds) versus *hiragana*. Classroom experience with similar learners indicated that *roomaji* made the vowel duration distinction clearer. Participants first read these instructions on the screen: "After a plus sign (+), a word will appear on the computer screen. When you are ready to say the word, press 'p' and say it." The plus sign appeared for 2 s. After the participant pressed the *p* key, the screen was cleared. Productions were digitally recorded and saved.

Perception Testing

After the production test, a forced-choice, four-alternative perceptual identification task was administered. Four practice stimuli familiarized participants with the task. The 18 testing stimuli were then presented. As with the production test, participants were alerted with a plus sign in the middle of the screen for 2 s, and instructed as follows: "You will hear a word. Please choose what you think you just heard from the list of words by pressing 1, 2, 3, or 4. Please choose as quickly and accurately as possible." While listening, participants were able to see the response options (similar to Figure 3, but without the waveform). Each stimulus was heard once. When it ended, the timer to measure the response time (RT) started. When a response was made, the timer stopped. The screen then showed the plus sign again and moved to the next trial. There was no feedback during testing. For each stimulus, a participant's identification accuracy and RT were recorded by the computer and saved.

Perception Training

Participants were divided into two groups: one received AV training where the visual input was a waveform display, and the other received A-only training. Between the pre-test and post-test, participants in both groups took 8 perception training sessions individually (approximately 25 min each, 4 days per week for 2 weeks). A forced-choice four-alternative identification task was used similar to the testing format. Before the first session, the AV training group received instruction for about 5 min, including a demonstration of how long and short vowels appeared in waveforms while they listened to audio files unrelated to the study. Participants in the AV training group listened to the stimulus and chose what they heard from the list of options while watching the associated waveform (see Figure 3, *kaaka*).

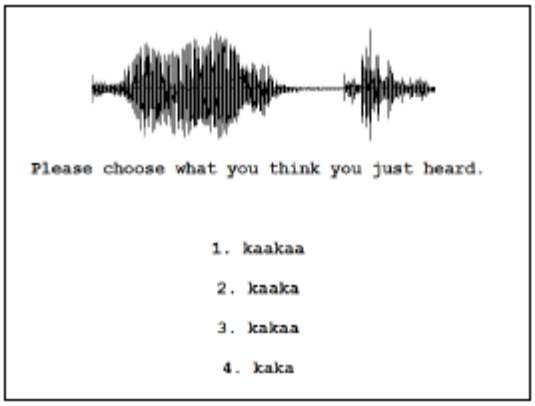

Figure 3. Perception training identification task for AV training group.

Participants in the A-only group did not see the waveform; otherwise, the procedure was the same. After participants in both groups made their selection, the correct item (and waveform for the AV group)

appeared on the screen and the audio was replayed. Identification accuracy and RTs were saved for analysis.

Tests of Generalization

TGs were administered to both training groups after the post-test. Stimulus presentation followed the same format and procedure as for the pre-test and post-test. Both perception accuracy and RT were recorded for comparison with the post-test data to see if improvement from training had generalized to novel stimuli and a new voice. At the conclusion of all testing, the first author interviewed each participant about their perceptions of the efficacy of the training.

RESULTS

It was determined a priori that participants whose scores were 90% or higher in the perception pre-test would be excluded from training. From the initial 64 participants, 12 were excluded due to this criterion. These 12 were comprised of three participants from each of the four years of Japanese study, including a first-year student who scored 100%. Of the 52 remaining participants, four did not complete all tasks; therefore, their data were not analyzed. For the final 48 participants, there was no correlation between pre-test identification accuracy and year of study; therefore, their data were combined for analysis.

Results are presented in the following order: (a) comparability of groups in pre-test perception accuracy and latency, (b) effects of training and stimulus variables on accuracy, (c) effects of training and stimulus variables on response latency, (d) analysis of training effects per group, (e) effects of perception training on production, and (f) TGs. For statistical analysis, the alpha level was set at .05. Data met the assumptions of all tests.

Perception

Comparability of Groups at Pre-test

The participants were divided into three groups: AV training group (n = 16), A-only training group (n = 16), and control (i.e., no training) group (n = 16). Each group included a similar range of pre-test perceptual identification accuracy scores. Both response accuracy and RT data were analyzed. In terms of accuracy, the participants' choice was coded as either correct (one point) or incorrect (zero). When a choice was not made, no point was given. RT was measured in ms using E-Prime. Two one-way ANOVAs confirmed the comparability of the groups at pre-test in identification accuracy, $F(2, 47) = 0.424$, $p = .657$ and RT, $F(2, 47) = 0.076$, $p = .927$.

Effects of Training and Stimulus Variables on Accuracy

The descriptive statistics for perception accuracy in the pre-test and post-test for each group (AV, A-only, control) are shown in Table 1. Both training groups increased in mean accuracy and the standard deviations decreased, especially for the AV group.

Table 1. *Descriptive Statistics for Perception Accuracy in the Pre-test and Post-test*

Group	*n*	Mean Percent Identification Accuracy (*SD*)	
		Pre-test	Post-test
AV	16	.69 (.16)	.97 (.09)
A-only	16	.71 (.15)	.88 (.13)
Control	16	.66 (.17)	.64 (.20)

An ANOVA was conducted to determine if the training had been effective in improving identification accuracy of vowel duration. The within-group factor was time (2); the between-groups factor was group type (3). Results indicated a significant effect of time for perception accuracy, $F(1, 45) = 68.275$, $p < .001$,

η_p^2 = .603[2], and group, $F(2, 45) = 6.956$, $p = .002$, $\eta_p^2 = .236$. The Time × Group interaction was also significant, $F(2, 45) = 25.271$, $p < .001$, $\eta_p^2 = .529$. Tukey's HSD tests indicated that the control group was significantly different from the AV group ($p = .003$) and the A-only group ($p = .018$). There was no significant difference between the two training groups ($p = .788$) although overall accuracy increased more for the AV group and at a faster rate as shown in Figure 4.

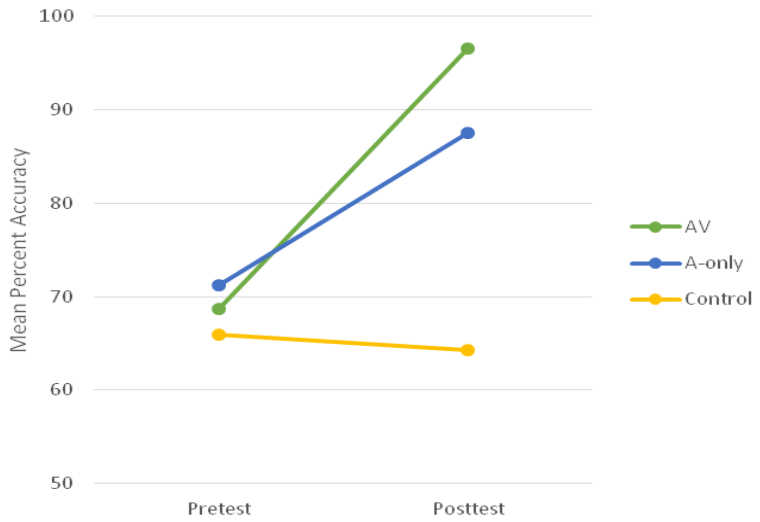

Figure 4. Mean percent accuracy for perception at pre-test and post-test by group.

We also examined how stimulus variables affected identification accuracy of vowel duration and how this changed from pre-test to post-test as a result of training. Earlier pilot testing suggested that the location of the long vowel in the second (vs. first) syllable was particularly problematic for learners. This finding was used in the present study to place the syllable structure types and pitch patterns into three categories to facilitate analysis. With reference to Figure 1, pitch patterns (1) LH.HH, (2) HL.LL, and (3) LH.HL formed Group I containing two long vowels (CVV.CVV); patterns (4) LH.H and (5) HL.L formed Group II with one long vowel in the first syllable (CVV.CV); and patterns (6) L.HH, (7) H.LL, and (8) L.HL formed Group III with one long vowel in the second syllable (CV.CVV). The pattern CV.CV was not included in testing but was used in training for contrastive purposes.

An ANOVA revealed a significant effect of pattern group, $F(2, 126) = 10.866$, $p < .001$, $\eta_p^2 = .147$. Data were further analyzed within each of the three pattern groups. Within-group factors were time (2), preceding consonant (2), vowel (2), and pattern type (for Group I, 3 patterns; Group II, 2 patterns; Group III, 3 patterns); the between-groups factor was training type (2). Discussion focuses on the significant findings.

For Group I (CVV.CVV), there was a significant effect of time, $F(1, 30) = 44.885$, $p < .001$, $\eta_p^2 = .599$, but not training type, $F(1, 30) = .839$, $p = .367$. Overall mean percent accuracy at pre-test was .62 ($SD = .15$) which increased to .91 ($SD = .06$) at post-test. There was a significant effect of pattern type, $F(2, 126) = 10.866$, $p < .001$, $\eta_p^2 = .147$. Pairwise comparisons indicated that LH.HH showed significantly greater accuracy compared to LH.HL ($p < .001$) and HL.LL ($p < .001$). LH.HH and LH.HL share the same pitch pattern on the first syllable but differ in the second syllable (HH and HL respectively), suggesting that this difference may have played a role.

For Group II (CVV.CV), results indicated significant effects of time, $F(1, 30) = 10.083$, $p = .003$, $\eta_p^2 = .252$; preceding consonant, $F(1, 63) = 7.471$, $p = .008$, $\eta_p^2 = .106$; pattern type, $F(1, 63) = 28.474$, $p < .001$, $\eta_p^2 = .311$; and vowel, $F(1, 63) = 10.938$, $p = .002$, $\eta_p^2 = .148$. There was also a significant effect of training type, $F(1, 30) = 6.127$, $p = .019$, $\eta_p^2 = .170$ with the AV group increasing in mean

percent accuracy from .78 (*SD* = .16) to .97 (*SD* = .05); the A-only group increased from .73 (*SD* = .19) to .78 (*SD* = .10). It was easier for learners to identify vowel duration when (a) the vowel was /a/ versus /u/, (b) the preceding consonant was /k/ versus /s/, and (c) when the pitch pattern was LH.H versus HL.L.

For Group III (CV.CVV), results indicated significant effects of time, $F(1, 30) = 24.083$, $p < .001$, $\eta_p^2 = .445$; vowel, $F(1, 63) = 5.154$, $p = .027$, $\eta_p^2 = .076$; and pattern type, $F(2, 126) = 5.586$, $p = .005$, $\eta_p^2 = .081$. Training type was not significant. Overall mean accuracy increased from .73 (*SD* = .17) to .91 (*SD* = .05) and was higher when (a) the vowel was /a/ versus /u/ and (b) when the pitch pattern was L.HH versus H.LL ($p < .01$). These two patterns are very distinct; L.HH starts with low pitch and remains high after the second mora; H.LL is the opposite.

Effects of Training and Stimulus Variables on Response Latency

The descriptive statistics for the RT data from the perception pre-test and post-test are shown in Table 2.

Table 2. *Descriptive Statistics for Perception Response Latency in the Pre-test and Post-test*

Group	*n*	Mean RT (*SD*) in ms	
		Pre-test	Post-test
AV	16	2856.63 (582.99)	3106.78 (530.25)
A-only	16	2805.25 (515.29)	3179.96 (564.17)
Control	16	2789.66 (410.47)	3139.41 (520.75)

As noted earlier, there was no significant difference across the groups at pre-test. As shown in Figure 5, RTs for the training groups significantly increased over time, $F(1, 45) = 10.748$, $p = .002$, $\eta_p^2 = .193$; even the control group's responses were slower. We address this issue later in the discussion.

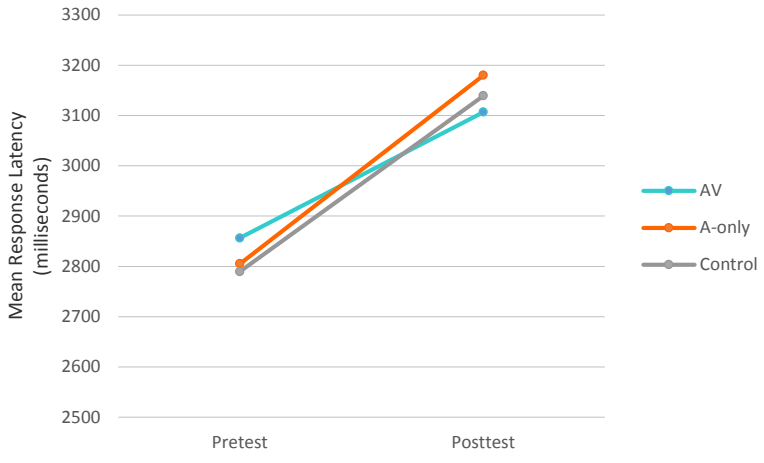

Figure 5. Mean response latency for perception at pre-test and post-test by group.

Further analysis focuses on the significant differences from ANOVAs for RTs within the pitch pattern groups shown in Figure 1. For each of the three groups, the within-group factors were time (2), consonant (2), vowel (2), and pattern type (Group I, 3 patterns; Group II, 2 patterns; Group III, 3 patterns); the between-groups factor was group type (AV, A-only, control).

For Group I (CVV.CVV), results revealed no significant main effects or interactions; however, there was a significant effect of time for Group II (CVV.CV), $F(1, 30) = 7.593$, $p = .010$, $\eta_p^2 = .202$, and Group III

(CV.CVV), $F(1, 30) = 16.515$, $p < .001$, $\eta_p^2 = .355$. For these two pattern groups, each having one long vowel, learners took more time to respond.

Analysis of Training Effects

To examine the effects of talker and stimulus factors during training, perception accuracy and RT data from the training sessions were analyzed per group. Results for accuracy are presented first followed by those for RTs. Figure 6 shows the mean perception accuracy for stimuli produced by each training talker. Generally, higher scores were found for the AV training group, which had access to waveforms as feedback during training.

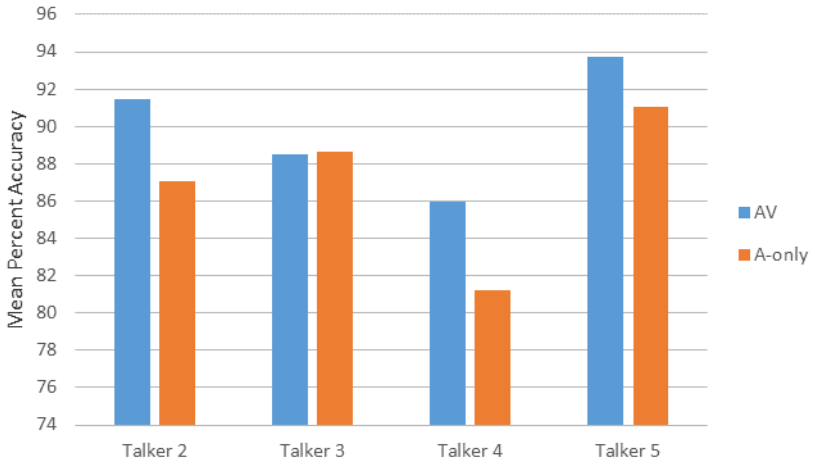

Figure 6. Perception accuracy by talker in training.

Accuracy of AV Training Group

A three-way ANOVA investigated the effects of training week (2), talker (4), and pattern type within each syllable structure group. For Group I (CVV.CVV), there was a marginally significant effect of week, $F(1, 15) = 4.444$, $p = .052$, with scores increasing in the second week. For Group II (CVV.CV), there was a significant effect of talker, $F(3, 45) = 19.056$, $p < .001$, $\eta_p^2 = .560$. Pairwise comparisons revealed that accuracy for stimuli produced by Talker 4 was significantly lower compared to the other talkers. Mean percent accuracy varied as follows: .63 (Talker 4), .82 (Talker 3), .89 (Talker 5), and .93 (Talker 2). While the precise source of the lower accuracy for Talker 4 is not known, a possible explanation appears in Figure 7, which shows the waveform and pitch track for *kaakaa* (HL.LL) produced by each training talker. Across talkers, these long vowels were of comparable duration; however, note that Talkers 2, 3, and 5 showed greater pitch movement (indicated by the blue line), especially in the initial-accented mora compared to Talker 4. Although all auditory stimuli were accurately identified by NSs prior to the study, the variability in the acoustic characteristics of voices as shown in this example may have influenced learner perception.

For Group III (CV.CVV), there was a significant effect of talker, $F(3, 45) = 4.470$, $p = .008$, $\eta_p^2 = .230$. Pairwise comparisons indicated that learners were more accurate in their perception of the duration of vowels produced by Talker 5 compared to the others although mean accuracy for stimuli in this group was relatively high, ranging from .89 (Talker 2) to .96 (Talker 5).

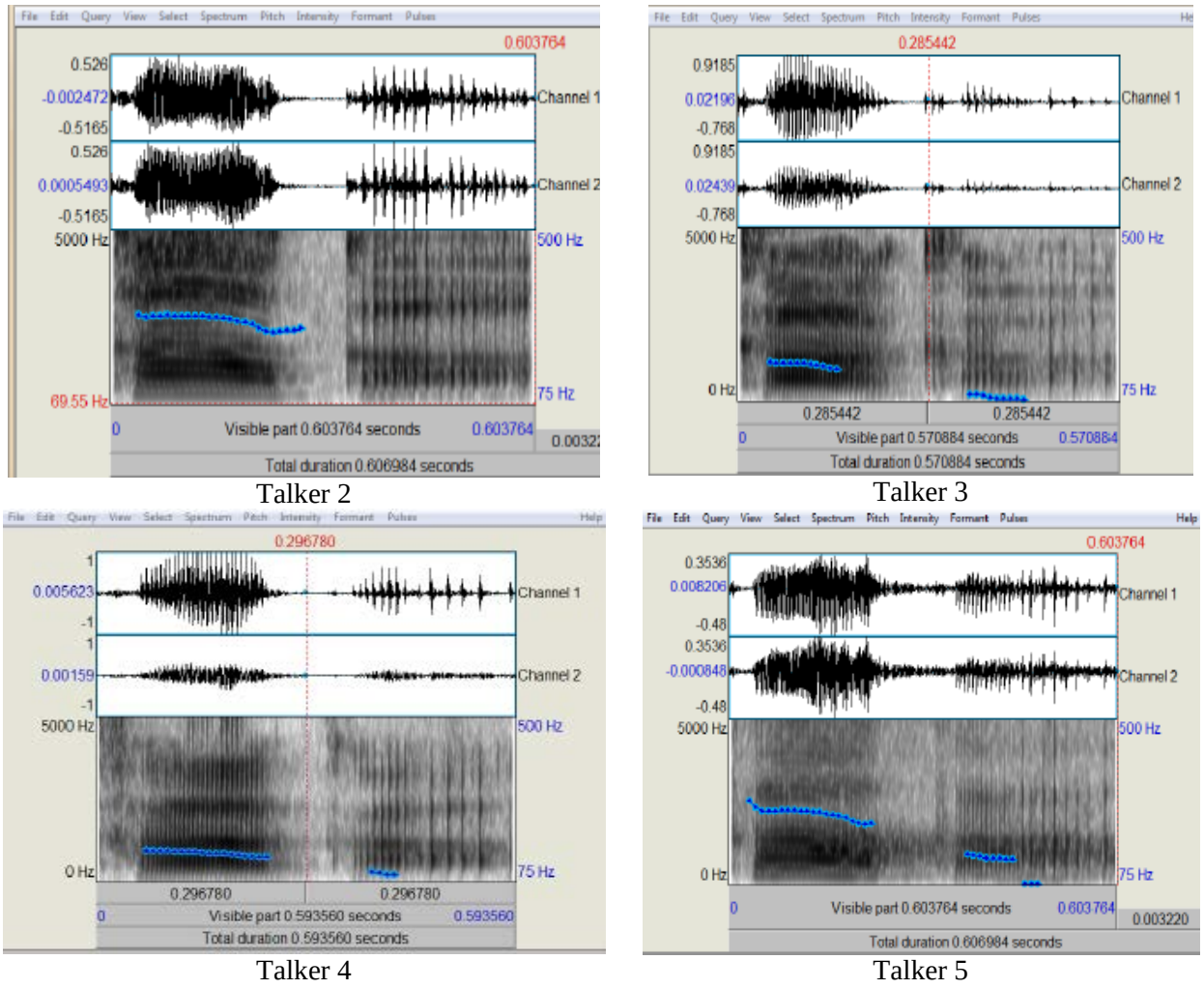

Figure 7. Talkers from training: Waveform and pitch displays for kaa.kaa (HL.LL).

Finally, for Group IV (CV.CV), there was a significant effect of week, $F(1, 15) = 6.363$, $p = .023$, $\eta_p^2 = .298$, with higher accuracy in the second week. Mean percent accuracy for all talkers was above .90.

Accuracy of A-only Training Group

A three-way ANOVA investigated the effects of training week (2), talker (4), and pattern type within each syllable structure group. For Group I (CVV.CVV), there was a significant effect of week, $F(1, 15) = 6.310$, $p = .024$, $\eta_p^2 = .296$, with mean accuracy rising from .83 to .88 by the end of the second week. There was substantial variability within the data as evidenced by a significant Week × Talker interaction, $F(3, 45) = 1.815$, $p = .049$, $\eta_p^2 = .108$. Simple effects tests pointed to lower accuracy again for Talker 4 in the first week of training.

For Group II (CVV.CV), there was a significant effect of talker, $F(3, 45) = 15.527$, $p < .001$, $\eta_p^2 = .509$, with stimuli produced by Talker 4 also showing a significantly lower percent accuracy (.58) compared to the other talkers for whom accuracy ranged from .84 to .91. There was a significantly lower accuracy for *saa.aa* (HL.L) at .65 compared to the other stimuli (accuracy of .84 to .86). For Group III (CV.CVV), there was a significant Talker × Pattern Type interaction, $F(3, 45) = 2.792$, $p = .002$, $\eta_p^2 = .157$. Simple effects tests revealed lower accuracy for *sa.saa* (L.HL) produced by Talker 4. For Group IV (CV.CV), there was a significant Talker × Pattern Type interaction, $F(3, 45) = 5.293$, $p = .003$, $\eta_p^2 = .261$; again,

vowel identification accuracy for stimuli with a H.L pattern (e.g., *ka.ka, sa.sa, ku.ku*) was lower when produced by Talker 4.

Response Latency of AV Training Group

Figure 8 shows the mean RTs for stimuli produced by the four training talkers. The AV (vs. A-only) group showed shorter RTs to stimuli produced by all the talkers. The RT data were analyzed the same way as the accuracy data. Discussion focuses on the significant findings. For Groups I (CVV.CVV) and II (CVV.CV), mean RTs were significantly shorter from Week 1 to Week 2, (e.g., $F_{Group\ I}(1, 15) = 19.363$, $p = .001$, $\eta_p^2 = .563$). For Group III (CV.CVV), there were significant effects of week, $F(1, 15) = 5.525$, $p = .033$, $\eta_p^2 = .269$, and talker, $F(3, 45) = 6.417$, $p = .001$, $\eta_p^2 = .300$. Mean RTs were shorter in the second week and longer for Talkers 3 and 5. For Group IV (CV.CV), there was a significant Week × Talker interaction, $F(3, 45) = 6.672$, $p = .001$, $\eta_p^2 = .308$. Simple effects tests found that learners' RTs significantly decreased for stimuli produced by Talker 2 from Week 1 to Week 2.

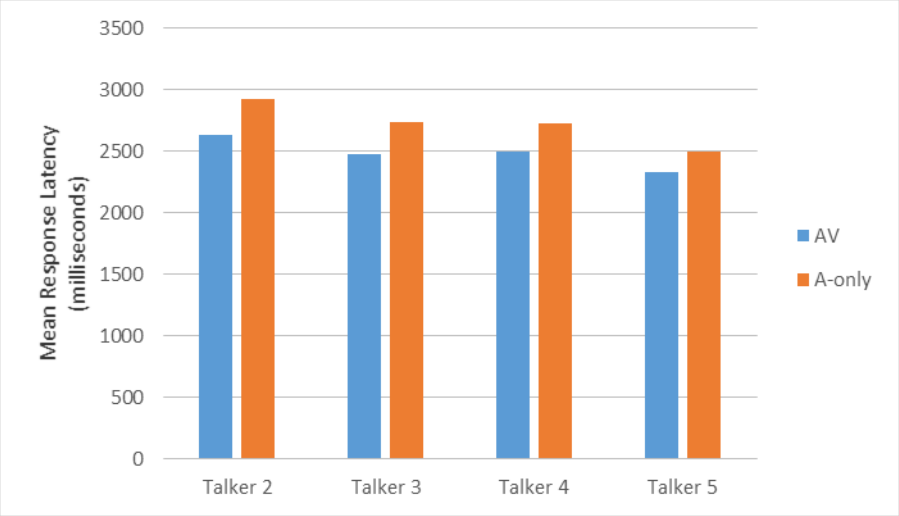

Figure 8. Mean response latency in training per talker

Response Latency of A-only Training Group

A three-way ANOVA was carried out on the RTs from the training data for the A-only group. For Group I (CVV.CVV), there was a significant effect of week, $F(1, 15) = 8.683$, $p = .010$, $\eta_p^2 = .367$, and a significant Week × Talker interaction, $F(3, 45) = 4.312$, $p = .011$, $\eta_p^2 = .216$. RTs to stimuli produced by Talker 2 were shorter in the second week. For Group II, there was a significant effect of talker, $F(3, 45) = 3.410$, $p = .025$, $\eta_p^2 = .185$. RTs to tokens produced by Talker 5 were shorter than those produced by Talker 3. For Group III (CV.CVV), there was also an effect of talker, $F(3, 45) = 7.610$, $p < .001$, $\eta_p^2 = .337$; RTs were shorter for Talker 5. Learners demonstrated shorter RTs to stimuli with the pattern L.HH versus L.HL or H.LL, especially if they involved /ku/ versus /sa/. Finally, for Group IV (CV.CV), there were significant effects of week and talker, and a significant Week × Talker interaction, $F(3, 45) = 12.816$, $p < .001$, $\eta_p^2 = .461$, with RTs decreasing in Week 2 for tokens produced by Talker 2.

Effects of Perception Training on Production

To examine whether perception training transferred to another skill, mean pre-test and post-test production accuracy ratings were provided by three raters from the Tokyo area who had a background in linguistics and experience teaching Japanese. They were instructed to consider only vowel duration in their evaluations. Raters selected what they heard from a list of options which varied according to the duration and syllable location of the vowels. Based on pilot testing, options included a geminate

consonant (e.g., kakka). For example, for the item *kaakaa*, the options were (a) kaakaa, (b) kaaka, (c) kakka, (d) kaka, and (e) other. Raters who chose (e) other were asked to write what they heard. One point was given for each correct pronunciation. Inter-rater reliability revealed a significant positive correlation between Raters 1 and 2 (r = .914, p < .001, R^2 = .84), Raters 1 and 3 (r = .930, p < .001, R^2 = .86) and Raters 2 and 3 (r = .906, p < .001, R^2 = .82). For all items, there was absolute agreement from at least two raters. Table 3 shows the descriptive statistics for each training group in the pre- and post-test.

Table 3. *Descriptive Statistics for Production Accuracy in the Pre-test and Post-test*

Group	*n*	Mean Percent Accuracy (*SD*)	
		Pre-test	Post-test
AV	16	.66 (.20)	.91 (.08)
A-only	16	.69 (.14)	.90 (.09)

The mean raters' scores were analyzed using ANOVA. Within-group factors were time (2), vowel (2), preceding consonant (2), and token type (4); the between-groups factor was training type (2). Results indicated a significant effect of time, $F(1, 30) = 67.148$, $p < .001$, $\eta_p^2 = .691$, and token type, $F(3, 90) = 5.392$, $p = .002$, $\eta_p^2 = .152$; however, vowel, $F(1, 30) = 1.815$, $p = .188$; training group, $F(1, 30) = 1.600$, $p = .216$; and preceding consonant, $F(1, 30) = .062$, $p = .806$ were not significant. Pairwise comparisons indicated that production of the token with only one long vowel in the first syllable (CVV.CV) was the most accurate.

In addition to the main effect of token type, the Time × Token Type interaction was significant, $F(3, 90) = 7.977$, $p < .001$, $\eta_p^2 = .210$. Results of simple effects tests revealed that production of CVV.CV was better than the other syllable types at pre-test (mean of .85) and improved to .97. CV.CVV and CV.CV showed comparable improvement (from a mean of .56 to .92, and .62 to .94, respectively); and CVV.CVV improved slightly (from .74 to .75). The Vowel × Token Type interaction, $F(3, 90) = 2.929$, $p = .038$, $\eta_p^2 = .089$ was also significant; production accuracy of CVV.CVV tokens was higher when the vowel was /a/.

In summary, production accuracy improved from pre-test to post-test with no significant difference between the two training groups. Three token types (CVV.CV, CV.CVV, and CV.CV) significantly improved after training, but not the CVV.CVV (two long vowels) although its accuracy was greater with the vowel /a/ versus /u/. In general, since the learners had received no specific production training or practice, results suggested that focused perception training had shown some transfer to improved production.

Tests of Generalization

Perception Accuracy

Analyses were performed to determine if training would generalize to the perception of unfamiliar stimuli spoken by a familiar talker (TG1; Appendix D), and familiar stimuli (i.e., used in testing) spoken by an unfamiliar talker (TG2). Descriptive statistics for perception accuracy in the TGs and post-test are shown in Table 4. An ANOVA was conducted with test (post-test, TG1) as the within-group variable and training type (2) as the between-groups variable. Results indicated no significant effect of test, $F(1, 30) = .438$, $p = .513$. The effect of training group, $F(1, 30) = 3.586$, $p = .068$, and the Test × Training Group interaction, $F(1, 30) = 3.800$, $p = .061$ approached significance.

To examine whether the post-test and TG2 (unfamiliar talker) were comparable, an ANOVA was conducted with test (post-test, TG2) as the within-group variable and training group type (2) as the between-groups variable. Results indicated no significant effect of test, $F(1, 30) = .786$, $p = .382$. Training group type approached significance, $F(1, 30) = 3.890$, $p = .058$; the AV group had higher scores.

The Test × Training Group interaction approached significance, $F(1, 30) = 3.610$, $p = .067$. Thus, the effects of the training showed some generalization to novel stimuli and a new talker's voice.

Table 4. *Descriptive Statistics for Perception Accuracy in the Post-test and Two TGs*

Group	*n*	Mean Percent Accuracy (*SD*)		
		Post-test	TG1 (novel tokens)	TG2 (new voice)
AV	16	.96 (.09)	.93 (.07)	.93 (.09)
A-only	16	.88 (.13)	.89 (.12)	.89 (.09)

Perception Latency

Table 5 shows descriptive statistics for the RT data. Separate ANOVAs were performed to compare the perception RT data from the post-test with those from TG1 and TG2. The within-group variable was test (2) and between-groups variable was training type (2). For TG1, results indicated a significant effect of test, $F(1, 30) = 92.711$, $p < .001$, $\eta_p^2 = .756$; RTs in TG1 were shorter. However, neither training type, $F(1, 30) = .796$, $p = .379$, nor the Test × Training Group interaction was significant, $F(1, 30) = 1.873$, $p = .181$. For TG2, results indicated significant effects of test, $F(1, 30) = 28.422$, $p < .001$, $\eta_p^2 = .486$; RTs in TG2 were shorter. However, neither training type, $F(1, 30) = 1.038$, $p = .316$ nor the Test × Training Type interaction, $F(1, 30) = .941$, $p = .340$ was significant. In summary, learners generally demonstrated shorter RTs in TGs versus the post-test; however, there were no significant differences between the two training groups.

Table 5. *Descriptive Statistics for Perception Response Latency (RT) in the Post-test and Two TGs*

Group	*n*	Mean RT (*SD*) in ms		
		Post-test	TG1 (novel tokens)	TG2 (new voice)
AV	16	3155.17 (532.95)	2435.90 (528.33)	2392.59 (571.46)
A-only	16	3241.33 (492.71)	2675.71 (477.38)	2685.66 (764.26)

GENERAL DISCUSSION

This study investigated the factors affecting the perception and production of vowel duration in L2 Japanese, and the effectiveness of waveform displays in training. Results demonstrated that perception training significantly improved overall accuracy in the identification of vowel duration. Although the AV group showed greater improvement at a faster rate, its performance was not significantly different from the A-only group. Greater accuracy was accompanied by longer RTs. Participants' comments following the current study suggest reasons for this finding. Those who received training, especially in the AV group, which saw waveforms, were faced with more information on which to base their post-test responses. Many commented that they were more confident in their post-test accuracy, which was supported by the data, but it came at a processing cost. It is possible we are seeing what transpires in the earlier stages of focused training. Consequently, if the training period had been longer, the RTs might have become shorter. The control group members indicated they were trying to do better in the post-test despite the lack of training, having felt less confident about their pre-test accuracy. Consequently, they took more time to respond hoping it would result in better accuracy, which was not the case. The training groups' RTs were also shorter in the TGs compared to the post-test. Participants in the training groups said that by the time they reached the end of the study, the tasks (i.e., the TGs) seemed easier, and they were feeling more confident.

Similar to previous training studies, results indicated a significant influence of preceding consonant, vowel type, and talker's voice in training. As hypothesized in the current study, there were also significant effects according to pitch pattern and syllable location of the long vowel or vowels. Prior to

training, learners exhibited more accurate perception for stimuli that had the pattern LH.H where there is only one long vowel in the first syllable, and the pitch remains high after it rises. The next best performance was for words with a L.HH pattern where the long vowel is in the second syllable with a consistently high pitch. These findings are compatible with English-based prosodic preferences for a rising intonation, and the salience created by a high pitch. In the post-test, there was generally less variability in accuracy across stimulus types although the best performance still tended to be for LH.HH followed by L.HH.

Consistent with other studies (e.g., Bradlow et al., 1999; Hardison, 2003), talker variability influenced perception, and perception training transferred to significant production improvement. The precise nature of the perception-production link is not known, and it is clear that not every learner follows the same path nor that these skills enjoy parallel development. However, this transfer to production improvement has been a consistent finding in studies whether the target sounds were AE /r/ and /l/ (e.g., Bradlow et al., 1999; Hardison, 2003), Japanese geminates (Motohashi-Saigo & Hardison, 2009), or vowels as in the current study.

Generalization of improvement to novel stimuli and a new voice is a valuable goal that takes skill improvement to new contexts. However, the findings should be interpreted with some caution as generalization was tested in this and other studies with a finite stimulus set and one new voice. There are some additional limitations in this study. The choice of stimuli for perception testing and training was based on the challenging syllable types and pitch patterns suggested by teaching experience and the results of pilot testing. This permitted focused training with a reasonable stimulus set and a timeframe that was appealing to participants, but may not have allowed us to see if improvement could have progressed even more. This is perhaps one advantage of allowing participants to access web-based training on their own time (e.g., Motohashi-Saigo & Hardison, 2009; Wang & Munro, 2004) although this approach involves some loss of control by the researchers. In addition, the current study's stimuli were presented in isolation. Further studies are needed to determine the effects of context on the perception and production of vowel duration. Another challenge for training studies is the issue of retention. Few studies (e.g., Bradlow et al., 1999; Wang & Munro, 2004) have reported findings on the retention of skill improvement although learners who continue to receive L2 input are well positioned to maintain their skills.

Finally, a recurring theme across L2 studies is variability. In fact, the successful auditory training studies that began 25 years ago were followed by others that often referred to adopting the high variability training paradigm. While it is important to recognize that including versus excluding variability in training likely contributes to more robust perceptual category development and the chances for generalization, this variability has been controlled. Studies generally record and present stimuli under sound-attenuated conditions where rate and style of speech are monitored, and talkers may not represent the full range of talker characteristics. This contrasts with the natural language environment where variability across multiple dimensions may impact the perceptual system simultaneously.

CONCLUSION AND PEDAGOGICAL IMPLICATIONS

Over the years, as more research has been conducted on the acquisition of nonnative sounds, more sources of variability have been found that affect learner performance. As with previous perception training studies, the current investigation on the acquisition of Japanese vowel duration found significant effects of vowel type, preceding consonant, and the talker's voice during training. To these, the current study added pitch pattern and syllable location of the vowels. Although these findings paint an even more complex picture of L2 perception and production, they help to explain variable learner performance.

Classroom teachers could use the above findings on factors affecting perceptual accuracy to guide their selection of materials for learners. To build perceptual categories that are robust to stimulus and talker variability, L2 learners need multimodal exposure to different types of speech, including a range of

talkers and phonetic contexts. Visual feedback from increasingly accessible technological tools can help to focus training on challenging features of L2 speech. It is encouraging to note the positive comments from participants in research studies toward the use of pitch displays (e.g., Hardison, 2004), spectrograms (Olson, 2014), and waveforms (Motohashi-Saigo & Hardison, 2009) in their L2 learning. In fact, some learners have asked why such tools are not a regular part of language instruction. Given the constraints on and expectations for the use of classroom time, teachers may wish to explore the use of these displays by learners on their own or in groups outside the classroom. Research suggests that learners need less instruction on the use of these tools and may benefit earlier in their learning than we had thought although some degree of guidance is necessary and must be tailored to the learner population and the purpose of training. As Olson (2014) noted, visual feedback may be more suitable for some features (e.g., duration and pitch) than others. These tools need not be viewed as opponents to a communicative focus, but as supplements to help learners benefit from focused training on challenging features.

APPENDIX A. Tokens for Production Testing

kaakaa	saasaa	kuukuu	suusuu
kaaka	saasa	kuuku	suusu
kakaa	sasaa	kukuu	susuu
kaka	sasa	kuku	susu

APPENDIX B. Stimuli Used in Perception Testing

	Pitch Pattern	Duration of First Vowel in ms	Duration of Second Vowel in ms
kaakaa	LH.HL	260	213
kaaka	LH.H	184	71
kaaka	HL.L	198	91
kakaa	L.HL	48	251
saasaa	LH.HH	239	261
saasaa	HL.LL	202	252
sasaa	L.HH	73	276
sasaa	H.LL	69	145
kuukuu	LH.HL	199	190
kuukuu	HL.LL	184	194
kuuku	LH.H	193	96
kuuku	HL.L	159	73
kukuu	H.LL	58	265
suusuu	LH.HH	216	251
suusu	LH.H	188	104
suusu	HL.L	178	89
susuu	L.HL	58	234
susuu	H.LL	52	233

Note. The period denotes a syllable boundary.

APPENDIX C. Stimuli Used in Training

	Pitch	Meaning
kaakaa	LH.HH	
kaakaa	HL.LL	
kakaa	L.HH	
kakaa	H.LL	
kaka	L.H	
kaka	H.L	flowers and fruits
saasaa	LH.HL	
saasa	LH.H	
saasa	HL.L	
sasaa	L.HL	
sasa	L.H	sake (Japanese beverage)
sasa	H.L	bamboo leaves
kuukuu	LH.HH	
kukuu	L.HH	
kukuu	L.HL	
kuku	L.H	cane
kuku	H.L	randomness
suusuu	LH.HL	
suusuu	HL.LL	
susuu	L.HH	
susu	L.H	
susu	H.L	

Note. The period denotes a syllable boundary.

APPENDIX D. Stimuli for Perception Task in TG1

	Pitch Pattern
seesee	LH.HH
seesee	HL.LL
seesee	LH.HL
seese	LH.H
seese	HL.L
sesee	L.HH
sesee	L.HL
sesee	H.LL
sese	L.H
sese	H.L
taataa	LH.HH
taataa	LH.HL
taataa	HL.LL
taata	LH.H
taata	HL.L
tataa	L.HH
tataa	L.HL

tataa	H.LL
tata	L.H
tata	H.L

Note. The period denotes a syllable boundary.

NOTES

1. Sonority is a property of a segment and refers to the degree of acoustic energy. On a continuum, voiceless stops have the least sonority and vowels have the most. The difference in sonority values between two segments may contribute to perceptual distance and facilitate the detection of boundaries. The sonority difference between a stop (e.g., /k/) and a vowel is greater than that between a fricative (e.g., /s/) and a vowel.
2. Effect sizes are reported as partial eta-squared (η_p^2 = .01 = small, .06 = medium, .14 = large; Cohen, 1988).

ACKNOWLEDGMENTS

Portions of this study were presented at the meeting of the American Association for Applied Linguistics in Portland, Oregon in March, 2014.

ABOUT THE AUTHORS

Tomoko Okuno received her Ph.D. in Second Language Studies from Michigan State University and is now lecturer and director of the Japanese Program in the Residential College at the University of Michigan. Her research interests include L2 perception and production, L2 phonology, Japanese linguistics, and Japanese pedagogy.

E-mail: okuno@umich.edu

Debra Hardison (Ph.D. Indiana University) is associate professor in the TESOL and Second Language Studies programs at Michigan State University. Her research focuses on auditory-visual integration in spoken language processing, learner variables in speech production, and the applications of technology in perception and production training involving segmental and suprasegmental features.

E-mail: hardiso2@msu.edu

REFERENCES

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer [Computer program]. Retrieved from http://www.fon.hum.uva.nl/praat/

Bradlow, A., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*, 977–985.

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories. *Applied Psycholinguistics*, *32*, 51–67.

Carey, M. (2004). CALL visual feedback for pronunciation of vowels: Kay Sona-Match. *CALICO Journal*, *21*, 571–601.

Chun, D. M. (1998). Signal analysis software for teaching discourse intonation. *Language Learning & Technology*, *2*(1), 61–77. Retrieved from http://llt.msu.edu/vol2num1/article4/

Chun, D. M. (2002). *Discourse intonation in L2: From theory and research to practice.* Amsterdam, Netherlands: Benjamins.

Chun, D. M., Hardison, D. M., & Pennington, M. C. (2008). Technologies for prosody in context: Past and future of L2 research and practice. In J. H. Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 323–346). Amsterdam, Netherlands: Benjamins.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

de Bot, K. (1983). Visual feedback of intonation: Effectiveness and induced practice behavior. *Language and Speech*, *26*, 331–350.

Haraguchi, S. (1999). Accent. In N. Tsujimura (Ed.), *The handbook of Japanese linguistics* (pp. 1–30). Malden, MA: Blackwell.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context and talker variability. *Applied Psycholinguistics*, *24*, 495–522.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training. Quantitative and qualitative findings. *Language Learning & Technology*, *8*(1), 34–52. Retrieved from http://llt.msu.edu/vol8num1/hardison/default.html

Hardison, D. M. (2005a). Contextualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input. *CALICO Journal*, *22*, 175–190.

Hardison, D. M. (2005b). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistic*s, *26*, 579–596.

Hardison, D. M. (2012). Second language speech perception: A cross-disciplinary perspective on challenges and accomplishments. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 349–363). London, UK: Routledge.

Hardison, D. M., & Motohashi-Saigo, M. (2010). Development of perception of second language Japanese geminates: Role of duration, sonority, and segmentation strategy. *Applied Psycholinguistics*, *31*, 81–99.

Hincks, R., & Edlund, J. (2009). Promoting decreased pitch variation in oral presentations with transient visual feedback. *Language Learning & Technology*, *13*(3), 32–50. Retrieved from http://llt.msu.edu/vol13num3/hincksedlund.pdf

Hirata, Y., & Kelly, S. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, *53*, 298–310.

Kipp. M. (2001). Anvil–A generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology* (pp. 1367–1370). Aalborg, Denmark: Eurospeech.

Koguma, R. (2000). Perception of Japanese short and long vowels by English-speaking learners. *Current Report on Japanese-Language Education around the Globe*, *10*, 43–55.

Kubozono, H. (1999). Mora and syllable. In N. Tsujimura (Ed.), *The handbook of Japanese linguistics* (pp. 31–61). Malden, MA: Blackwell.

Leather, J. (1990). Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In J. Leather & A. James (Eds.), *New Sounds 90* (pp. 72–97). Amsterdam, Netherlands: University of Amsterdam.

Levis, J. M., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, *32*, 505–524.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242–1255.

Minagawa, Y. (1997). Sokuon no shikibetsu niokeru accent-gaga to shiinshu no yooin: kankoku, tai, chuugoku, ei, seigo bogowasha no baai. [Accent patterns and segment places as a factor for perceiving Japanese long and short vowels by native speakers of Korean, Thai, Chinese, English, and Spanish]. In *Proceedings of the Spring Meeting of the Society for Teaching Japanese as a Foreign Language* (pp. 123–128).

Motohashi-Saigo, M., & Hardison, D. M. (2009). Acquisition of L2 Japanese geminates: Training with waveform displays. *Language Learning & Technology*, *13*(2), 29–47. Retrieved from http://llt.msu.edu/vol13num2/motohashisaigohardison.pdf

Olson, D. J. (2014). Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning & Technology*, *18*(3), 173–192. Retrieved from http://llt.msu.edu/issues/october2014/olson.pdf

Pennington, M. C., & Esling, J. H. (1996). Computer-assisted development of spoken language skills. In M. C. Pennington (Ed.), *The power of CALL* (pp. 153–189). Houston, TX: Athelstan.

Shibatani, M. (1990). *The languages of Japan*. New York: Cambridge University Press.

Tsujimura, N. (2007). *An introduction to Japanese linguistics* (2nd ed.). Malden, MA: Blackwell Publishing.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, *32*, 539–552.

Yoshida, Y. A. (2006). Accents in Tokyo and Kyoto Japanese vowel quality in terms of duration and licensing potency. *SOAS Working Papers in Linguistics*, *14*, 249–264.