ACQUISITION OF L2 VOWEL DURATION IN JAPANESE BY NATIVE ENGLISH SPEAKERS

By

Tomoko Okuno

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – DOCTOR OF PHILOSOPHY

2013

ABSTRACT

ACQUISITION OF L2 VOWEL DURATION IN JAPANESE
BY NATIVE ENGLISH SPEAKERS

By

Tomoko Okuno

Research has demonstrated that focused perceptual training facilitates L2 learners' segmental perception and spoken word identification. Hardison (2003) and Motohashi-Saigo and Hardison (2009) found benefits of visual cues in the training for acquisition of L2 contrasts. The present study examined factors affecting perception and production of vowel duration (i.e., long versus short) in Japanese and benefits of waveform displays as visual cues on the acquisition of vowel duration in L2 Japanese by native speakers of L1 English, and transfer to production. Vowel length in Japanese is a contrastive feature, important for communication, and a challenge for many L2 learners.

A pretest-posttest design with controls was used. A between-subject variable was training type: auditory visual (AV), auditory-only (A-only), and no training (controls). Within-subject variables were vowel type, preceding consonant, and pitch pattern. Participants were 64 learners of Japanese whose L1 was American English. Testing and training materials were 40 bisyllabic-words containing long and short vowels. To create the stimuli, two Japanese vowels (/a, u/), two consonants (/k, s/), and 10 pitch patterns were selected. The stimuli, produced by six NSs of Japanese, were recorded.

Production and perception pre- and post-tests were administered to assess the effects of training on perception accuracy and reaction time (RT). During production testing, participants produced 16 bisyllabic words in isolation. For perception testing, they completed a forced-

choice, four-alternative identification task for 18 stimuli, the bisyllabic words. Perception training, conducted between the pre- and post-tests, involved eight sessions, each 25 minutes; the participants also completed the same identification task, using a computer. During training, feedback was provided on both correct and incorrect responses; immediately after the choice, correct words appeared on the screen.

Results indicated significant improvement on identification accuracy for both groups, but the rate of improvement of the AV group was greater. On the other hand, RTs of the two groups became slower after the training. In addition, it was found that vowel type, preceding consonant, and pitch patterns in addition to the talker's voice in the training together affected L2 learners' perception of vowel duration. The results suggested that the learners' stages of L2 perceptual development involve the evaluation of input based on context- and talker-dependent perceptual categories.

To my family

ACKNOWLEDGMENTS

I could not complete this dissertation without help, support, and encouragement from many people. I would like to show my deepest appreciation to Dr. Debra Hardison, the co-chair of my dissertation committee, for her support and guidance and feedback on the proposal, draft, and statistical analysis. In addition, she helped me to revise the dissertation by giving me specific comments and feedback. I think that I learned a lot through working on and revising this dissertation. I would like to extend my appreciation to Dr. Mutsuko Endo Hudson, the other co-chair, for her support and feedback on my dissertation and giving me opportunities to expand my teaching career. In addition, I would like to thank to Dr. Susan Gass, and Dr. Yen-Hwei Lin, the dissertation committee, for patiently reading my dissertation and giving me helpful feedback.

I would like to thank the Department of Linguistics and Languages and Second Language Studies Program for giving me a lot of support, funding, and opportunities to teach Japanese courses as an instructor.

Students who were taking Japanese language courses at Michigan State University were very supportive. They were motivated to learn Japanese and to improve their Japanese language, and they participated in my study. I would like to thank them for their participation and encouragement to me to finish my dissertation.

I thank my dissertation working group with Soo Hyon Kim and Baburhan Uzum. We met twice a week and wrote our dissertations. It was a very good way to motivate myself to continue to work on my dissertation and to give encouragement to each other. I am very happy to graduate with Soo Hyon and Baburhan.

Finally, I would like to thank to my family and friends at Michigan State University. My family, including my two cats, Moggy and Syllable, always emotionally supported me since I came to Michigan State University. Also, I want to thank my friends, including Masae Yasuda, Nao Nakano, Misako Matsubara, Chien Hsiung (Scott) Chiu, Nobuhiro Kamiya, Kanako Kamiya, Tsuyoshi Oshita, Junkyuu Lee, Shaofeng Lee, Seongmee Ahn, Marthe Russell, Grace Lee Amuzie, Solène Inceoglu, Jimin Kahng, and Chiung Wang.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND REVIEW OF THE LITERATURE

Introduction

Second language (L2) learners have difficulties in perceiving and producing new L2 contrasts once they have established a phonological system for their first language (L1) (e.g., Archibald, 2005; Flege, 1995). The learners need to modify the existing system or establish a new one in order to be able to perceive or produce a new contrast in the L2, such as the contrast between English /l/ and /r/ for Japanese and Korean native speakers (NSs) (e.g., Ingram & Park, 1998). One of the common cases that L2 learners of Japanese encounter is acquisition of durational contrasts (e.g., Asano, 2005; Enomoto, 1992; Hirata, 1990; Hirata & Kelly, 2010; Minagawa, 1997; Motohashi, 2007; Motohashi-Saigo & Hardison, 2009; Toda, 1998, 2003, and 2009). For English native speakers (NSs), acquiring the contrasts between geminates and singletons as well as long vowels and short vowels in Japanese is a challenge. According to Toda's (2009) study, L2 learners experienced communication breakdown due to the failure of correctly identifying or pronouncing the durational contrasts. Thus, it is important to acquire L2 durational contrasts for communication.

In order to help the acquisition of L2 contrasts, several researchers have examined and found the effectiveness of focused perceptual training for the acquisition of L2 phonetic contrasts or segmental perception, using auditory-only (A-only) training (e.g., Borden, Gerber, & Milsark, 1983; Bradlow & Pisoni, 1999; Ingram & Park, 1998; Jamieson & Moroson, 1986; Lively, Logan & Pisoni, 1993; Logan, Lively, & Pisoni, 1991; McCandliss, Fiez, Protopapas & Conway, 2002; Morosan & Jamieson, 1989; Sheldon, 1985; Sheldon & Strange, 1982; and Strange & Dittman, 1984) and auditory-visual (AV) training (e.g., Hardison, 1999, 2003, 2005a, 2005b).

Other studies have paired waveforms with auditory information to train durational contrasts (e.g., Motohashi, 2007; Motohashi-Saigo & Hardison, 2009) or hand gestures (Hirata & Kelly, 2010). The studies in the previous literature suggest that training on durational contrasts is easier than spectrographic contrasts (Bohn, 1995). In addition, auditory as well as visual information helped L2 learners to improve their correct identification of L2 contrasts in the training. The bimodal training was particularly effective for the phonologically challenging segments based on the learners' L1 (Hardison, 2003).

The current project investigated the factors affecting acquisition of L2 durational contrasts and how perceptual training can contribute to it. Specifically, the focus was the factors affecting identification accuracy of vowel duration in Japanese by L1 American English learners. In order to investigate the issue, four factors, including vowel type, pitch pattern, preceding consonant, and learners' L2 proficiency, were treated as independent variables in Experiment 1. Experiment 2 then examined the effectiveness of two weeks of focused perceptual training using AV versus A-only input, in order to improve L2 learners' correct identification of L2 vowel duration. Visual input was a waveform display. Participants were English NSs who were studying Japanese as a foreign language in the U.S. The study examined the effectiveness of input type on identification accuracy and response time before and after the training in order to see how the training affected perceptual development of L2 vowel duration.

Review of the Literature

A Model for Speech Perception

It has been reported in the previous literature and in the foreign language classrooms that durational contrasts in Japanese are difficult for L2 learners, particularly for English NSs. Flege

(1995) proposed a model for speech perception and production, called the Speech Learning Model (SLM) to suggest why nonnative contrasts cause challenges for learners.

The SLM predicts two kinds of difficulties in acquiring L2 contrasts. First, it is argued that it is difficult to acquire novel L2 contrasts such as English /l/ and /r/ for Japanese and Korean learners (e.g., Aoyama, Flege, Guion, Akanahe-Yamada, & Yamada, 2004, Flege, 1995; Ingvalson, McClelland, & Holt, 2011). For instance, Japanese has only one liquid which is perceptually more similar to the flap in English (Price, 1981). Therefore, in order to acquire the novel contrast, it is necessary to create two new categories for English liquids to distinguish the contrast between /l/ and /r/ (e.g., Ingram & Park, 1998; Lively, Logan, & Pisoni, 1993; Sekiyama & Tohkura, 1993; Takagi, 1993).

Flege (1995) also claims that it is difficult to acquire 'similar L2 contrasts' (i.e., two segments that are contrastive in the L2, but not in L1). For example, the contrast between English /i/ and /ɪ/ is difficult to acquire for Italian NSs because the L1 has only /i/ in its phonological system (Flege & MacKay, 2004). This second category of difficulty described by Flege can be found when NSs of English acquire L2 Japanese durational contrasts, including the contrast between a geminate and a singleton consonant, as well as the contrast between a long and a short vowel. The durational contrasts are contrastive in Japanese (Shibatani, 1990; Kubozono, 1999b), but not in English. For example, Motohashi (1997) showed that English NSs have difficulties in acquiring the durational contrast between geminates as in Table 1.

Table 1: Examples of words with geminates and singletons (Motohashi, 2007)

Words with Geminates			Words with Singletons		
geminates	Japanese	English Gloss	singletons	Japanese	English Gloss
kk	*ka**kk**o*	parenthesis	*k*	*ka**k**o*	past
tt	*ko**tt**oo*	antique	*t*	*ko**t**oo*	isolated-island
ss	*sa**ss**u*	to infer	*s*	*sa**s**u*	to bite

In addition, Asano (2005) reported that distinguishing vowel duration (i.e., long and short vowels) such as *ojiisan* 'grandfather' and *ojisan* 'uncle' is difficult for native English speakers. The distinction between a short and long vowel is not contrastive in L1.

Status of Vowel Duration in Japanese: Issue of the Mora

Another factor involved in the difficulty of acquiring L2 durational contrasts in Japanese is the role of the mora as a unit of timing. English is a stress-timed language and employs a syllable as a basic unit for timing (Pennington, 1996). Stressed vowels have longer duration than unstressed vowels when they are spoken in isolation; unstressed vowels go through the process of lenition and are reduced to schwa (Hayes, Kirchner, & Steriade, 2004). Thus, a key factor determining the length of vowels in English is whether stress falls on the vowel or not. Vowels also tend to lengthen before a voiced consonant. In Japanese, on the other hand, word stress is not a key to determining the length of vowels; neighboring moraic units tend to show equal duration (Port, Dalby, & O'Dell, 1987). The mora in Japanese is the key unit of timing (e.g., Kubozono, 1999a; Tsujimura, 2007). Following Hayes (1989), Figure 1 shows one way to represent the phonological structures of a geminate, singleton, and long vowel (Hardison & Motohashi, 2010, p. 82) incorporating both the moraic and syllabic levels of representation.

Figure (1a), (1b), and (1c) represent a geminate consonant (tt), a singleton consonant (t), and a long vowel (ii) respectively.

Figure 1: One type of phonological structure for geminates, singletons, and long vowels (σ: syllable; μ: mora)

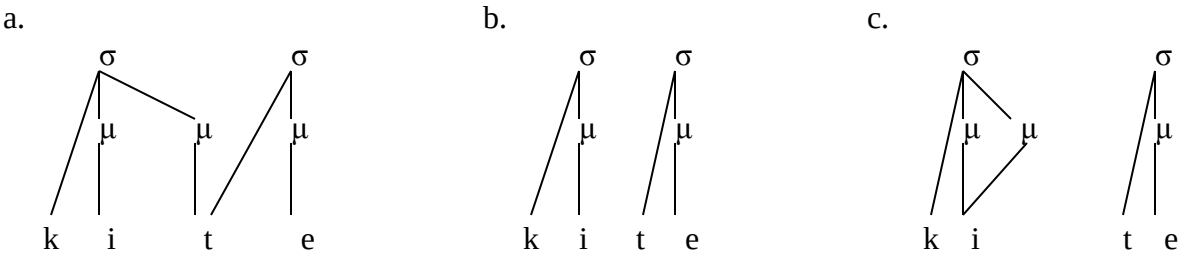


Figure (1a) is a phonological structure of the word *kitte* 'a stamp;' it has two syllables, but /t/ in the coda position of the first syllable also forms the onset of the second syllable and there are three morae. On the other hand, Figure (1b) is the structure of the word *kite* 'coming;' it also has two syllables, but only two morae as /t/ only forms the onset of the second syllable. Figure (1c) is the structure of the word *kiite* 'listening;' the vowel in the nucleus position constitutes two morae. The difference in the basic units of timing, a syllable versus a mora, may contribute to the difficulty English NSs have in acquiring durational contrasts in Japanese. A number of research studies investigated factors affecting perception of these moraic units by both NSs and NNSs.

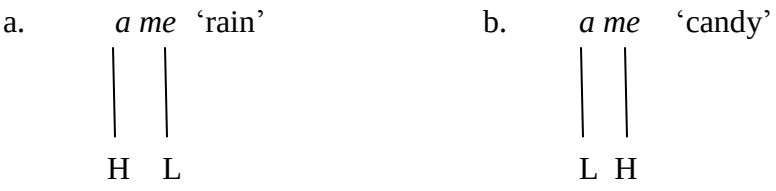
Factors Affecting Perception of Segment Length for NSs

Regarding Japanese NSs' perception, Fujisaki, Nakamura, and Imoto (1973, cited in Toda, 2003) found that the actual length of the special morae (i.e., morae that consist of a

geminate, Figure 1a, a long vowel, Figure 1c, and moraic nasal such as *hoN.ya* 'a bookstore') plays an important role. The special morae have one syllable but two morae and perception of special morae is categorical, not continuous. Fujisaki and Sugifuji (1977) examined the Japanese NSs' perception of geminates using synthesized stimuli where the closure duration of a stop consonant was manipulated. The NSs were asked to discriminate between geminates and singletons. They found that the closure duration was a key for the NSs to correctly discriminate the two segments.

In addition to the duration itself, other studies (e.g., Nagano-Madsen, 1992; Ofuka, 2003) found that other factors such as pitch accent patterns can affect NSs' perception. In Japanese, each mora receives either High (H) or Low (L) pitch as in Figure 2, and pitch accent is contrastive (Shibatani, 1990).

Figure 2: Pitch assignment in the Tokyo dialect



In Figure (2a), the word *ame* has a HL pitch pattern, which means 'rain' in the Tokyo dialect. On the other hand, in Figure (2b), the same segmental sequence *ame* has a LH pitch pattern, and means 'candy.' Ofuka (2003) investigated how different pitch accents, HL or LH, affected Japanese NSs' perception of geminates and singletons. She manipulated the closure duration of the stop /t/ in words such as *katta* [HL(L)] 'won' and *katta* [LH(H)] "bought" to create geminates and *kata* (HL) 'shoulder' and *kata* (LH) 'pattern' to create singletons. Her findings

demonstrated that the NSs needed a longer closure duration for a word with LH(H) to be perceived as a geminate, compared to a word with a HL(L) pattern.

Factors Affecting Perception of Segment Length for NNSs

Regarding NNSs' perception, Toda (1998), Enomoto (1992), Hardison and Motohashi-Saigo (2010) reported that L2 proficiency can affect perception of the durational contrast. In her study of English NSs' perception of Japanese vowel duration, Toda found that the NSs and beginning level learners required a different duration for a vowel to be judged as long. Enomoto found that the advanced learners of Japanese showed similar perceptual boundaries for geminates and long vowels as Japanese NSs; however, the beginning learners did not. Thus, perception of durational contrasts may progress along with overall L2 language proficiency. Hardison and Motohashi-Saigo's findings also concluded that correct identification of geminates with three different consonants (i.e., /t/, /k/, and /s/) was affected by learners' proficiency. For beginners, segmental duration significantly affected the identification of all types of geminates. Yet for low-intermediate and advanced learners, geminates with /s/, particularly geminates with /s/ followed by the vowel /u/[1], were significantly more difficult to identify than others.

In addition to proficiency, pitch-accent pattern and position in a word affect perception of vowel duration. Minagawa (1997) investigated whether pitch patterns (HH, LL, HL, and LH) affected the perception of vowel duration for L2 learners whose L1s were Korean, Chinese, English, Spanish, and Thai, and found that they (1) had greater perception accuracy of long vowels when a word had a high-high (HH) pitch pattern, and (2) showed a tendency to perceive a long vowel as a short vowel when a word had a low-low (LL) pitch pattern. Koguma (2000)

[1] In this dissertation, /u/ is used as a typographical convention, but the Japanese vowel is [ɯ̟̥].

investigated how L2 learners (L1 English) perceived long vowels in various positions in a word (i.e., word-initial, word-medial, and word-final) and found that word-final position was the most difficult and word-initial position was the easiest.

L2 Research with a Focus on Training Studies Involving Spectral Differences

In the literature, several perceptual training studies were conducted in order to examine whether training helps L2 learners to develop their ability to correctly perceive L2 contrasts (e.g., Bradlow & Pisoni, 1999; Ingram & Park, 1998; Lively, Logan & Pisoni, 1993; Logan, Lively, & Pisoni, 1991; Sheldon, 1985; and Strange & Dittman, 1984). Successful perceptual training studies have reported the improvement of correct identification accuracy of L2 contrasts. The first successful perceptual training reported in the literature was a study by Logan et al. (1991). They conducted training for three weeks (i.e., a total of approximately 7.5 hours) to train L1 Japanese learners of L2 English to correctly identify /l/ and /r/. They found significant improvement in ESL learners' identification of the sounds.

Following the study by Logan et al. (1991), subsequent studies by Pisoni and colleagues demonstrated the facilitative effects of auditory perceptual training (Lively et al., 1993; Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). Lively et al. (1993) reported that their perceptual training of L2 learners facilitated correct segmental identification of English /l/ and /r/ and suggested the benefits of having training stimuli produced by multiple talkers. In addition, Lively et al. (1993) found that the effects of perceptual training could be retained for three months in a setting where English was a foreign language. Bradlow et al. (1999) examined whether the facilitative effects of perceptual training could be retained and transferred to production. They found that perceptual training enhanced correct identification of English /l/

and /r/ as well as improved production of the segments without explicit production training. Additionally, they discovered that development in both perception and production was retained even after three months.

The above perceptual training studies suggested the factors necessary to make perceptual training successful. First, they emphasize that an identification task (vs. a discrimination task) with stimuli containing high variability should be used because the identification task promotes learners' classification of the target sounds into appropriate categories. Logan et al. also used different phonetic environments (i.e., different positions in a word such as initial/final clusters and singletons) so that learners were exposed to a full range of cues and the development of robust perceptual categories was enhanced. In addition, it is important to provide immediate feedback during training because it can "enhance the learning process by allowing observations of within-category similarities and between-category distinctions across contexts and talkers" (Hardison, 2003, p. 515).

L2 Research with a Focus on Training Studies Involving Exaggerated Stimuli

Although most of the training studies employed /l/ and /r/ as the targets for training, there are a few studies with other approaches, including the use of exaggerated acoustic cues. Jamieson and Morosan (1989) conducted short perceptual training, including two training sessions lasting 90 minutes with voiced and voiceless interdental fricatives using natural and synthetic stimuli. Synthetic stimuli were created by exaggerating the amount of frication. Their results indicated that (1) identification accuracy improved and (2) training with exaggerated stimuli generalized to natural speech as well as a new talker. One of the limitations of their study was the failure of training using the word-initial position to generalize to other positions in

the word such as word medial or final positions. Also, the results did not generalize to improved performance with the [ð] and [d] confusion.

The efficacy of exaggerated cues was suggested by Kuhl, Andruski, Chistovich, Chistovich, Kozhevnikova, Ryskina, Stolyarova, Sundberg, and Lacerda (1997). Mothers who were NSs of English, Russian, and Swedish talked to their infants using hyperarticulated vowels (/i/, /a/, /u/) in contrast to vowels in their speech to other adults. Kuhl et al.'s results may have implications for L2 acquisition: hyperarticulated input may be adopted at the beginning of learning an L2 so that it is easier to draw learners' attention to the critical features in the input. However, it is also important to give learners natural speech as input because they have to deal with natural speech in communication. Therefore, hyperarticulated input should be changed to natural speech over time. Uther, Knoll, and Burnham (2007) also found that female speakers of Southern British English showed hyperarticulation of vowels in infant-directed speech as well as speech directed to adult nonnative speakers of English compared to other adult English speakers.

McCandliss et al. (2002) examined the effectiveness of modified speech in the development of perception of L2 contrasts between English /l/ and /r/ for L1 Japanese learners of L2 English. They compared adaptive (i.e., exaggerated input; F3 of /l/ and /r/ are exaggerated) and fixed (i.e., natural input) training for L2 learners with and without feedback for perception of English /l/ and /r/. Results indicated that the most effective training condition was natural input with feedback; exaggerated cues were not necessary. However, they did not examine the effects of neighboring vowels on the segments. In addition, their perceptual training involved self-controlled sessions; therefore, it is unknown how much the participants paid attention to the stimuli during the training or how they carried out the training.

L2 Research with a Focus on AV Training Studies

In addition to auditory perceptual training, a few researchers examined auditory-visual (AV) training on the development of L2 perception (e.g., Hardison, 2003; Hirata & Kelly, 2010; Motohashi, 2007; Motohashi-Saigo & Hardison, 2009). Different from unimodal language input such as listening to speech sounds, bimodal input involves auditory information as well as visual cues such as facial cues and/or hand gestures, which can be additional resources for the learners to identify contrasts. Hardison (2003) compared two types of perceptual training (i.e., AV using articulatory gestures with auditory information, and A-only) on the identification of L2 English /r/ and /l/ by NSs of Japanese and Korean. She found that both training types brought improvement in identification accuracy; however, the AV training provided significantly greater improvement. Based on the study, visual input facilitated perception of the segments "in the most challenging phonetic environments for each L1 group" (p. 514). In addition, she also discovered that production of /l/ and /r/ improved significantly as a result of perceptual training. Thus, similar to the successful A-only studies described earlier (e.g., Logan et al., 1991), the effects of AV training can also be transferred to other skills. In this way, Hardison has shown the advantage of bimodal input (i.e. audio-visual input) over unimodal input in identifying different L2 consonants such as /l/ and /r/.

Motohashi-Saigo and Hardison (2009) also examined the effects of visual input on the acquisition of Japanese durational contrasts. They used waveform displays along with the auditory information in AV training, compared it with A-only training, and examined how the visual cues helped the development of correct identification of Japanese geminate consonants by NSs of English. They found that learners with AV training improved identification accuracy

significantly, generalized to novel stimuli, and transferred to production skill improvement. There were significant advantages of AV training with waveforms over A-only training.

Hirata and Kelly (2012) also investigated the effect of multimodal information on the perception of vowel durations in Japanese by NSs of English. The perceptual training (4 sessions, 120 minutes total) included four types of input: "A-only" (audio with visual image of speaker with no movement), audio with lip movements, audio with hand gestures, audio with hand gestures and lip movements. During the training, non-words were embedded in carrier sentences, and produced at a slower pace by four different talkers. The researchers used identification tasks in both testing and training. The participants listened to the input and decided whether the second vowel in each target word was short or long. The results showed that there were statistically significant effects of training so that the participants improved their ability to identify vowel duration after the training. The audio with lip movement condition was significantly better than A-only. The authors concluded that mouth movements were beneficial, but the hand gestures had not helped perceptual learning. There are several methodological issues with this study: a) participants were not learners of Japanese and had never been exposed to the language so that it was difficult to compare or generalize their results/findings with other studies involving learners of the target language, b) several stimulus factors such as rate of speech, voice, and varying context of carrier sentence were not treated as variables, c) the hand gesture involved the type of stroke associated with the given vowel duration and the hand's location in the speaker's gesture space, which were not markedly different between the short and long vowels, d) training involved four sessions, and e) pre- and post-test data were based only on auditory information.

Okuno (2009) investigated the most effective training type for the correct identification of L2 vowel duration (i.e., long and short vowels) in Japanese, using four different types of perceptual training (i.e., AV and A-only training with hyperarticulated or natural speech). Participants were 29 learners of Japanese as a FL (L1 English) at the beginning level. AV input was a speaker's face. The learners took a total of eight training sessions. In order to examine the efficacy of the training, perception accuracy scores before and after training were compared. The results indicated that all the learners improved in identification accuracy after the training; however, no advantage was found for hyperarticulated speech over natural speech. One of the possible explanations for the finding is that the study did not involve perceptual fading moving from exaggerated speech to normal speech that other studies such as Morosan and Jamieson (1989) had incorporated. Since the participants were not presented with graduated stimuli from exaggerated to natural, they did not adjust their skills to correctly identify different lengths of vowels in natural speech. In addition, the pretest scores may have reflected a ceiling effect. Therefore, it was difficult to conclude whether hyperarticulation was effective for the development of correct identification of L2 durational contrasts.

Exemplar-Based Model

The L2 learners' performance in the previous studies that investigated the effects of perceptual training (e.g., Logan et al., 1991; Hardison, 2003) was affected significantly by the context in which the contrasts were embedded and talker variables. Findings in Hardison's (2003) studies revealed that "the context- and talker-dependent nature of speech processing support the view that sources of variability or complexities in the speech signal are not merely noise discarded from the signal during processing, but are a part of subsequent neural

representations" (p. 515). Perceptual training which provides the learners with multiple exemplars in visual and/or speech input and feedback can enhance development of identifying L2 contrasts.

Research Questions and Hypotheses

To sum up, the success of auditory and auditory-visual training for correct identification of L2 segments has been established in the literature. Lively et al. (1993) concluded that training should include stimulus variability, multiple talkers, identification tasks, and feedback in order to develop robust perceptual categories. L2 learners have shown variable performance according to phonetic environment and talker. This indicates that the learners use context- and talker-dependent exemplars. Most of the previous investigations have paid closer attention to the perception of consonants in the L2, including /l/ and /r/ or /θ/ and /s/, as a focus of training. On the other hand, few studies have focused on the effects of perceptual training on vowel identification. Except for Hirata and Kelly (2010), no study has yet reported the effects of training on vowel duration, which is a contrastive feature in Japanese, important for communication, and a challenge for many L2 learners. Learners need to modify their perceptual system to perceive vowel duration accurately in the L2. Perceptual training can provide focused, identifiable input, which can shift their attention to relevant cues. The shift could, in turn, promote a reorganization of perceptual distances in psychophysical space (Hardison, 2003). By examining the efficacy of perceptual training on the identification of vowel duration and the possibility of reorganizing perceptual distances, the present study seeks to fill a gap in the previous literature.

This project investigates the effects of visual cues on the acquisition of vowel duration in L2 Japanese by English NSs. Following Motohashi-Saigo and Hardison (2009), waveforms were used as visual cues because they contain visual information on vowel duration. Also, pseudo words (i.e., words that can be pronounced but do not have any meanings) were used in order to avoid effects of neighborhood density, word frequency, and size of vocabulary. Previous psycholinguistic research (e.g., Bundgaard-Nielsen, Best, & Tyler, 20011; Imai, Walley, & Flege, 2005; Metsala, 1997; Ziegler, Muneaux, & Grainger, 2003) has shown that neighborhood density and a learner's size of vocabulary significantly affected word recognition and determination of the phonological contrasts. For measurement, in addition to accuracy of perception and production, reaction times (RTs) were measured when L2 learners identified vowel duration both in testing and training. The proposed study is designed to investigate the following five main research questions.

Research Question1: What factors affect perception accuracy, perception latency, and production accuracy of vowel duration in L2 Japanese?

Hypothesis 1a: Based on Minagawa (1997), I hypothesized that pitch pattern could affect the perception of vowel duration. In Minagawa's study, it was easier to identify long vowels with a HH pitch pattern and short vowels with a LL pitch pattern. Thus, tokens with the high pitch pattern would have higher accuracy and shorter RT than the low pitch or falling pitch (HL) if the pitch height is a key for L1 English learners.

Hypothesis1b: Regarding the types of vowels, high vowels such as /u/ have shorter duration than the low vowel /a/ in the Tokyo dialect. The duration of the long vowel /u/ could be very close to that of the short vowel /a/. As a result, NNSs may demonstrate difficulties in determining the correct identification of vowel duration for the high vowels. Thus, I hypothesized that the type of vowel could affect NNSs' perception of vowel duration, and identification accuracy and RT of the low vowel would be higher than that of the low vowel.

Hypothesis 1c: Based on Hardison and Motohashi-Saigo (2010), I hypothesized that proficiency would affect the identification of long vowels. In this study, pseudo-words were used in order to remove possible influences of vocabulary size, word familiarity, and neighborhood density. Therefore, the ability to correctly identify the durational contrast could be related to the length and overall L2 proficiency. Thus, it was predicted that identification accuracy would be higher and RT would be shorter if the learners' proficiency was higher.

Research Question 2: Is focused perceptual training effective for the acquisition of vowel duration? How do perceptual accuracy and RT vary across the period of training? Do they vary according to talker and/or other stimulus factors?

Hypothesis 2a: Based on the previous training studies including Hardison (2003) and Motohashi-Saigo and Hardison (2009), I hypothesized that focused perceptual training could be effective for the correct identification of vowel duration. In other words, L2 learners would have higher accuracy in identifying the correct length of vowels after training.

Hypothesis 2b: Based on the previous training studies (e.g., Lively et al., 1993), I hypothesized that L2 learners' accuracy in identifying vowel length would increase, and response time (i.e., RT) would decrease as they progressed in training. As the other studies show, the largest improvement in accuracy and RT could take place between Week 1 and Week 2 of the training or from the pretest to the end of Week1.

Research Question 3: Which type of input in training, AV (with waveform display) or A-only, is more effective for development of identification accuracy of durational contrasts in L2 vowels? Does the effectiveness vary with proficiency level, vowel type, and preceding consonant?

Hypothesis 3: Based on Hardison (2003) and Hardison and Motohashi-Saigo (2010), I hypothesized that the most effective type of training would be AV training. Hardison and Motohashi-Saigo suggested that L2 learners can use visual cues, specifically waveforms as "a valuable source of input in L2 learning" (p. 42).

Research Question 4: Does perception training transfer to production improvement?

Hypothesis 4: Based on Hardison (2003) and Bradlow, Akahane-Yamada, Pisoni, and Tohkura (1999), I hypothesized that the effect of the training would transfer to another skill (i.e., production) if the training was effective.

Research Question 5: Does training generalize to novel stimuli spoken by a familiar talker from training as well as stimuli spoken by an unfamiliar voice? Does the ability to generalize vary according to the modality of training input? Do other stimulus factors affect the process?

Hypothesis 5: Based on Hardison (2003), I hypothesized that the effect of the training would generalize to correct identification of new tokens and a new voice if the training was effective.

Overview of Study Design

Two experiments were conducted for this study. Experiment 1 was designed to investigate factors affecting the identification and production of L2 vowel duration in Japanese. In addition, it had the objective of potentially reducing the number of factors and/or levels for analysis of the effects of training (Experiment 2) if they were not statistically significant. A cross-sectional design was adopted for the experiment. A between-subject factor was L2 proficiency (i.e., High, Mid, Low). Within-subject factors were vowel type: /a/, /u/ (one high and one low vowel), pitch pattern (where the dot represents a syllable boundary): LH.HH, LH.HL, HL.LL, LH.H, HL.L, L.HH, L.HL, H.LL, L.H, H.L, and preceding consonant: /k/, /s/ (one stop and one fricative). Dependent variables were perception accuracy (i.e., percentage of correct identification of vowel length), production accuracy (i.e., based on NSs' ratings of correct pronunciation), and perception reaction time (RT) (i.e., RT in milliseconds). Independent and dependent variables are summarized in Table 2 below.

Table 2: Summary of independent and dependent variables for Experiment 1

	Variables	Description
Between-Subject	L2 Proficiency (3)	High, Mid, Low
Within-Subject	Vowel Type (2)	Low, High (/a/, /u/)
	Pitch Pattern (10)	LH.HH, LH.HL, HL.LL LH.H, HL.L L.HH, L.HL, H.LL L.H, H.L
	Preceding Consonant (2)	/k/, /s/
Dependent Variables	Perceptual Identification Accuracy	Percentages of correct identification
	Production Accuracy	NSs' ratings of correct pronunciation
	Perception RT	RT in milliseconds

CHAPTER 2: EXPERIMENT 1

Method

Participants

Participants were 64 L2 learners, whose L1 was American English, studying Japanese as a foreign language at a large Midwestern university in the U.S. They were enrolled in the first year (n=24), second year (n=17), third year (n=16), and fourth year (n=7) Japanese courses at the time of the experiment. The participants enrolled in the first year Japanese language course (12 females and 12 males) did not have previous knowledge of Japanese when they started to study it. At the time of participation, they had studied Japanese for about three months. The participants enrolled in the second year Japanese language course (9 females and 8 males) had passed the first year course (i.e., a total of 125 hours instruction in class) and were in the third semester. The participants enrolled in the third year Japanese language course (9 females and 7 males) had passed the second year course (i.e., a total of 250 hours instruction in class since the beginning of their study) and were in the fifth semester. The participants in the fourth year Japanese course (6 females and 1 male) had passed the third year course (i.e., a total of 350 hours instruction in class since the beginning of their study) and were in the seventh semester. No heritage learners participated in this study, and all of the participants reported normal hearing and vision.

In the elementary Japanese language courses, the first- and second-year courses, the contact hours of the class were 50 minutes per day, five times per week (a total of 125 hours of instruction per year). In class, an instructor corrected the students' inaccurate pronunciation during oral drills and communicative activities; however, no special training for discriminating particular phonemic contrasts was usually provided.

Generally speaking, the longer they study Japanese, the more interactions they have with Japanese NSs. However, it was not necessarily the case that those interactions led to development in Japanese proficiency because of individual differences such as motivation, L2 use, and L2 exposure.

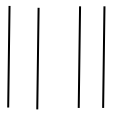
Materials

Production Test: Target materials included 16 tokens contrasting long and short vowels (Appendix A). High and low vowels, /a, u/, and two consonants /k, s/ were used to construct target stimuli. The two consonants, a voiceless velar stop and a voiceless fricative, were selected for this experiment based on the potential role played by consonant-vowel sonority difference on learner perception (Hardison & Hotohashi-Saigo, 2010). The vowels /a, u/ represent the longest and shortest vowels respectively in the Tokyo dialect. In addition to the target tokens in Appendix A, four practice trials in Appendix B were prepared to familiarize participants with the task.

Perception Test: Target materials included 40 tokens contrasting long and short vowels (Appendix C). High and low vowels, /a, u/, and two consonants /k, s/ were used to construct target stimuli. Also, 10 pitch patterns that occur in the language were used in this study as shown in Figure 3. Each target was assigned one of the 10 pitch patterns. As a result, the target stimuli included both real words and pseudo-words as in Appendix C.

Figure 3: Pitch patterns used in this study with an example word with the pitch pattern

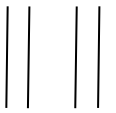
1. CVV.CVV



LH HH

koo.hoo
'official information'

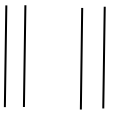
2. CVV.CVV



LH H L

koo.hii
'coffee'

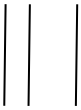
3. CVV.CVV



HL LL

kee.zai
"economics'

4. CVV.CV



LH H

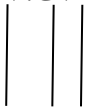
ii.e
'no'

5. CVV.CV

HL L

aa.to
'art'

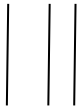
6. CV.CVV



L HH

ji.koo
'statute of limitation'

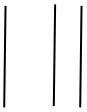
7. CV.CVV



L HL

i.suu
'heteromerous'

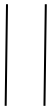
8. CV.CVV



H LL

ma.naa
'manner'

9. CV.CV



L H

ha.na
'flower'

10. CV.CV

H L

u.mi
'sea'

In the current project, mostly pseudo words were used (i.e., words that can be pronounced in terms of the phonology of Japanese; however, they do not have a meaning). Based on the psycholinguistics research (Bundgaard-Nielsen, Best, & Tyler, 20011; Imai, Walley, & Flege, 2005; Metsala, 1997; Ziegler, Muneaux, & Grainger, 2003), it was found that neighborhood density and a learner's size of vocabulary significantly affected the word recognition and determination of the phonological contrasts. Therefore, in order to avoid the effects, most of the stimuli in the current project were pseudo words. There were 10 real words in order to balance the stimuli; however, their frequency was not high and the learners may have had limited exposure to them, if any. In the analysis, they will be compared with pseudo words and used if there are no statistical differences between the two types of words. In addition to the target tokens in Appendix C, four practice trials in Appendix D were prepared to familiarize participants with the task.

Six NSs of Japanese, whose ages ranged from 18 to 35 years old and who were born in Tokyo or near the Tokyo area of Japan, were recruited (4 females and 2 males) to record the stimuli. In this project, pitch patterns used in *kyootsuu-go*, a dialect spoken in the Tokyo area (Shibatani, 1990), were used. Therefore, the NSs who were born in the area were recruited. While the NSs were bilinguals who speak English and Japanese, their dominant language is Japanese. One of the female speakers, Talker 1, produced the testing and practice tokens for the perception test in Experiment 1.

Procedures

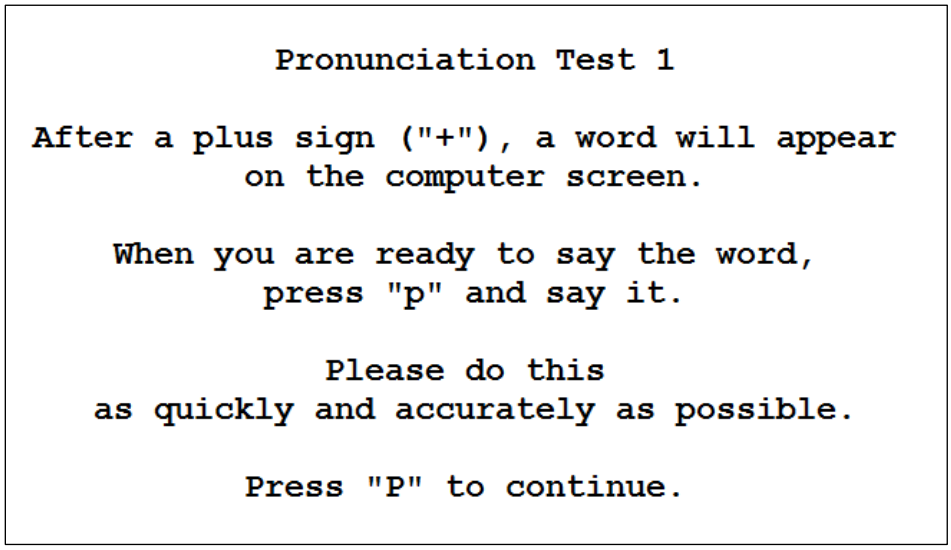
Production Test: Computerized production test was created using *E-Prime*. The production test was administered prior to the perception test. This order was adopted in order to avoid providing

participants with auditory input of the target tokens prior to the production tests, which could influence the participants' correct pronunciation of the target tokens.

During production testing, a visual prompt task of 16 tokens, listed in Appendix A, was given to participants. Prior to the target stimuli, practice tokens, listed in Appendix B, were given in order to familiarize participants with the task. The stimuli were written in *roomaji* (i.e., the alphabet representation of Japanese sounds), not *hiragana*, because the distinction between long and short vowels was clearer (e.g., *kaakaa* vs. かぁかぁ 'high school') for some participants whose proficiency was lower. The experiment was conducted in a quiet room.

The procedure of production testing is described below. First, participants read the instructions on the computer screen (Figure 4).

Figure 4: Instructions for the production test for Experiment 1



Then, a plus sign ('+') appeared on the computer screen for two seconds (Figure 5) followed by the target word while the participant was asked to read aloud. Then, a stimulus appeared on the

screen (Figure 6) and a participant was asked to read. When the participants were ready to pronounce the word, they were asked to press ‘P’ to move to the next screen.

Figure 5: A “+” sign shown before the presentation of stimuli

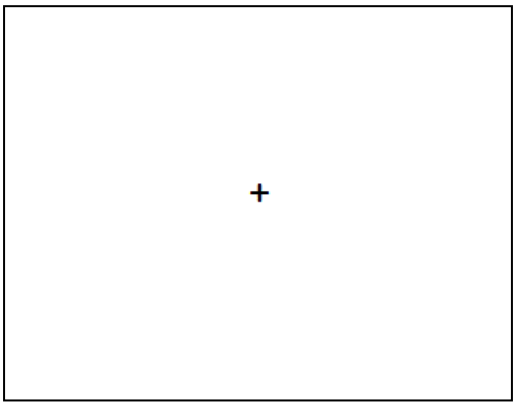
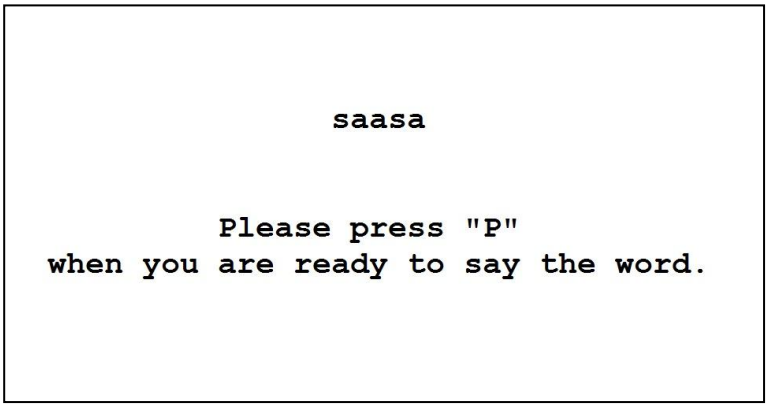


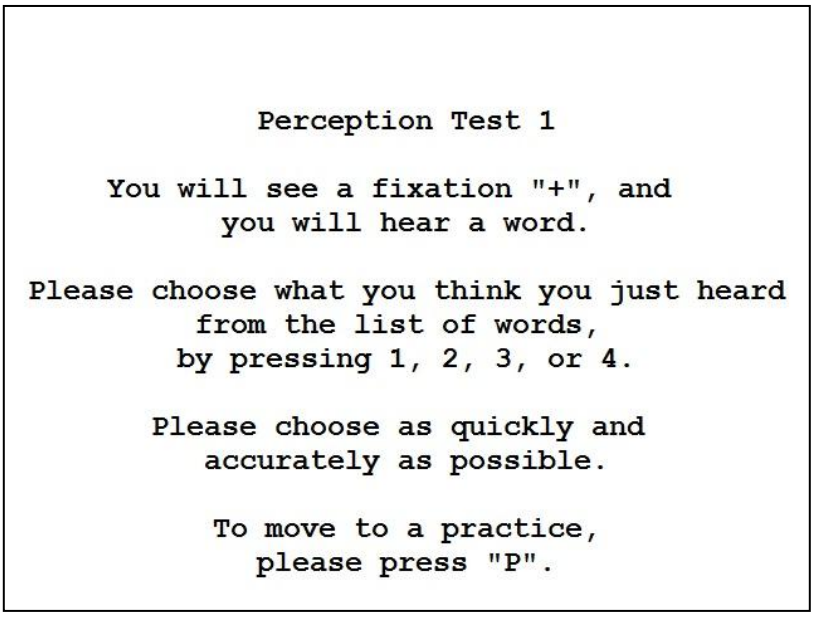
Figure 6: Presentation of the stimuli for the production test in Experiment 1



In the next screen, the participants were asked to pronounce the word. The participants were asked to press the key ‘P’ to move to the next stimulus.

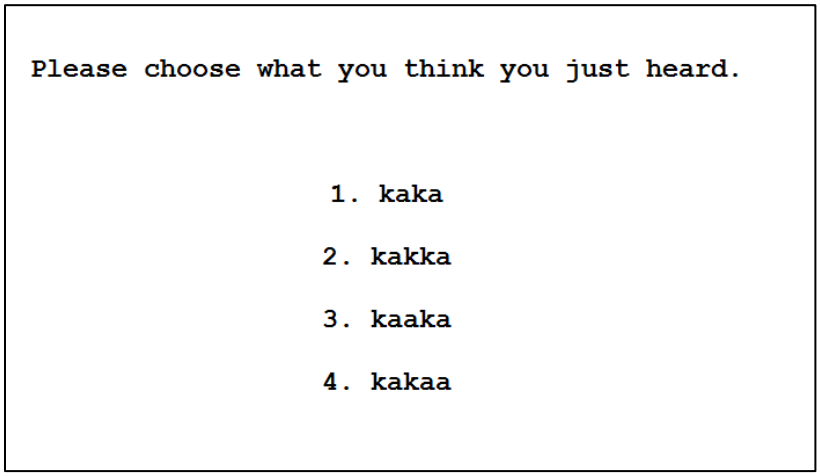
Perception Test: After the production test, a perception test was given. During perception testing, participants were given a forced-choice, four-alternative identification task involving a total of 40 target stimuli (see Appendix C). The rationale for using the identification task rather than a discrimination task was based on previous studies (e.g., Logan et al., 1991). The choices were written in *romanization* to make the distinction between long and short vowels clearer. First, participants read the instructions on the computer screen (Figure 7).

Figure 7: Instructions for the perception test in Experiment 1



Then, a plus sign ('+') appeared on the computer screen (Figure 5) for two seconds. After participants listened to a word played on the computer, they were asked to choose one option that they thought matched what they heard from the list provided on the computer screen (Figure 8). The participants were able to see the choices while they were listening to the stimuli.

Figure 8: Presentation of the auditory stimuli and identification task for the perception test in Experiment 1



When the auditory stimulus ended, the timer to measure RT started. As soon as the participant made a choice, the timer to measure RT stopped. Then, the computer screen showed the plus sign and moved to the next stimulus. There was no feedback given in Experiment 1.

Prior to the target stimuli, the practice tokens in Appendix D were given to the participants in order to familiarize them with the task. For each stimulus, a participant's responses, identification accuracy, and RT were recorded on the computer and saved for later analysis. It was determined that the participants whose scores were 90% or higher in the perception test would be excluded from this study in order to avoid ceiling effects.

Results

Identification accuracy scores (i.e., percentages of correct responses), production accuracy scores (i.e., Japanese NSs' rating), and Reaction Time (RT) in milliseconds were tabulated. The data were analyzed and are reported in the following order: (1) the overall results

of the perception and production tests, (2) factors affecting production accuracy of vowel duration, (3) factors affecting perception accuracy of vowel duration, and (4) factors affecting perception latency. For the statistical analysis, alpha level was set at .05 (α = .05).

Overall Results of the Production Test: A total of 64 participants took a production test in Experiment 1. A total of 16 items in Appendix A were used, and accuracy of correct pronunciation was measured using NSs' judgment. Three female NSs of Japanese rated the participants' pronunciation. The NSs, whose ages ranged from 30 to 40 years old, were born in Japan and lived in the US. All of the three raters had taken linguistics courses and had Japanese teaching experience. For rating, the raters were asked to listen to the words pronounced by the participants and choose what they thought they heard from the list provided as in Table 3 below.

Table 3: A sample task for the raters

Item to be Rated	A List of Choices
kaakaa	(a) kaakaa (b) kaaka (c) kakka (d) kaka (e) other: ( )

When the rater chose (e) 'other', she was told to write down what she thought she heard. When the rater judged that the participants pronounced the word correctly, one point was given for the token; otherwise, no point was given. The pitch pattern was not measured because it was not a focus in the production part and as pseudo words, learners would not have known what pattern to use. The three raters coded each production individually, and the result on which at least two

raters agreed was used as the basis for the production score for the item. Interrater reliability was checked using Pearson Correlation/Coefficient. There was a significant positive correlation between Rater 1 and Rater 2 (r = .896, p = .001, R^2 = .80), between Rater 1 and Rater 3 (r = .895, p = .001, R^2 = .80), as well as between Rater 2 and Rater 3 (r = .887, p = .001, R^2 = .79); the correlation was strong. For all the items, there was an agreement from at least two raters; therefore, there was no need to resolve any ambiguous items.

The 16 tokens produced by learners were divided into four types depending on the location of the long vowels: (1) CVV.CVV, contained long vowels in the first and second syllables, (2) CVV.CV, contained long vowels in the first syllable, (3) CV.CVV contained long vowels in the second syllable, and (4) CV.CV, contained no long vowels. Table 4 shows mean scores of production accuracy sorted by the preceding consonant, vowel type, and token type, obtained from 64 participants. The mean production accuracy was 70.38% (*s.d.* 16.96).

Table 4: Mean accuracy for the production accuracy in Experiment 1

Preceding Consonant /k/				Preceding Consonant /s/			
Vowel /a/		Vowel /u/		Vowel /a/		Vowel /u/	
Item	Mean (*s.d.*)	Item	Mean (*s.d.*)	Item	Mean (*s.d.*)	Item	Mean (*s.d.*)
kaa.kaa	.83 (.38)	kuu.kuu	.72 (.45)	saa.saa	.84 (.37)	suu.suu	.70 (.46)
kaa.ka	.83 (.38)	kuu.ku	.94 (.24)	saa.sa	.77 (.43)	suu.su	.72 (.45)
ka.kaa	.66 (.48)	ku.kuu	.53 (.50)	sa.saa	.66 (.48)	su.suu	.67 (.47)
ka.ka	.73 (.45)	ku.ku	.67 (.47)	sa.sa	.63 (.49)	su.su	.58 (.50)

Overall Results of the Perception Test: A total of 64 participants took a perception test in Experiment 1. A total of 40 items in Appendix C were used, and perception accuracy and latency were measured, using *E-prime*. For the perception accuracy, the participants' choice was

coded either correct (one point) or wrong (zero). When a participant did not make a choice, no point was given for the specific token. The perception reaction time (RT) was measured in milliseconds using *E-Prime*.

Originally, I planned to treat the participants' L2 proficiency as a between-subject factor, with the intention of dividing the participants into three groups using the results of Experiment 1 in order to examine how proficiency affected correct identification of vowel duration in Japanese. However, the use of test scores to assess proficiency is arbitrary because it is not clear what scores can indicate the proficiency level. In addition, there is no appropriate independent measurement available. *The Japanese Language Proficiency Test* (JLPT) has a listening section; however, it measures holistic skills in listening. Therefore, the measurement is not directly related to the issue of vowel duration. Finally, the courses that the participants were enrolled in were not valid estimates of their ability to identify vowel duration as shown in Figure 9. Figure 9 and Table 5 shows the distribution of accuracy scores according to the participants' length of time studying Japanese at the college level (i.e., 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} year of the Japanese classes). Even some 1^{st} year learners obtained more than 90% identification accuracy, which was equal to the accuracy of more advanced learners. Therefore, the data were collapsed into one group and the analysis focused on the remaining within-subject variables.

Figure 9: Distribution of perception accuracy by course enrollment. (For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.)

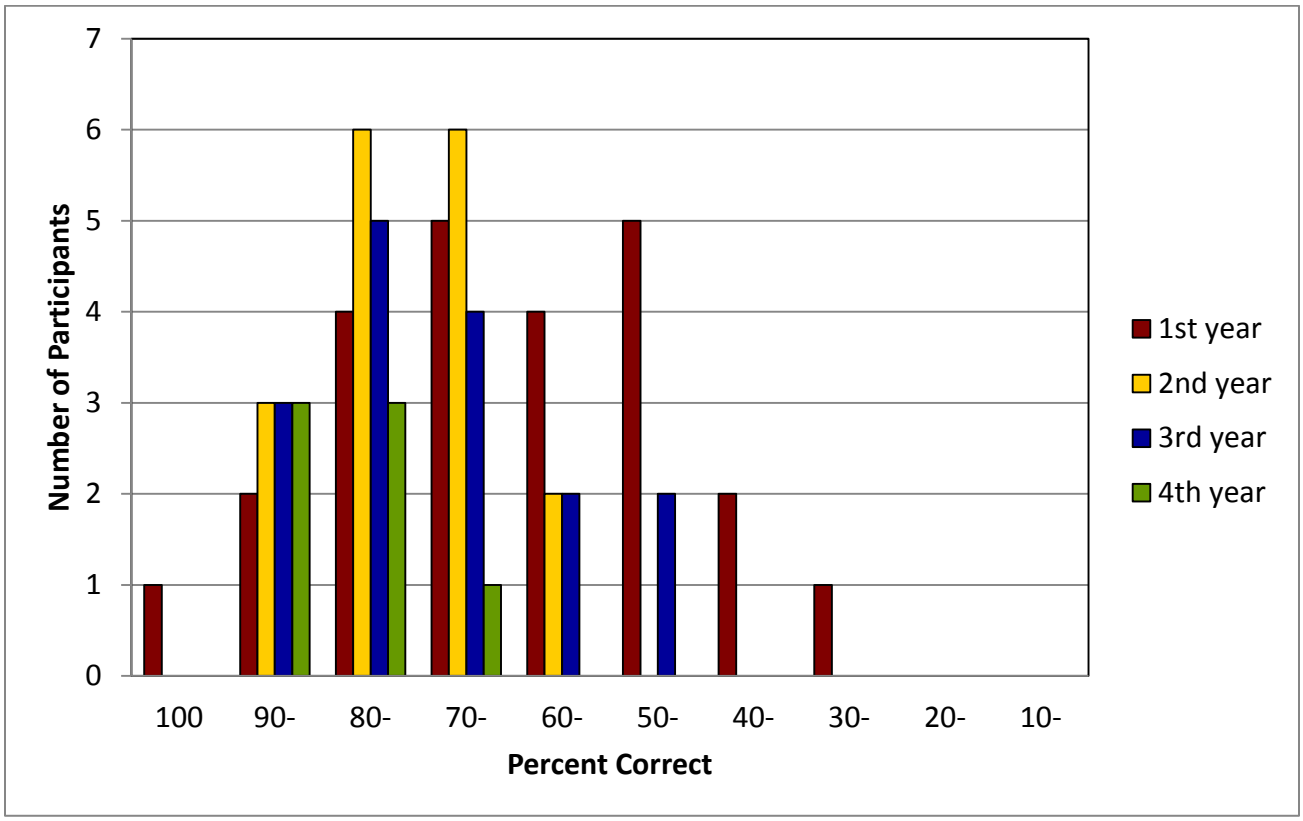


Table 5: Distribution of perception accuracy by course enrollment in percentages

Courses	100%	90 – 99.99%	80 – 89.99%	70 – 79.99%	60 – 69.99%	50 – 59.99%	40 – 49.99%	30 – 39.99%	0 – 29.99%
1st year	1	2	4	5	4	5	2	1	---
2nd year	---	3	6	6	2	---	---	---	---
3rd year	---	3	5	4	2	2	---	---	---
4th year	---	3	3	1	---	---	---	---	---
Total	1	11	18	17	8	7	2	1	---

Analysis of Production Data: A three-way design ANOVA was used to test whether the preceding consonant, type of vowel, or token type significantly affected accuracy in pronouncing vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s/), vowel type (2; /a/ and /u/), and token type (4: CVV.CVV, CVV.CV, CV.CVV, CV.CV). The dependent variable was production accuracy. Results indicated significant main effects of vowel type, $F_{\text{Vowel}}(1, 63) = 5.063$, $p = .028$, $\eta_p^2 = .074$, and token type, $F_{\text{Type}}(3, 189) = 6.290$, $p < .001$, $\eta_p^2 = .091$; however, preceding consonant was marginally significant, $F_{\text{PreC}}(1, 63) = 3.768$, $p = .057$. Thus, it was found that vowel type had a significant influence, but not the preceding consonant. The mean accuracy scores for the tokens with the vowel /a/ and /u/ were .74 (*s.d.* .21) and .69 (*s.d.* .19) respectively. Thus, it was easier for the learners to correctly pronounce vowel duration when the vowel was /a/. In addition, it was found that token type had a significant influence. In order to locate where the differences existed in the four token types, pairwise comparisons were performed using the Bonferroni correction. The mean accuracy for the CVV.CVV, CVV.CV, CV.CVV, and CV.CV is tabulated in Table 6 below.

Table 6: Mean production accuracy of the four tokens in Experiment 1

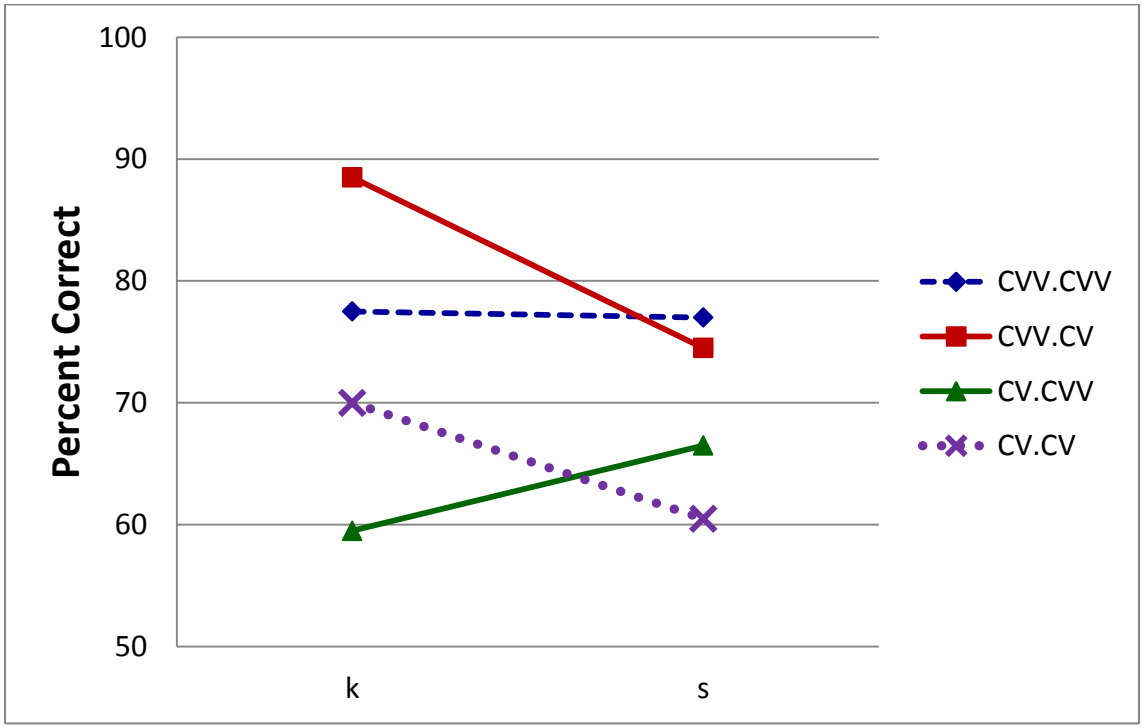
Token	Mean (*s.d.*)
CVV.CVV	.77 (.26)
CVV.CV	.81 (.24)
CV.CVV	.62 (.36)
CV.CV	.65 (.35)

Results indicated that the (2) CVV.CV type was significantly different from the (3) CV.CVV type ($p = .005$) as well as the (4) CV.CV type ($p = .011$). The mean scores of the (2)

CVV.CV were higher than those of the (3) CV.CVV as well as the (4) CV.CV. Therefore, it was concluded that the (2) CVV.CV type in which the long vowel is in the first syllable was comparable to CVV.CVV and easier to produce correctly than (3) CV.CVV and (4) CV.CV.

In addition to the main effects above, the Preceding Consonant x Token Type interaction was significant, $F(3, 189) = 4.002$, $p = .009$, $\eta_p^2 = .061$. The results of simple effects tests indicated that the CV tokens (Type 4) revealed significant effects on the two consonants /k/ and /s/, $F(1, 63) = 7.87$, $p = .007$, as shown in Figure 10. The CV.CV tokens with the consonant /k/ (a stop) had higher accuracy than those with the consonant /s/ (a fricative).

Figure 10: Effects of consonant and token type on production accuracy in Experiment 1



An error analysis was then conducted on the production data. There were three cases in which the participants did not produce anything. Excluding these errors, there were 298 incorrect productions and they are summarized in Table 7 below.

Table 7: Errors observed in the production data in Experiment 1

Token with /a/	Errors	Number	Token with /u/	Errors	Number
CaaCaa	CaaCa	15	CuuCuu	CuuCu	35
	CaCaa	5		CuCuu	2
	CaCCaa	1			
CaaCa	CaaCaa	17	CuuCu	CuuCuu	17
	CaCa	3		CuCu	4
	CaCaa	1		CuCuu	1
	CaCCa	1			
	CaCCaa	1			
CaCaa	CaaCaa	29	CuCuu	CuuCuu	35
	CaCCaa	6		CuCCuu	8
	CaaCa	5		CuuCu	3
	CaCa	2		CuCu	3
CaCa	CaCaa	21	CuCu	CuCuu	26
	CaaCa	19		CuuCu	16
	CaaCaa	1		CuuCuu	4
	CaCCa	1		CuCCu	1
	CaCCaa	1		CuCCuu	1

Note: C = consonant (/k/ or/s/)

As shown in the table above, the most common errors observed for the CVV.CVV tokens were CVV.CV; a long vowel in the second syllable[2] was shortened to a short vowel. For the CVV.CV tokens, the most common error was CVV.CVV; a short vowel in the second syllable was lengthened. For the CV.CVV tokens, the most common error was CVV.CVV; a short vowel in the first syllable was lengthened. Finally, the two major errors for the CV.CV tokens were CV.CVV and CVV.CV; one of the short vowels was lengthened. Based on this error analysis, it was concluded that when a token contained two long vowels, the learner shortened the one on the second syllable. On the other hand, when a token contained both short and long vowels, the learner lengthened the short vowel. When a token contained two short vowels, the learner lengthened the short vowel either on the first or second syllable.

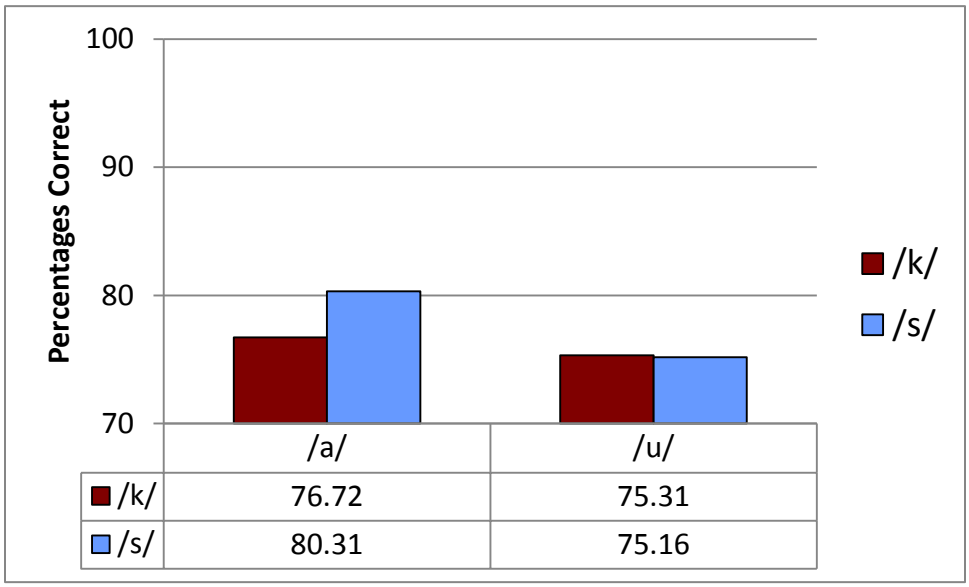
In conclusion, in this section, factors affecting production accuracy of vowel duration were examined. It was found that vowel and token type had significant main effects on producing tokens containing vowel duration. In addition, the interaction between the preceding consonant and token type was found; the CV.CV token with the consonant /k/ had higher production accuracy than those with the consonant /s/.

Analysis of Factors Affecting Perception Accuracy: As possible factors that affected identification accuracy of vowel duration, preceding consonant (2; /k/, /s/), vowel type (2; /a/, /u/), and pitch patterns (10) were examined. As shown in Figure 3, not every token type occurs in the language in conjunction with every possible pitch pattern. The overall mean score for perception accuracy was 76.04% (*s.d.* 15.17). The mean identification accuracy for words with the preceding consonants /k/ and /s/ was 76.02% (*s.d.* 17.89) and 77.73% (*s.d.* 14.64)

[2] In this dissertation, the word final position is represented as the second syllable in order to make a contrast to the first syllable.

respectively; the mean identification accuracy for words with the vowels /a/ and /u/ were 78.52% (*s.d.* 14.98) and 75.23% (*s.d.* 17.35) respectively. Then, the preceding consonant and vowel type were combined. Mean scores for identification accuracy for /ka/, /ku/, /sa/, and /su/ were 76.72% (*s.d.* 20.55), 75.31% (*s.d.* 18.17), 80.31% (*s.d.* 13.33), and 75.16% (*s.d.* 20.31) respectively (Figure 11).

Figure 11: Mean perception accuracy by preceding consonant and vowel type



	/a/	/u/
/k/	76.72	75.31
/s/	80.31	75.16

Descriptive statistics were then conducted on the responses to the 10 pitch patterns (1: LH.HH, 2: LH.HL, 3: HL.LL, 4: LH.H, 5: HL.L, 6: L.HH, 7: L.HL, 8: H.LL, 9: L.H, 10: H.L) assigned to each combination of the consonants and vowels: /ka/, /sa/, /ku/, and /su/. Table 8 and Figure 12 show the descriptive statistics for perception accuracy by pitch pattern, preceding consonant, and vowel type.

Table 8: Descriptive statistics for perception accuracy by pitch pattern, preceding consonant, and vowel type

Pitch Pattern	Preceding Consonant /k/				Preceding Consonant /s/			
	Vowel /a/		Vowel /u/		Vowel /a/		Vowel /u/	
	Mean	(*s.d.*)	Mean	(*s.d.*)	Mean	(*s.d.*)	Mean	(*s.d.*)
1	.83	(.38)	.63	(.49)	.73	(.45)	.84	(.37)
2	.53	(.50)	.70	(.46)	.58	(.50)	.58	(.50)
3	.53	(.50)	.61	(.49)	.69	(.47)	.66	(.48)
4	.92	(.27)	.95	(.21)	.91	(.29)	.83	(.38)
5	.83	(.38)	.66	(.48)	.75	(.44)	.55	(.50)
6	.84	(.37)	.73	(.45)	.83	(.38)	.84	(.37)
7	.80	(.41)	.77	(.43)	.67	(.47)	.77	(.43)
8	.59	(.50)	.70	(.46)	.98	(.13)	.56	(.50)
9	.86	(.35)	.92	(.27)	.94	(.24)	.95	(.21)
10	.94	(.24)	.86	(.35)	.95	(.21)	.94	(.24)

Figure 12: Mean perception accuracy by pitch pattern, preceding consonant, and vowel type

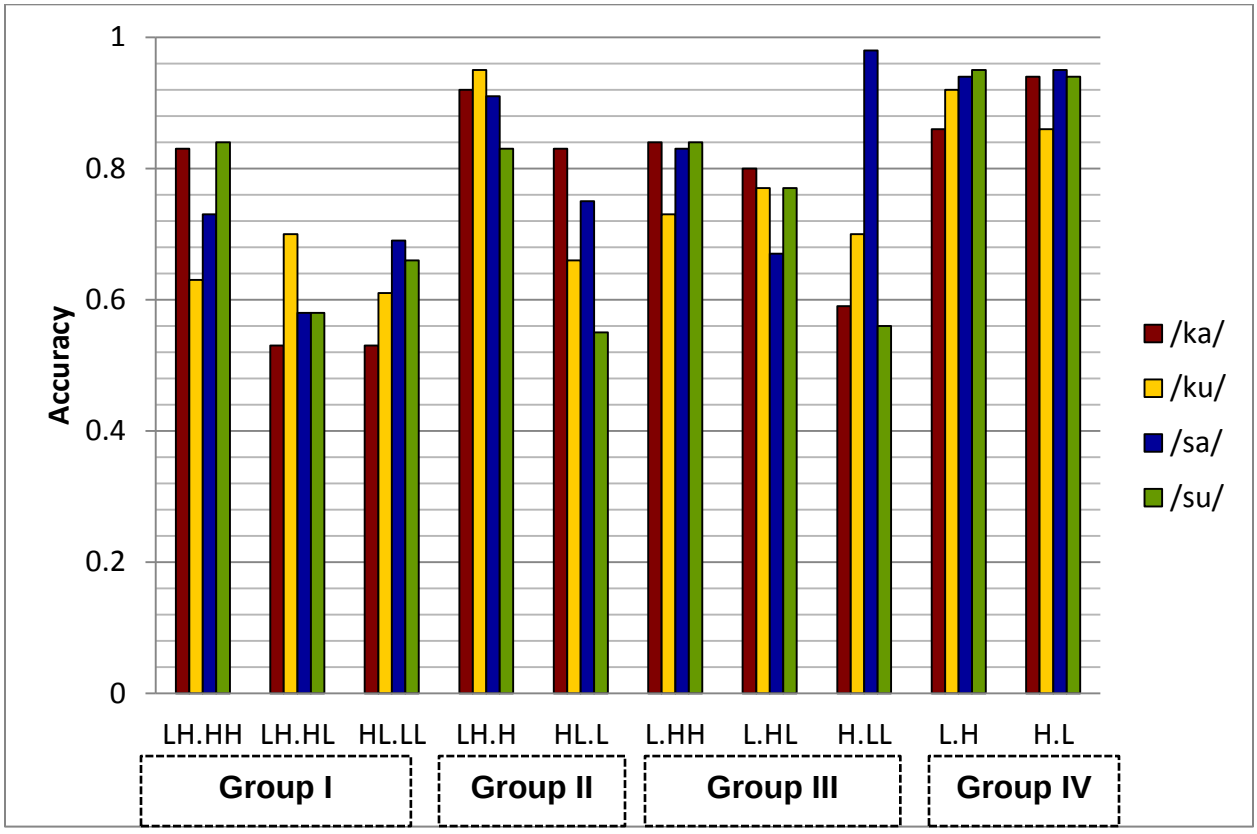


Figure 12 above shows that perception accuracy for the L.H, H.L, and LH.H pitch was higher than the other pitch patterns. The LH.H and HL.L patterns also had relatively higher perception accuracy; the LH.HH, LH.HL, and HL.LL patterns had relatively lower perception accuracy.

In order to examine whether preceding consonant, vowel type, and pitch pattern significantly affected the correct identification of vowel duration in Japanese, the 10 pitch patterns were divided into 4 categories according to the location of the long vowels (i.e., first and/or second syllables) as shown in Figure 12. Pitch patterns (1) LH.HH, (2) LH.HL, and (3) HL.LL were categorized into Group I which contained two long vowels (CVV.CVV); pitch patterns (4) LH.H and (5) HL.L were categorized into Group II which contained one long vowel in the first syllable (CVV.CV); pitch patterns (6) L.HH, (7) L.HL, (8) H.LL were categorized into Group III which contained one long vowel in the second syllable (CV.CVV); and pitch patterns (9) L.H and (10) H.L were grouped into Group IV which did not contain any long vowels and was used as baseline information.

A three-way ANOVA was used to test whether preceding consonant, vowel type, and/or pitch pattern in Group I (CVV.CVV) significantly affected the correct identification of vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s), vowel type (2; /a/ and /u/), and pitch pattern (3: LH.HH, LH.HL, HL.LL). The dependent variable was perception accuracy. Results indicated significant main effects of pitch pattern, $F_{\text{Pitch}}(2, 126) = 10.866$, $p < .001$, $\eta_p^2 = .147$; however, preceding consonant, $F_{\text{PreC}}(1, 63) = 1.726$, $p = .194$, and vowel type, $F_{\text{Vowel}}(1, 63) = .578$, $p = .450$ were not significant.

It was found that neither the type of vowel nor the preceding consonant affected identification of the vowel duration of the tokens in Group I. However, the pitch patterns

affected the correct identification. In order to locate where the differences existed among the three pitch patterns, pairwise comparisons were performed using the Bonferroni correction. Results indicated that (1) LH.HH was significantly different from (2) LH.HL ($p < .001$) and (3) HL.LL ($p = .001$). The pitch pattern (1) LH.HH, a mean of .76, had significantly higher accuracy than (2) LH.HL, a mean of .60 and (3) HL.LL, a mean of .62. Then, these three pitch patterns were compared. The pitch patterns (1) and (2) shared the same pitch on the first syllable and only had a difference in the pitch pattern on the second syllable, HH and HL respectively. Yet, the pitch patterns (1) and (3) did not share any similarity. The pattern (1) started with low pitch and kept high pitch after the second mora; the pattern (3) had the opposite pattern. The comparison between (1) and (2) suggested that the differences in the pitch pattern on the second syllable were the key.

In addition to the main effect of pitch pattern, the results indicated that the Preceding Consonant x Vowel Type x Pitch Pattern interaction was significant, $F(2, 126) = 7.322$, $p = .001$, $\eta_p^2 = .104$. The results of simple effects tests revealed that perception accuracy of /ka/ was higher than /ku/ with the LH.HH pitch while that of /ku/ was higher than /ka/ with the LH.HL pitch.

Error analysis was conducted on the responses to the CVV.CVV tokens. Table 9 below shows the four choices used in the identification task for the tokens with the CVV.CVV structure. The order of presentation of the four choices was randomized; therefore, each token had a different order of options. Among the four choices, (A) was the correct response; (B) was different from (A) in terms of the vowel length of the second syllable, (C) was different because the first syllable contains a geminate, instead of a long vowel, and (D) was different in terms of the vowel length of the first syllable. Previous literature (e.g., Motohashi-Saigo & Hardison,

2009) found that L2 learners misperceived long vowels as geminates. Also, the token with the CV.CV structure was considered too different from the CVV.CVV structure; therefore, geminates were included as one of the choices.

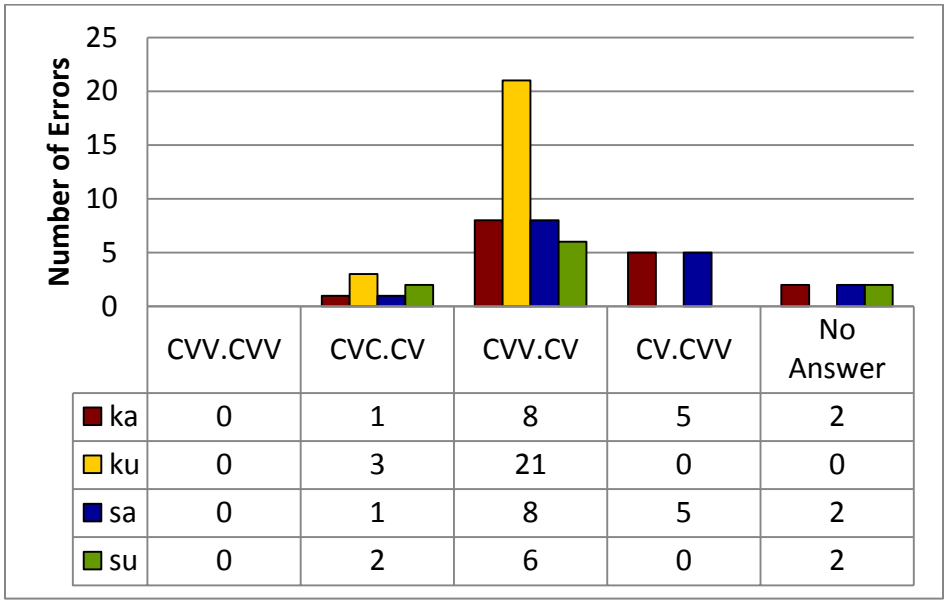
Table 9: An example of choices used in the identification task for CVV.CVV tokens

Choices in the Identification Task	Examples	Selection would indicate:
A. CVV.CVV	kaa.kaa	-correct
B. CVV.CV	kaa.ka	-misperception of long vowel in second syllable
C. CVC.CV	kak.ka	-misperception of long vowel in second syllable -misperception of long vowel in first syllable as geminate
D. CV.CVV	ka.kaa	-misperception of long vowel in first syllable

Note: Syllable boundaries were not marked in the experiment.

Figure 13 shows the number of errors that the participants made for the tokens with the (1) LH.HH pitch pattern. There were 66 errors in total; approximately 65.15% of the errors were observed for the choice CVV.CV (misperception of long vowels in second syllable).

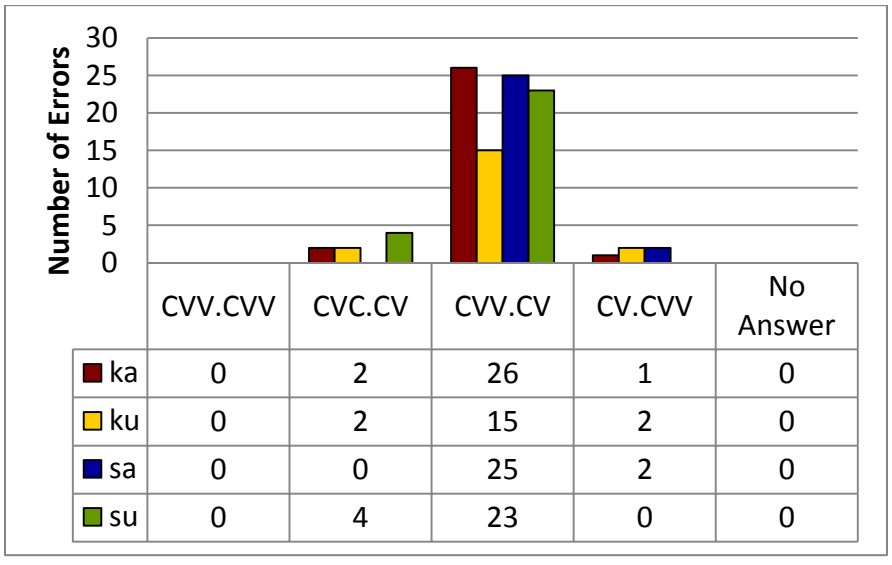
Figure 13: Errors for the tokens CVV.CVV with the (1) LH.HH pitch pattern



	CVV.CVV	CVC.CV	CVV.CV	CV.CVV	No Answer
ka	0	1	8	5	2
ku	0	3	21	0	0
sa	0	1	8	5	2
su	0	2	6	0	2

Next, Figure 14 shows the number of errors that the participants made for the tokens with the (2) LH.HL pitch pattern. There were 102 errors in total; approximately 88.25% of the errors were again observed for the choice CVV.CV (misperception of long vowel in second syllable).

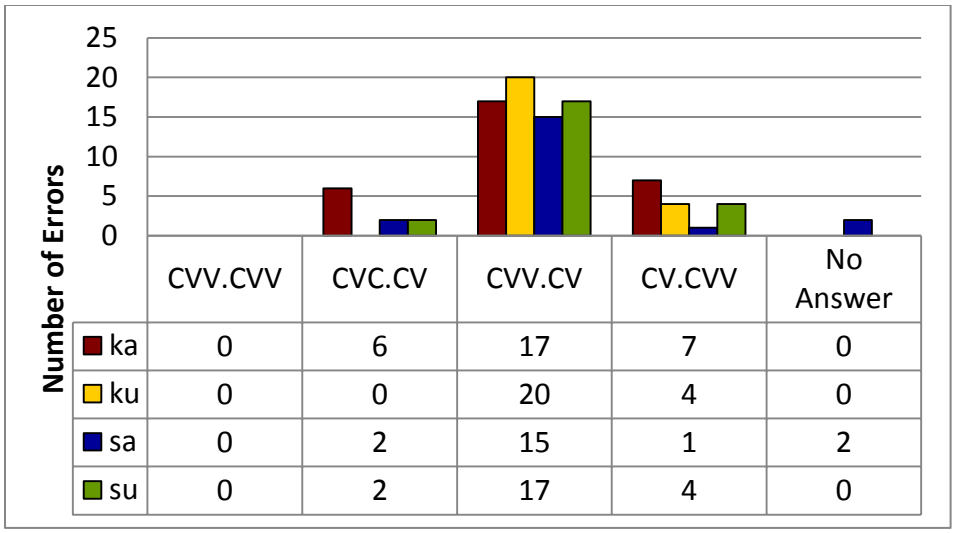
Figure 14: Errors for the tokens CVV.CVV with the (2) LH.HL pitch pattern



	CVV.CVV	CVC.CV	CVV.CV	CV.CVV	No Answer
ka	0	2	26	1	0
ku	0	2	15	2	0
sa	0	0	25	2	0
su	0	4	23	0	0

Finally, Figure 15 shows the number of errors that the participants made for the tokens with the (3) HL.LL pitch pattern. There were 97 errors in total; approximately 71.13% of the errors were observed for the choice CVV.CV, similar to the other two patterns.

Figure 15: Errors for the tokens CVV.CVV with the (3) HL.LL pitch pattern



	CVV.CVV	CVC.CV	CVV.CV	CV.CVV	No Answer
ka	0	6	17	7	0
ku	0	0	20	4	0
sa	0	2	15	1	2
su	0	2	17	4	0

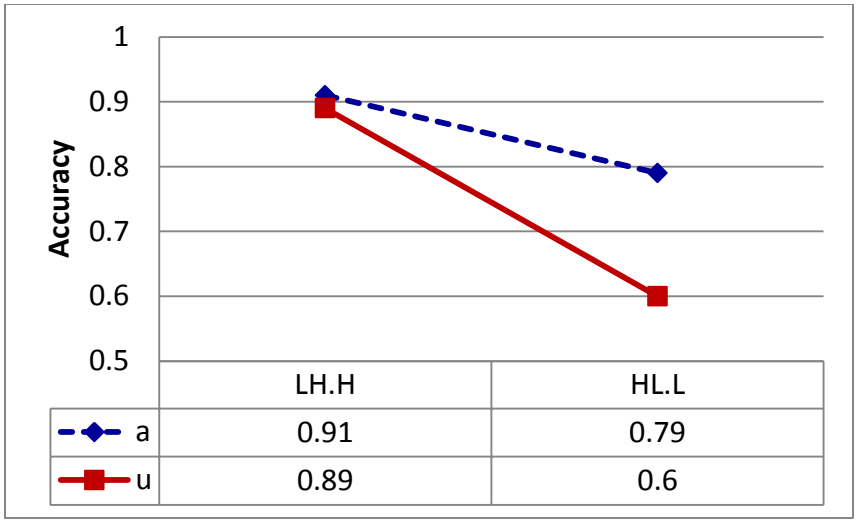
The error analysis revealed that the learners had a tendency to incorrectly perceive the CVV.CVV tokens as CVV.CV. This error pattern suggested that a long vowel in the second syllable was perceived as a short vowel. In addition, there were more errors observed for the vowel /u/ compared to the vowel /a/. The simple effects analysis of the interaction also suggested that the vowel /a/ had higher accuracy than the vowel /u/ for the three pitch patterns in this group.

Next, a three-way ANOVA was used to test whether the preceding consonant, vowel type, and/or pitch pattern in Group II (CVV.CV) significantly affected the correct identification of vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s), vowel type (2; /a/ and /u/), and pitch pattern (2: LH.H, HL.L). The dependent variable was

perception accuracy. Results indicated significant main effects of preceding consonant, $F_{PreC}(1, 63) = 7.471$, $p = .008$, $\eta_p^2 = .106$, pitch pattern, $F_{Pitch}(1, 63) = 28.474$, $p < .001$, $\eta_p^2 = .311$, and vowel type, $F_{Vowel}(1, 63) = 10.938$, $p = .002$, $\eta_p^2 = .148$. Based on the results, the type of vowel, preceding consonant, and pitch pattern affected the identification of L2 vowel duration of the tokens in Group II. The mean accuracy scores of the tokens with the vowel /a/ and /u/ were .85 and .75 respectively. Therefore, it was easier for the learners to identify vowel duration when the vowel was /a/, compared to /u/. The mean accuracy scores of the tokens with the consonant /k/ and /s/ were .84 and .76 respectively. Thus, it was easier for the learners to identify vowel duration when the preceding consonant was /k/, compared to /s/. The mean accuracy scores of the tokens with the pitch pattern (4) LH.H and (5) HL.L were .90 and .70 respectively; therefore, the pattern (4) was easier than (5). Similar to the previous comparison between (1) LH.HH and (3) HL.LL, (4) LH.H and (5) HL.L did not share any similarity; the two tokens were very distinct.

In addition to the significant main effects above, the Vowel Type x Pitch Pattern interaction was significant, $F(1, 63) = 8.663$, $p = .005$, $\eta_p^2 = .121$. Results of simple effects tests revealed that accuracy for the vowel /u/ was significantly lower in the pitch pattern HL.L as shown in Figure 16.

Figure 16: Effects of vowel type and pitch pattern in Group II on perception accuracy



Error analysis was conducted on the responses to the CVV.CV tokens. Table 10 below shows the four choices used in the identification task for the tokens with the CVV.CV structure. Among the four choices, (B) was the correct response; (A) was different from (B) in terms of the vowel length in the second syllable, (C) was different because the first syllable contains a geminate, instead of a long vowel, and (D) was different in terms of the vowel length of the first syllable.

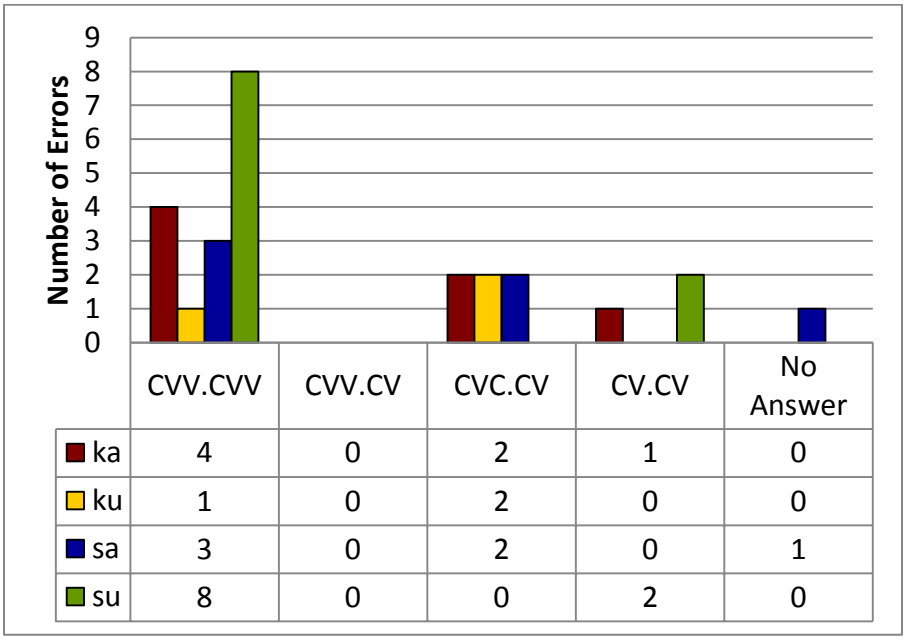
Table 10: An example of choices used in the identification task for CVV.CV tokens

Choices in the Identification Task	Examples	Selection would indicate:
A. CVV.CVV	kaa.kaa	-misperception of vowel length in second syllable
B. CVV.CV	kaa.ka	-correct
C. CVC.CV	kak.ka	-misperception of long vowel as geminate
D. CV.CV	ka.ka	-misperception of vowel length in first syllable

Note: Syllable boundaries were not marked in the experiment.

Figure 17 shows the number of errors that the participants made for the tokens with the (4) LH.H pitch pattern. There were 26 errors in total; approximately 61.54% of the errors were observed for the choice CVV.CVV (misperception of vowel length in the second syllable).

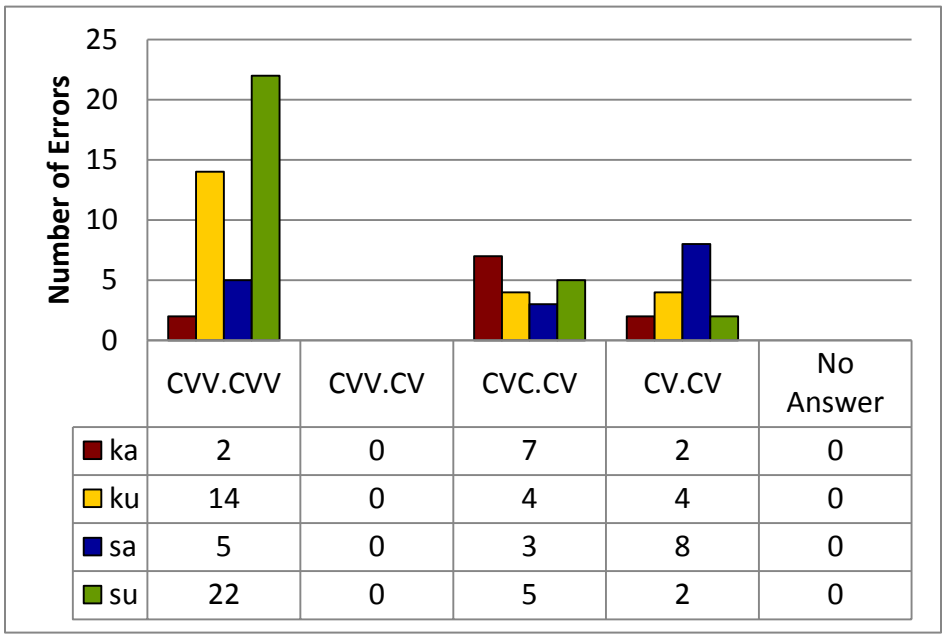
Figure 17: Errors for the tokens CVV.CV with the (4) LH.H pitch pattern



	CVV.CVV	CVV.CV	CVC.CV	CV.CV	No Answer
ka	4	0	2	1	0
ku	1	0	2	0	0
sa	3	0	2	0	1
su	8	0	0	2	0

Second, Figure 18 shows the number of errors that the participants made for the tokens with the (5) HL.L pitch pattern. There were 78 errors in total; the majority, approximately 55.12% of the errors were observed for the choice CVV.CVV, similar to errors for pitch pattern LH.H.

Figure 18: Errors for the tokens CVV.CV with the (5) HL.L pitch pattern



	CVV.CVV	CVV.CV	CVC.CV	CV.CV	No Answer
ka	2	0	7	2	0
ku	14	0	4	4	0
sa	5	0	3	8	0
su	22	0	5	2	0

The error analysis also indicates that the HL.L pitch pattern as shown in Figure 18 revealed more errors for the tokens with the vowel /u/ (i.e., a total of 36) than those with /a/ (i.e., a total of 7). In addition, more errors were observed for the tokens with HL.L pitch (i.e., a total of 16) as shown in Figure 17 than LH.H pitch (i.e., a total of 43) as shown in Figure 18.

Next, a three-way ANOVA was used to test whether preceding consonant, vowel type, and/or pitch pattern in Group III (CV.CVV) significantly affected the correct identification of vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s), vowel type (2; /a/ and /u/), and pitch pattern (3: L.HH, L.HL, H.LL). The dependent variable was perception accuracy. Results indicated significant main effects of vowel type, $F_{\text{Vowel}}(1, 63) = 5.154$, $p = .027$, $\eta_p^2 = .076$, and pitch pattern, $F_{\text{Pitch}}(2, 126) = 5.586$, $p = .005$, $\eta_p^2 = .081$; however, preceding consonant was not significant, $F_{\text{PreC}}(1, 63) = 1.595$, $p = .211$. Based on the

findings, the type of vowel and pitch pattern affected the identification of L2 vowel duration of the tokens in Group III. The mean accuracy scores of the tokens with the vowel /a/ and /u/ were .79 and .73 respectively. Therefore, it was easier for the learners to identify vowel duration when the vowel was /a/, compared to /u/. In order to locate where the differences existed in the three pitch patterns, pairwise comparisons were performed using the Bonferroni correction. Results indicated that (6) L.HH was significantly different from (8) H.LL ($p < .01$). The pitch pattern (6) L.HH was significantly easier than (8) H.LL. The two pitch patterns were very distinct; (8) L.HH starts with low pitch and remains high after the second mora; (10) H.LL is the opposite pattern.

In addition to the main effects, the following interactions were significant: the Vowel Type x Pitch Pattern interaction, $F(2, 126) = 4.759$, $p = .01$, $\eta_p^2 = .070$, the Preceding Consonant x Pitch Pattern interaction, $F(2, 126) = 3.759$, $p = .026$, $\eta_p^2 = .056$, and the Preceding Consonant x Vowel Type x Pitch Pattern interaction, $F(2, 126) = 18.990$, $p < .001$, $\eta_p^2 = .232$. In order to analyze the three-way interaction, a simple effects test was conducted. Based on the results, it was found that the LH.L pitch pattern had higher accuracy with the vowel and consonant combination of /ka/, than the other consonant-vowel combinations such as /ku/, /sa/, and /su/.

Error analysis was conducted on the responses to the CV.CVV tokens. Table 11 below shows the four choices used in the identification task for the tokens with the CV.CVV structure. Among these choices, (C) was the correct response; (A) was different from (C) in terms of the vowel length in the first syllable, (B) was different in terms of the vowel length in both the first and second syllable, and (D) was different in terms of the vowel length on the second syllable.

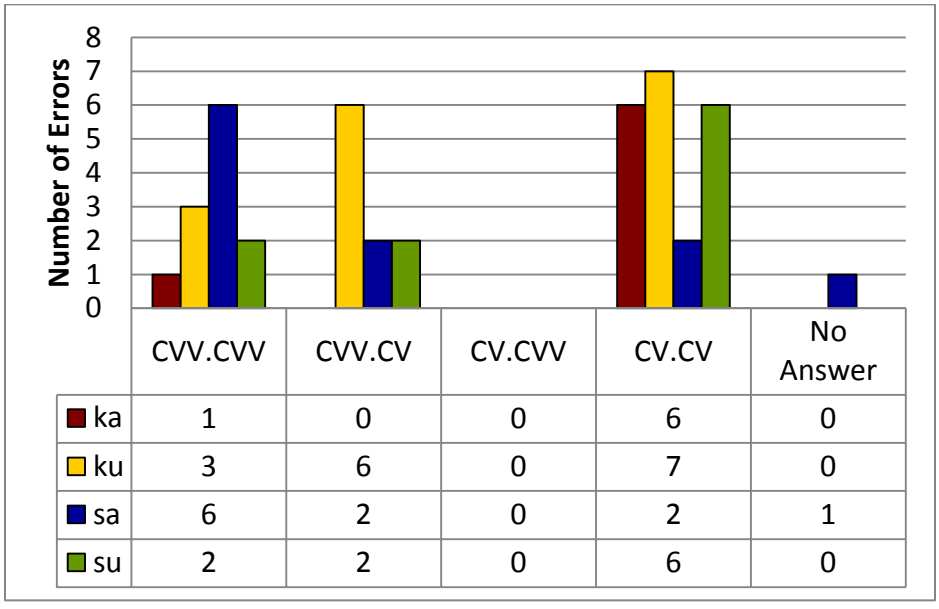
Table 11: An example of choices used in the identification task for CV.CVV tokens

Choices in the Identification Task	Examples	Selection would indicate:
A. CVV.CVV	kaa.kaa	-misperception of vowel length in first syllable
B. CVV.CV	kaa.ka	-misperception of vowel length in both syllables
C. CV.CVV	ka.kaa	-correct
D. CV.CV	ka.ka	-misperception of vowel length in second syllable

Note: Syllable boundaries were not marked in the experiment.

Figure 19 shows the number of errors that the participants made for the CV.CVV tokens with the (6) L.HH pitch pattern. There were 44 errors in total; approximately 47.73% of the errors were observed for the choice CV.CV; approximately 27.27% of the errors were observed for CVV.CVV; and approximately 22.72% of the errors were observed for CVV.CV. The majority was observed for the choice CV.CV (misperception of vowel length in the second syllable).

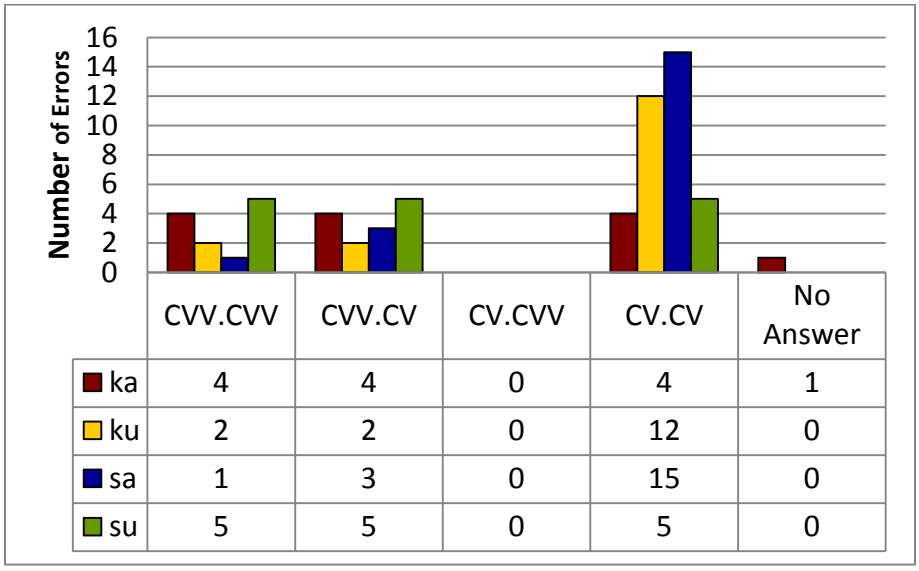
Figure 19: Errors for the CV.CVV tokens with the (6) L.HH pitch pattern



	CVV.CVV	CVV.CV	CV.CVV	CV.CV	No Answer
ka	1	0	0	6	0
ku	3	6	0	7	0
sa	6	2	0	2	1
su	2	2	0	6	0

Next, Figure 20 shows the number of errors that the participants made for the CV.CVV tokens with the (7) L.HL pitch pattern. There were 63 errors in total; approximately 57.14% of the errors were observed for the choice CV.CV (misperception of the vowel length in the second syllable).

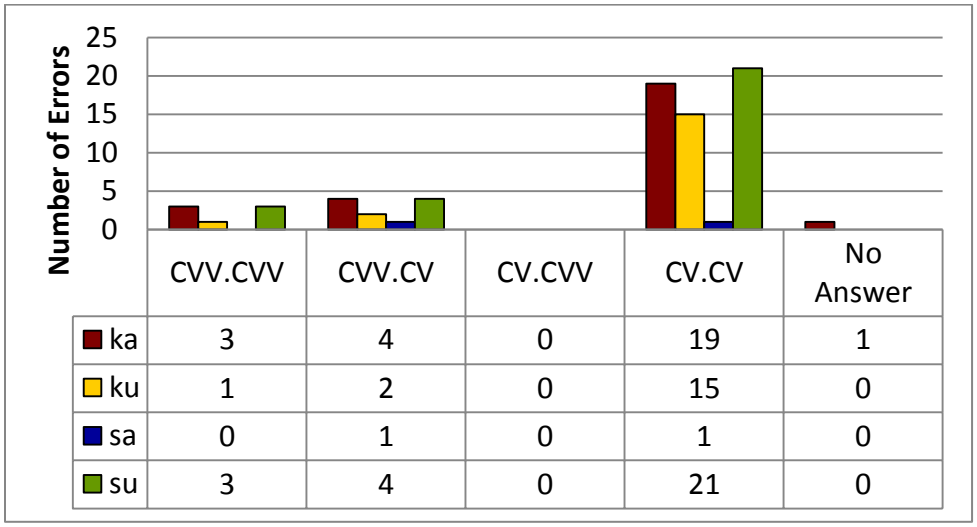
Figure 20: Errors for the CV.CVV tokens with the (7) L.HL pitch pattern



	CVV.CVV	CVV.CV	CV.CVV	CV.CV	No Answer
ka	4	4	0	4	1
ku	2	2	0	12	0
sa	1	3	0	15	0
su	5	5	0	5	0

Finally, Figure 21 shows the number of errors that the participants made for the CV.CVV tokens with the (8) H.LL pitch pattern. There were 75 errors in total; approximately 74.67% of the errors were observed for the choice CV.CV, similar to the other two patterns.

Figure 21: Errors for the CV.CVV tokens with the (8) H.LL pitch pattern



	CVV.CVV	CVV.CV	CV.CVV	CV.CV	No Answer
ka	3	4	0	19	1
ku	1	2	0	15	0
sa	0	1	0	1	0
su	3	4	0	21	0

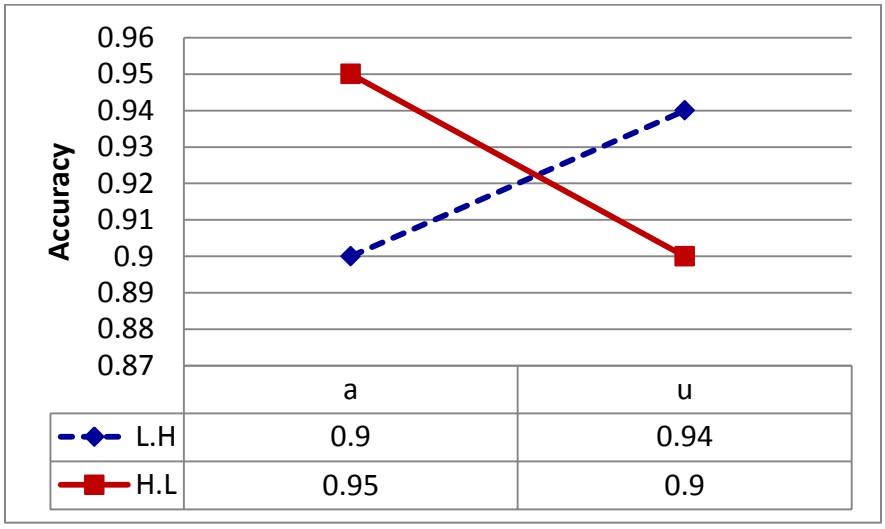
The error analysis revealed that the majority of the learners incorrectly perceived the CV.CVV tokens as CV.CV. In other words, the learners misperceived a long vowel on the second syllable as a short vowel.

Finally, a three-way ANOVA was used to test whether preceding consonant, vowel type, and/or pitch pattern in Group IV (CV.CV) significantly affected the correct identification of vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s), vowel type (2; /a/ and /u/), and pitch pattern (2: L.H, H.L). The dependent variable was perception accuracy. Results indicated significant main effects of preceding consonant, $F_{\text{PreC}}(1, 63) = 9.061$, $p = .004$, $\eta_p^2 = .126$; however, vowel type, $F_{\text{PreC}}(1, 63) = .047$, $p = .829$, and pitch pattern, $F_{\text{Pitch}}(1, 63) = .034$, $p = .854$, were not significant. Based on the findings, the preceding consonant affected the identification of L2 vowel duration of the tokens in Group IV (CV.CV). The mean accuracy scores of the tokens with the consonant /k/ and /s/ were .90 and .95

respectively. Therefore, it was easier for the learners to identify vowel duration when the preceding consonant was /s/, compared to /k/.

In addition to the main effects, the Vowel Type x Pitch Pattern interaction was significant, $F(1, 63) = 5.154$, $p = .027$, $\eta_p^2 = .076$ (Figure 22). Simple effects tests were conducted, and the results revealed that the accuracy of the L.H pitch with the vowel /u/ was higher than that with the vowel /a/. The figure suggests that the biggest difference is greater accuracy for /a/ with H.L versus L.H.

Figure 22: Effects of vowel type and pitch pattern in Group IV in Experiment 1



Error analysis was conducted on the responses to the CV.CV tokens. Table 12 below shows the four choices used in the identification task for the tokens with the CV.CV pitch pattern. Among the four choices, (D) was the correct response; (A) was different from (D) in terms of the vowel length in the first syllable, (B) was different because of the geminate, and (D) was different in terms of the vowel length of the second syllable.

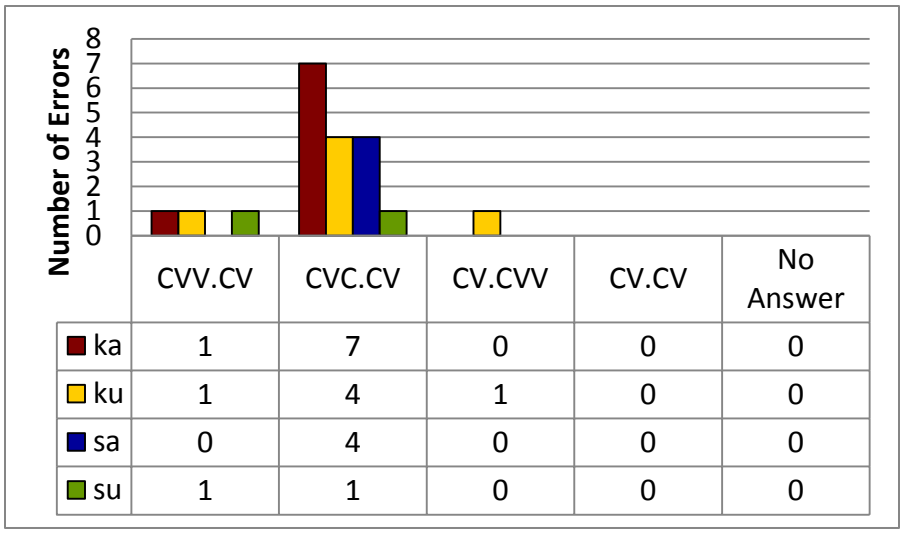
Table 12: An example of choices used in the identification task for CV.CV tokens

Choices in the Identification Task	Examples	Selection would indicate:
A. CVV.CV	kaa.ka	-misperception of vowel length in first syllable
B. CVC.CV	kak.ka	-misperception as a geminate
C. CV.CVV	ka.kaa	-misperception of vowel length in second syllable
D. CV.CV	ka.ka	-correct

Note: Syllable boundaries were not marked in the experiment.

Figure 23 shows the number of errors that the participants made for the CV.CV tokens with the (9) L.H pitch pattern. There were 20 errors in total; approximately 80% of the errors were observed for the choice CVC.CV (misperception as a geminate).

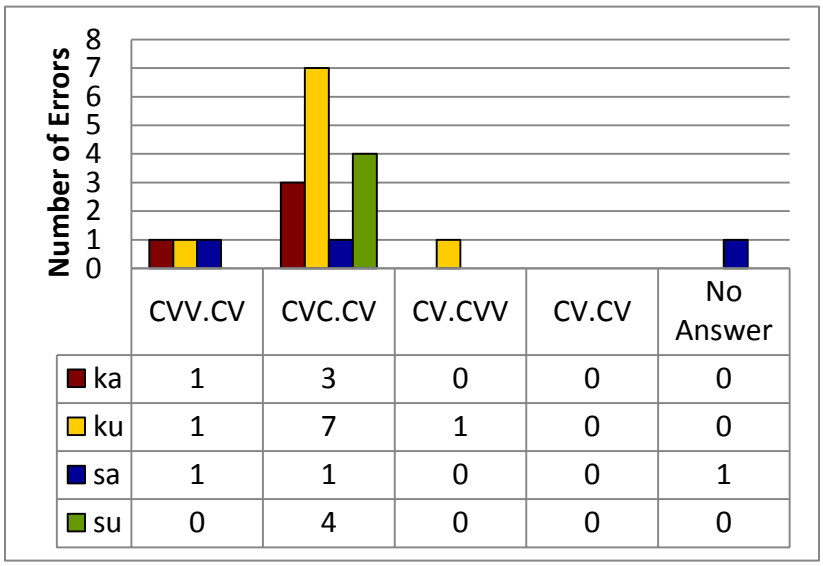
Figure 23: Errors for the CV.CV tokens with the (9) L.H pitch pattern



	CVV.CV	CVC.CV	CV.CVV	CV.CV	No Answer
ka	1	7	0	0	0
ku	1	4	1	0	0
sa	0	4	0	0	0
su	1	1	0	0	0

Figure 24 shows the number of errors that the participants made for the CV.CV tokens with the (10) H.L pitch pattern. There were 20 errors in total; approximately 75% of the errors were observed for the choice CVC.CV, following the data for the L.H pattern.

Figure 24: Errors for the CV.CV tokens with the (10) H.L pitch pattern



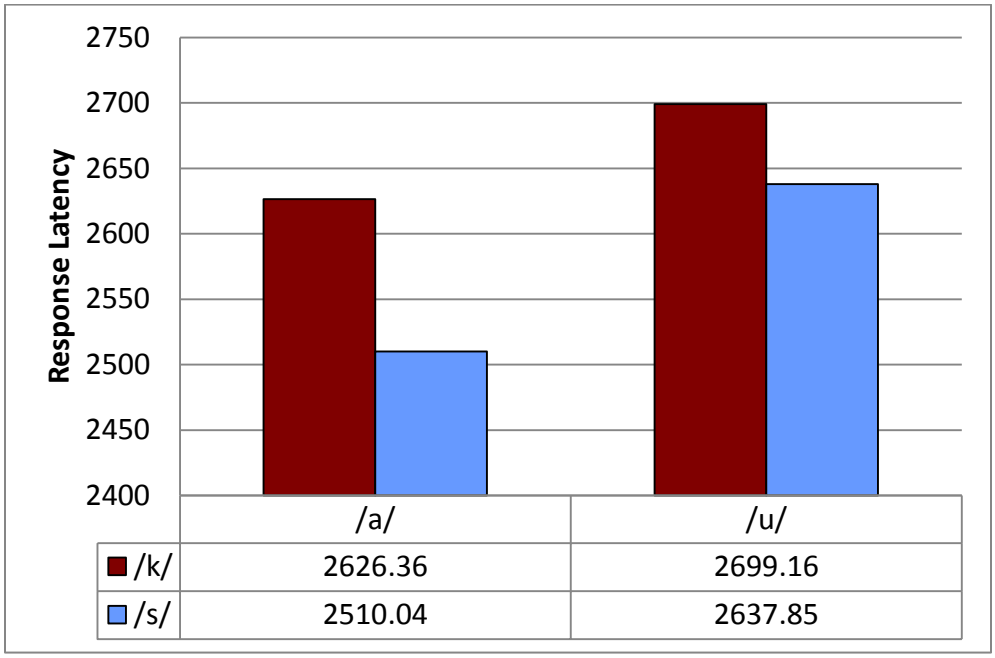
	CVV.CV	CVC.CV	CV.CVV	CV.CV	No Answer
ka	1	3	0	0	0
ku	1	7	1	0	0
sa	1	1	0	0	1
su	0	4	0	0	0

The error analysis revealed that the majority of the learners incorrectly perceived the CV.CV token as CVC.CV. This error pattern suggested that the perception of duration in the first syllable was misperceived as a geminate. For the L.H pitch pattern, there were more errors on the tokens with the vowel /a/; however, for the H.L pitch pattern, there were more errors on the tokens with the vowel /u/. The simple effect tests also suggested that the accuracy was higher with the vowel /u/, compared to the vowel /a/.

Analysis of Factors Affecting Perception RT: As possible factors that could affect the perception RT, preceding consonant (2; stop /k/, fricative /s/), vowel type (2; /a/, /u/), and pitch pattern (10 patterns) were examined. The overall mean of perception RT was 2652.42 milliseconds (*s.d.* 429.11). The mean identification RT for words with the preceding consonant /k/ and /s/ was 2662.76 milliseconds (*s.d.* 437.85) and 2573.95 milliseconds (*s.d.* 441.45) respectively. The mean identification RT for stimuli with the vowel /a/ and /u/ was 2568.20 milliseconds

(*s.d.*387.41) and 2668.51 milliseconds (*s.d.* 482.92) respectively. The mean RT for the CVcombinations /ka/, /ku/, /sa/, and /su/ was 2626.36 milliseconds (*s.d.* 482.01), 2699.16 milliseconds (*s.d.* 484.04), 2510.04 milliseconds (*s.d.* 404.78), and 2637.85 milliseconds (*s.d.* 551.65) respectively as in Figure 25.

Figure 25: Mean perception RTs by preceding consonant and vowel type

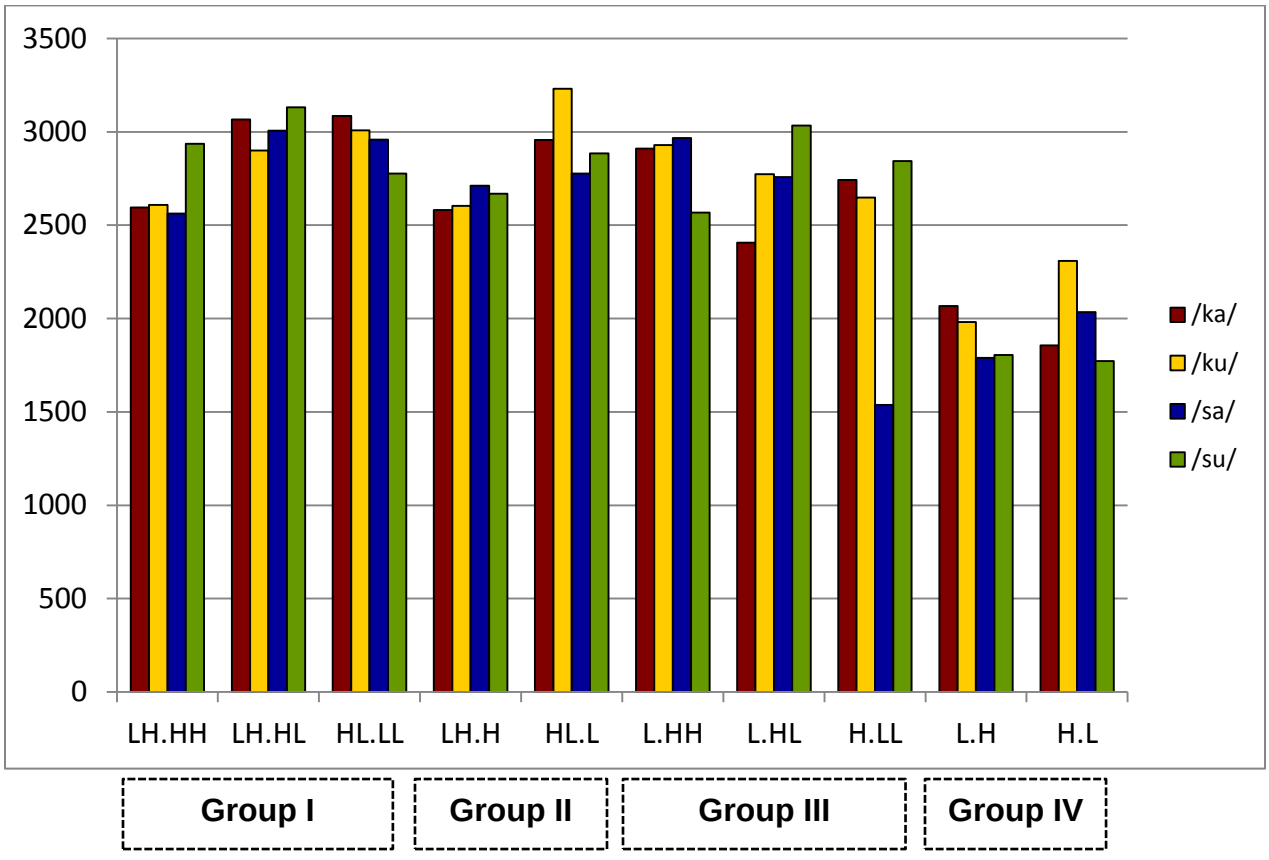


Then, 10 pitch patterns (1: LH.HH, 2: LH.HL, 3: HL.LL, 4: LH.H, 5: HL.L, 6: L.HH, 7: L.HL, 8: H.LL, 9: L.H, 10: H.L) were assigned to each combination of consonant and vowel: /ka/, /ku/, /sa/, and /su/. Table 13 and Figure 26 show the descriptive statistics for the perception accuracy by pitch pattern, preceding consonant, and vowel type.

Table 13: Descriptive Statistics for perception RT by pitch pattern and CV combination (in milliseconds)

Pitch	Preceding Consonant /k/				Preceding Consonant /s/			
	Vowel /a/		Vowel /u/		Vowel /a/		Vowel /u/	
	Mean	(*s.d.*)	Mean	(*s.d.*)	Mean	(*s.d.*)	Mean	(*s.d.*)
1	2594.05	(1105.82)	2608.91	(877.32)	2562.36	(1318.76)	2936.28	(1363.07)
2	3066.11	(1000.43)	2900.61	(956.43)	3006.17	(1013.56)	3130.66	(1075.40)
4	3084.25	(1129.87)	3007.59	(1040.13)	2958.61	(1262.16	2776.81	(1152.75)
5	2581.39	(703.08)	2602.78	(1010.40)	2712.05	(916.47)	2669.13	(782.26)
7	2956.86	(661.15)	3230.97	(993.23)	2776.03	(1050.64)	2884.75	(1020.95)
8	2909.63	(963.02)	2928.56	(965.29)	2966.53	(1004.49)	2567.23	(877.07)
9	2405.42	(956.37)	2772.88	(784.10)	2757.14	(920.75)	3033.09	(1143.37)
10	2742.77	(1137.97)	2648.70	(879.31)	1537.41	(603.29)	2843.41	(1039.80)
11	2067.11	(834.36)	1981.69	(692.84)	1789.23	(748.91)	1804.94	(779.67)
12	1855.98	(903.65)	2308.94	(979.32)	2034.88	(749.48)	1772.22	(858.63)

Figure 26: Mean RTs by pitch pattern, consonant, and vowel combination (in milliseconds)



As Figure 26 shows, the RT is shortest for the pitch pattern L.H and H.L (CV.CV). Also, when the token has high pitch at the end such as LH.HH and LH.H, the RT tends to be shorter than the other patterns such as LH.HL and HL.L respectively.

It was examined whether preceding consonant, vowel type, and/or pitch pattern significantly affected the response latency for the identification of vowel duration in Japanese. In order to examine the effects in detail, the 10 pitch patterns were divided into 4 categories (Group I, II, III, and IV) as indicated in Figure 26, according to the location of the long vowels as described earlier.

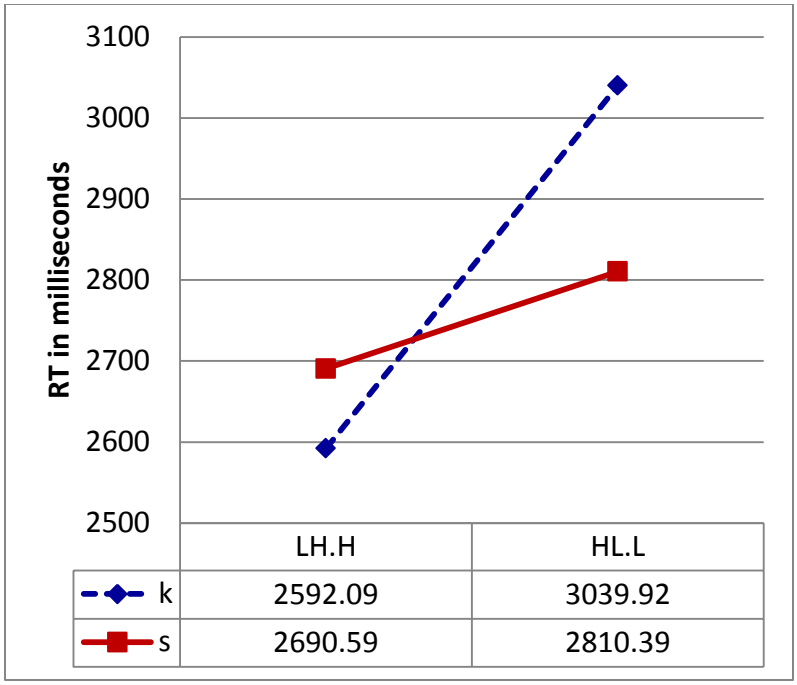
A three-way ANOVA was used to test whether preceding consonant, vowel type, and/or pitch pattern in Group I (CVV.CVV) significantly affected the RT in identifying vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s/), vowel type (2; /a/ and /u/), and pitch pattern (3; LH.HH, LH.HL, HL.LL). The dependent variable was perception RT. Results indicated significant main effects of pitch pattern, $F_{\text{Pitch}}(2, 126) = 7.884$, $p = .001$, $\eta_{\text{p}}^2 = .111$; however, vowel type, $F_{\text{Vowel}}(1, 63) = .046$, $p = .810$, and preceding consonant, $F_{\text{PreC}}(1, 63) = .058$, $p = 831$, were not significant. None of the interactions was significant.

It was found that the four pitch patterns significantly affected the RT to perceive vowel length. In order to locate where the differences existed in the four pitch patterns, pairwise comparisons were performed using the Bonferroni correction. Results indicated that (1) LH.HH was significantly different from (2) LH.HL ($p < .01$) as well as (3) HL.LL ($p < .01$). The mean RT for LH.HH was 2675.40 milliseconds, for LH.HL was 3025.89 milliseconds, and for HL.LL was 2956.82. Thus, the learners identified the vowel duration for the LH.HH pitch pattern more quickly than the other two patterns.

Next, a three-way ANOVA was used to test whether preceding consonant, vowel type, and/or pitch pattern in Group II (CVV.CV) significantly affected the RT in identifying vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s/), vowel type (2; /a/ and /u/), and pitch pattern (2; LH.H, HL.L). The dependent variable was perception RT. Results indicated significant main effects of pitch pattern, $F_{\text{Pitch}}(1, 63) = 16.853$, $p < .001$, $\eta_p^2 = .211$; however, preceding consonant, $F_{\text{PreC}}(1, 63) = 1.409$, $p = .240$, and vowel type, $F_{\text{Vowel}}(1, 63) = 1.098$, $p = .299$, were not significant. It was found that the two pitch patterns significantly affected the RT to perceive vowel length. The mean RT for (4) LH.H was 2641.34 milliseconds and that for (5) HL.L was 2952.15 milliseconds. Therefore, the learners identified the vowel duration for the token with the LH.H pattern faster than the ones with the HL.L pattern.

In addition to the main effects, the Preceding Consonant x Pitch Pattern interaction was significant, $F(1, 63) = 7.259$, $p = .099$, $\eta_p^2 = .103$. Simple effects tests were conducted, and as shown in Figure 27 results suggest that the RTs for /s/ vs. /k/ showed a greater difference with HL.L than LH.H.

Figure 27: Effects of preceding consonant and pitch pattern in Group II on RT in Experiment 1



Third, a three-way ANOVA was used to test whether preceding consonant, vowel type, and/or pitch pattern in Group III (CV.CVV) significantly affected the RT in identifying vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s/), vowel type (2; /a/ and /u/), and pitch pattern (3; L.HH, L.HL, H.LL). The dependent variable was perception RT. Results indicated significant main effects of pitch pattern, $F_{\text{Pitch}}(2, 126) = 12.120$ $p < .001$, $\eta_p^2 = .161$, and vowel type, $F_{\text{Vowel}}(1, 63) = 17.403$, $p < .001$, $\eta_p^2 = .216$; however, preceding consonant was not significant, $F_{\text{PreC}}(1, 63) = 3.025$, $p = .078$. It was found that the type of vowel significantly affected the response speed of L2 vowel duration. The mean response speed of the tokens with the vowel /a/ and /u/ were 2553.15 milliseconds and 2798.98 milliseconds respectively. Therefore, the L2 learners identified the vowel duration for the tokens with the vowel /a/ significantly faster than ones with the vowel /u/. It was also found that the

three pitch patterns significantly affected the RT to perceive vowel length. In order to locate where the differences existed in the three pitch patterns, pairwise comparisons were performed using the Bonferroni correction. Results indicated that (8) H.LL was significantly different from (6) L.HH ($p < .001$) as well as (7) L.HL ($p < .001$). The mean RT of (8) was 2443.07 milliseconds and was faster than that of (6) L.HH (2842.99 milliseconds) and (7) L.HL (2742.13 milliseconds). Thus, the L2 learners responded to the tokens with the H.LL pitch patterns faster than the other two pitch patterns.

In addition to the main effects, all of the interactions were significant: the Vowel Type x Pitch pattern, $F(2, 126) = 20.297$, $p < .001$, $\eta_p^2 = .244$, the Preceding Consonant x Vowel Type, $F(1, 63) = 4.391$, $p = .042$, $\eta_p^2 = .064$, the Preceding Consonant x Pitch Pattern, $F(2,126) = 20.587$, $p < .001$, $\eta_p^2 = .246$, and the Preceding Consonant x Vowel Type x Pitch Pattern, $F(2, 126) = 26.166$, $p < .001$, $\eta_p^2 = .293$. In order to analyze the three-way interaction in detail, simple effects tests were conducted. Basically, it was found that the token with the L.HL pitch pattern had a faster RT with /ka/ compared to /ku/.

Finally, a three-way ANOVA was used to test whether preceding consonant, vowel type, and/or pitch pattern in Group IV (CV.CV) significantly affected the RT in identifying vowel duration in Japanese. Independent variables were preceding consonant (2; /k/ and /s/), vowel type (2; /a/ and /u/), and pitch pattern (2; L.H and H.L). The dependent variable was perception RT. Results indicated significant main effects of preceding consonant, $F_{PreC}(1, 63) = 8.944$, $p = .004$, $\eta_p^2 = .124$; however, vowel type, $F_{Vowel}(1, 63) = .139$, $p = .710$, and pitch pattern, $F_{Pitch}(1, 63) = 1.928$, $p = .170$, were not significant. It was found that the preceding consonant

significantly affected the RT. The mean RTs for the token with the consonant /k/ and /s/ were 2053.43 milliseconds and 1850.31 milliseconds respectively. Thus, the learners could identify the vowel duration with the preceding consonant /s/ faster than the consonant /k/.

In addition to the main effects, the Preceding Consonant x Vowel Type interaction, $F(1, 63) = 5.271$, $p = .025$, $\eta_p^2 = .107$, and the Preceding Consonant x Vowel Type x Pitch Pattern interaction, $F(1, 63) = 9.704$, $p = .003$, $\eta_p^2 = .133$, were significant. The simple effects tests were conducted, and the results revealed that for the H.L pitch patterns the RT was shorter when the vowel and consonant combination was /su/ compared to /sa/.

Conclusion of Experiment 1

In conclusion, in this section, factors affecting accurate production, correct identification, and response latency of vowel duration were examined. Regarding the production of the vowel duration, it was found that vowel type and token type had significant main effects. In addition, a significant interaction between the preceding consonant and token type was found; the stop /k/ had higher accuracy than the fricative /s/ for the CV.CV token.

The important pattern that emerges from the perception accuracy data involves the influence of structural position of the long vowel, i.e., overall, there is misperception of vowel length in the second syllable regardless of pitch pattern. In the case of CV.CV, the pattern of errors suggests participants thought they perceived a longer duration but assigned it to a geminate.

In the next section, the data obtained from the perceptual training in Experiment 2 were analyzed. One of the objectives of Experiment 1 was to explore the potential effects of variables prior to the training study in Experiment 2. For the perception accuracy, perception latency, and

production accuracy, three variables (preceding consonant, vowel type, and pitch pattern) were analyzed and found to affect the identification of vowel duration. Therefore, the variables were included in the analysis of the data in Experiment 2.

CHAPTER 3: EXPERIMENT 2

Experiment 2 investigated the effects of auditory-visual input (i.e., waveform displays) and auditory-only in the training of L1 English speakers to identify L2 Japanese vowel duration.

Method

Participants

Participants were the same as in Experiment 1, which served as an exploratory study. A total of 12 participants received 90% or higher on the identification task; therefore, they were excluded from the study in order to avoid ceiling effects. The remaining 52 learners participated in Experiment 2.

Materials

Production Test: Materials included 16 tokens contrasting long and short vowels (Appendix A). High and low vowels /a, u/ and two consonants /s, k/ were used to construct the target stimuli. As with Experiment 1, pitch production was not treated as a variable in production.

Perception Test: Target stimuli for testing and treatment (i.e., perceptual training) were different. Out of 40 tokens used in Experiment 1, 18 with long and short vowels were used for testing in Experiment 2 (see Appendix E). A total of six NSs of Japanese (M=2; F=4) pronounced the stimuli as shown in Table 14; Talker 1 was used for the testing stimuli (same with Experiment 1), Talker 6 was used for the Test of Generalization 2 (TG2) which contained familiar stimuli produced by a novel talker, Talker 2 was used for perception training and TG1, which involved

novel stimuli produced by a familiar talker, and the remaining three talkers (Talker 3, 4, and 5) were used for training stimuli.

Table 14: Talker assignment for recording stimuli used in identification tasks

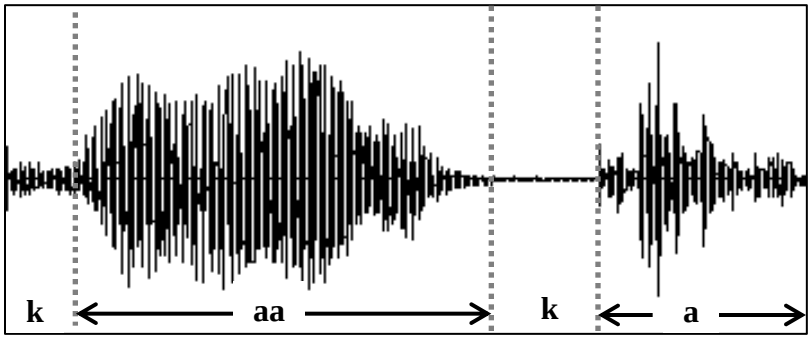
Talker	Gender	Experiment	Task
1	Female	Experiment 1	Perception Test
		Experiment 2	Perception Pretest and Posttest
2	Female	Experiment 2	Training 1 & 5
			TG1
3	Male	Experiment 2	Training 2 & 6
4	Male	Experiment 2	Training 3 & 7
5	Female	Experiment 2	Training 4 & 8
6	Female	Experiment 2	TG2

Notes: TG1: generalization test with novel tokens produced by a familiar talker
TG2: generalization test with familiar tokens produces by a new talker

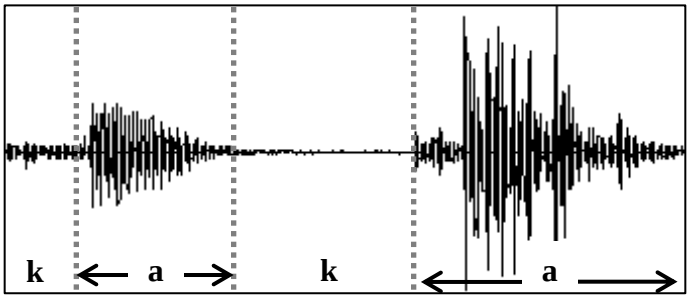
Perception Training: Out of 40 tokens used in Experiment 1, a total of 22 stimuli were used for the perceptual training (see Appendix F). The tokens were produced by four talkers as shown in Table 14 above. For both AV and A-only training conditions, the stimuli were audiorecorded using a digital voice recorder. For the AV condition, waveforms as shown in Figure 28 were generated using *Praat.*

Figure 28: Examples of the waveform displays

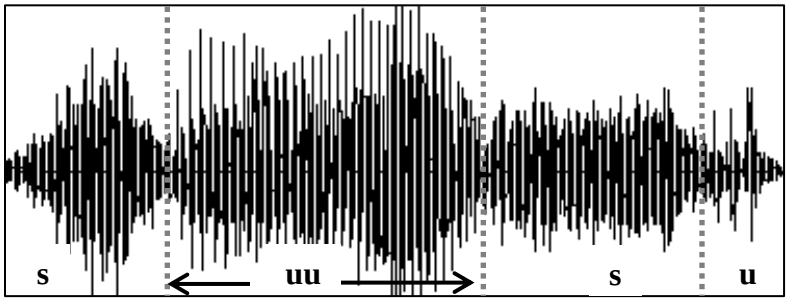
(a) *kaaka*



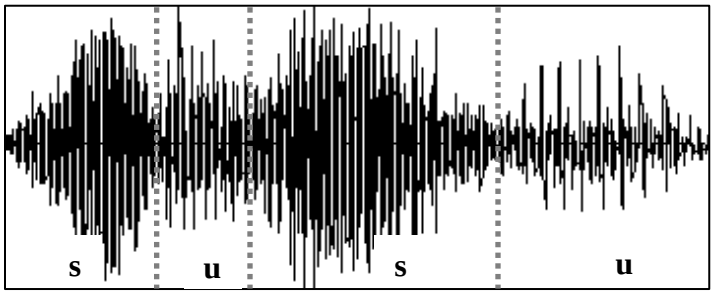
(b) *kaka*



(c) *suusu*



(d) *susu*



Procedures

Production Test: Computerized production test was created using *E-Prime*. The production test was administered prior to the perception test. The procedure was the same as in Experiment 1. During production testing, a visual prompt task of 16 tokens, listed in Appendix A, was given to participants. The stimuli were written in *roomaji* (i.e., the alphabet representation of Japanese sounds), not *hiragana*. The experiment was conducted in a quiet room. The participants' responses were recorded using a digital voice recorder and saved for later analyses.

Perception Test: After the production test, a perception test was given. Computerized perception test was created using *E-Prime*. During perception testing, participants were given a forced-choice, four-alternative identification task involving a total of 18 target stimuli (see Appendix E). The rationale for using the identification task rather than a discrimination task was based on previous studies (e.g., Logan et al., 1991). The choices were written in *romanization*, not *hiragana*. The procedure was the same as in Experiment 1. Identification accuracy, the participants' responses, and RTs were recorded on the computer and saved for later analysis.

In studies with a person's face as the AV input, testing stimuli are often presented in AV, A-only, and V-only conditions for the group that receives AV training (e.g., Hardison, 2003). The A-only test scores for the AV and A-only training groups are then compared since it is the only modality they share. However, in the current study, a waveform was used as the visual input, and it was not reasonable to test V-only accuracy. In addition, there was no rationale for AV testing because the waveform was essentially a training tool to facilitate the perception of duration. Therefore, in the current study, the two groups differed on the type of training, but were tested with A-only input.

Perception Training: Eight training sessions (approximately 25 minutes each), totaling approximately 3.5 hours in length, were administered individually, depending on participants' schedules. A forced-choice identification task was used. Prior to perception training, all the participants in the AV training group received waveform instruction for about five minutes, which included demonstration of how long and short vowels appeared in waveforms while listening to audio files. The purpose of this instruction was to "help learners understand the relation between the acoustic signal they [are] receiving and the electronic visual representation" (Motohashi-Saigo, 2007, p. 72). Five practice tokens were used in order to familiarize participants with the task (Appendix G). The participants in the AV training group listened to the stimulus and were asked to choose what they heard from the list provided while watching the associated waveform. On the other hand, the participants in the A-only group listened to the stimuli and were asked to choose from the options. Feedback was provided in the training, regardless of whether the responses were correct or not; the correct stimulus appeared as feedback on the computer screen after the participants selected their response. After receiving the feedback, participants in the A-only group had another chance to listen to each stimulus again. Participants in the AV group had another chance to listen to each stimulus again with the display of the associated waveform. The waveform was shown with the feedback so that the participants in the AV group could use the visual information to pay more attention to the form when their answers were wrong and the input type was always consistent with the type of training they were receiving. Their responses and RTs were recorded on the computer.

The detailed procedure for the perceptual training is described below. First, a participant read the instructions on the computer screen as in Figure 29 for the A-only training group and Figure 30 for the AV group.

Figure 29: Instructions for perceptual training for A-only training group

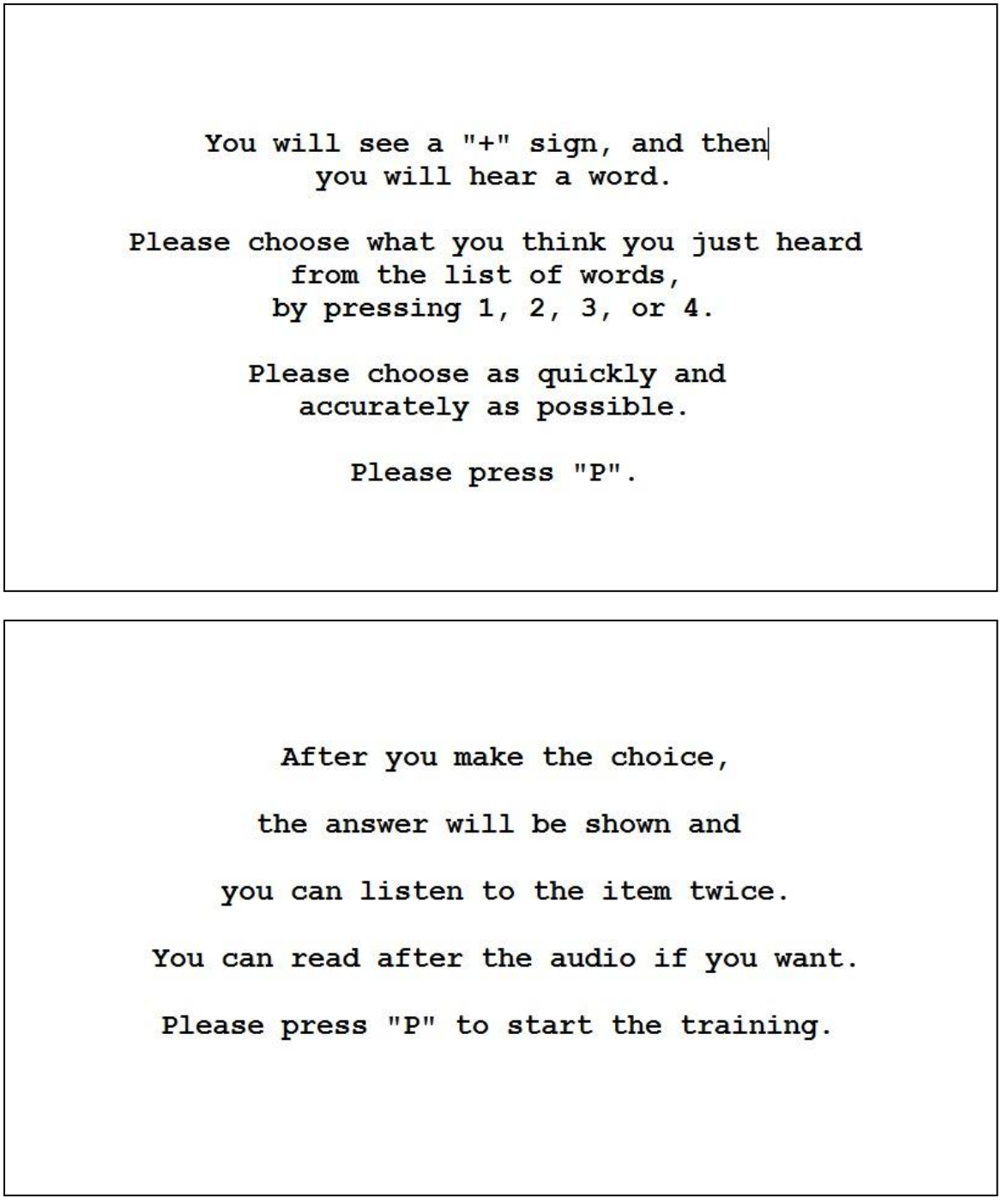
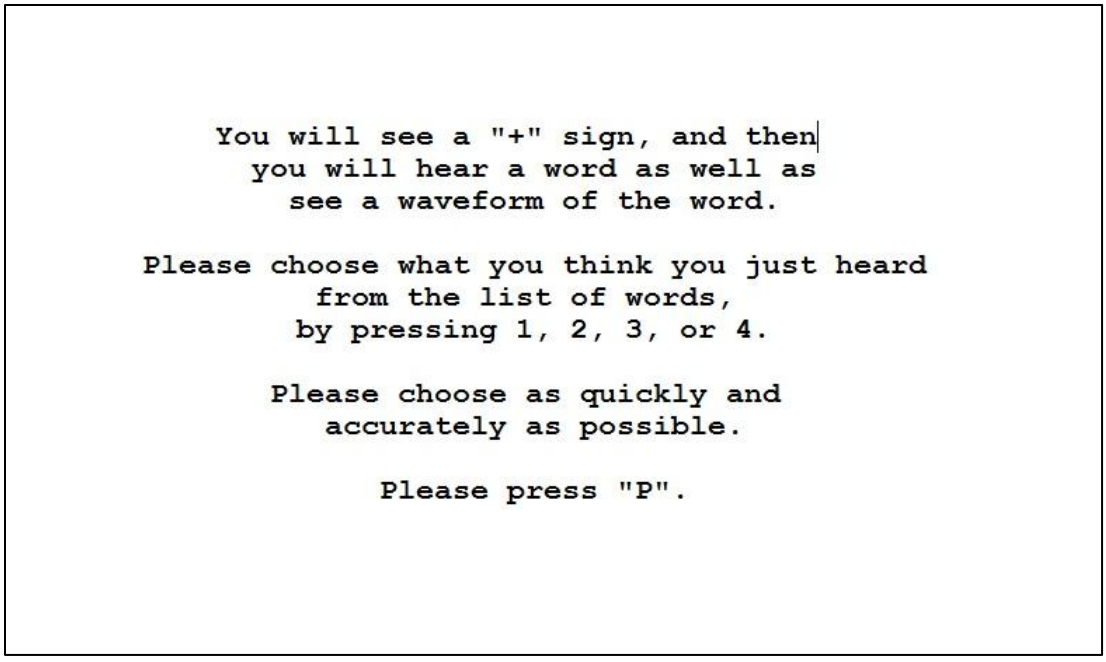


Figure 30: Instructions for perceptual training for AV training group



Then, the plus sign (+) appeared on the computer screen for four seconds before the participant listened to the stimulus presented as an isolated word. The participant listened to a stimulus and was asked to choose the correct response from the list provided as in Figure 31 for the A-only training group and Figure 32 for the AV training group.

Figure 31: Identification task for perceptual training for A-only training group

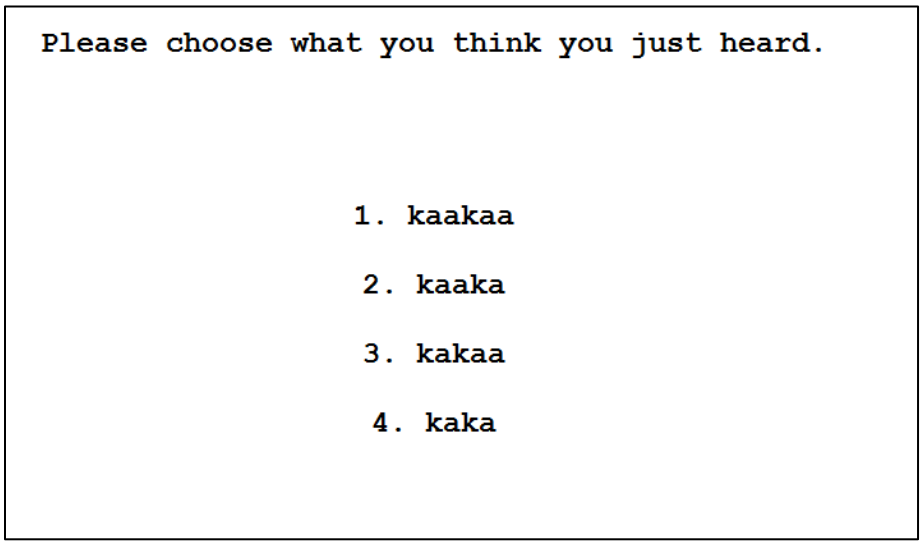
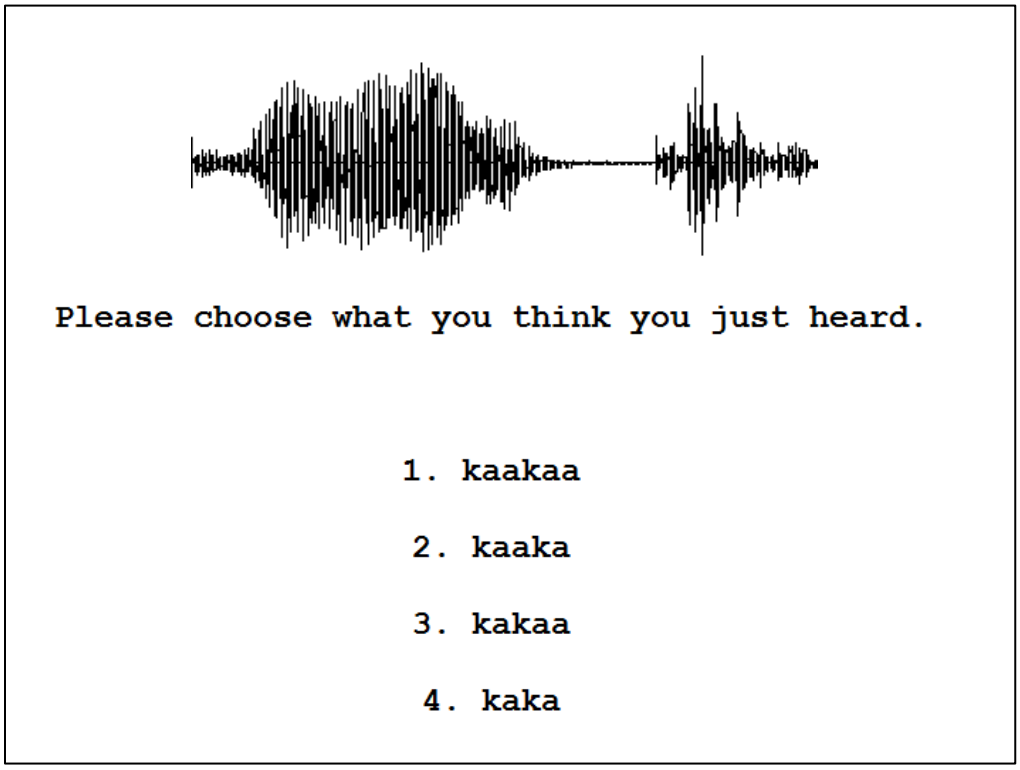


Figure 32: Identification task for perceptual training for AV training group



As shown in Figure 32, the waveform was also provided when the participants in the AV training group worked on the identification task. As soon as the participant made a choice, the computer screen showed a correct answer, and the participant listened to the stimulus again. As soon as the feedback was finished, the computer screen showed a plus sign again and continued the task for the rest of the stimuli.

Test of Generalization (TG): In order to see whether the participants' improvement in identifying vowel duration could be extended to novel stimuli produced by a familiar voice (TG1) and familiar stimuli (i.e., stimuli used in training sessions) produced by an unfamiliar voice (TG2), TGs that involve production and perception tests were given to the AV and A-only training groups. The novel stimuli for TG1 are listed in Appendix H; the familiar stimuli for TG2 were the same as the posttest in Appendix E. These involve a vowel not presented in testing or training /e/ and a new consonant /t/.

All the procedures and formats of the tests were the same as the pretest/posttest described earlier. A familiar and an unfamiliar voice were operationalized in the following way. A familiar voice was a talker who produced tokens for training; therefore, Talker 2 (female) produced the target stimuli for TG1. On the other hand, an unfamiliar voice was a talker who had not produced tokens for either training or testing; therefore, a new talker (Talker 6) produced target stimuli for TG2. Production accuracy, perception accuracy, and perception RT were compared with (1) the pretest data to see if there was a significant improvement for these new materials, and (2) the posttest data to see if the TG data were comparable and any improvement noted between pretest and posttest could be generalized.

Results

A total of 4 participants did not complete all the tasks in Experiment 2 (i.e., perceptual training, posttests, and TGs); therefore, their data were removed from the analysis. As a result, the data from the remaining 48 participants were used for the analysis for Experiment 2. The data were analyzed following Hardison (2003) and Motohashi-Saigo and Hardison (2009) and are presented in the following order: (1) comparability of groups at pretest, (2) overall effectiveness of perceptual training, (3) influence of stimulus variables on perception accuracy, (3) influence of stimulus variables on perception latency, (3) the effects of perceptual training on production, (4) the effect of training per group, and (5) tests of generalization. For the statistical analysis, the alpha level was set as .05 ($\alpha = .05$).

Comparability of Groups at Pretest: The 48 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). The mean accuracy scores in the pretest for the AV, A-only, and Control groups were 68.75% (*s.d.* 16.21), 71.78% (*s.d.* 14.94), and 65.97% (*s.d.* 16.84) respectively. The mean RT scores in the pretest for the AV, and A-only, and Control groups were 2856.63 milliseconds (*s.d.* 582.99), 2805.25 milliseconds (*s.d.* 515.29), and 2789.66 milliseconds (*s.d.* 410.47) respectively. In order to examine whether the three groups were statistically equivalent at the time of pretest, two one-way ANOVAs were performed. The independent variables for both were group type (AV, A-only, Control); dependent variables were perception accuracy and RT. The results of the ANOVAs confirmed that the three groups were statistically equivalent before perceptual training: $F_{Accuracy}(2, 47) = 0.424$, $p = .657$; $F_{RT}(2, 47) = 0.076$, $p = .927$.

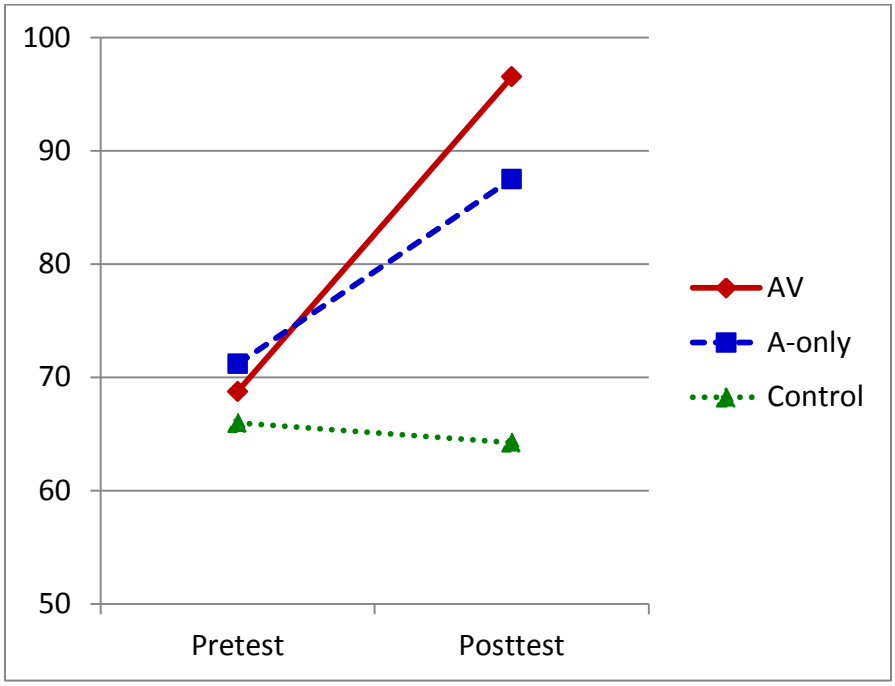
Analysis of Overall Effectiveness of the Perception Training: The descriptive statistics of perception accuracy and RT in the pretest and posttest for each training group are shown in Table 15 below.

Table 15: Descriptive statistics for the perception pre/post-tests per group

Group	Sample Size	Accuracy [Mean % (*s.d.*)]		RT [Mean in milliseconds (*s.d.*)]	
		Pretest	Posttest	Pretest	Posttest
AV	16	68.75 (16.21)	96.53 (8.58)	2856.63 (582.99)	3106.78 (530.25)
A-Only	16	71.18 (14.94)	87.50 (12.91)	2805.25 (515.29)	3179.96 (564.17)
Control	16	65.97 (16.84)	64.24 (19.98)	2789.66 (410.47)	3139.41 (520.75)

Mixed ANOVA was used to test whether the training itself was effective for improving the accuracy of identifying vowel duration and its response speed, compared to no training. The within-subject factor was time (2; pretest and posttest); the between-subject factor was group type (3; AV, A-Only, Control). The dependent variable was perception accuracy. Results indicated significant main effects of time, $F_{\text{Time}}(1, 45) = 68.275$, $p < .001$, $\eta_p^2 = .603$, and group type, $F_{\text{Group}}(2, 45) = 6.956$, $p = .002$, $\eta_p^2 = .236$. The Time x Training Modality interaction was also significant, $F(2, 45) = 25.271$, $p < .001$, $\eta_p^2 = .529$. In order to locate where the difference existed among the three groups, post-hoc comparison was performed using Tukey HSD. Results indicated that the control group was significantly different from the AV group ($p = .003$) and the A-only group ($p = .018$); however, there was no statistically significant difference between the two experimental groups ($p = .788$) (Figure 33) although overall accuracy increased more for the AV group.

Figure 33: The comparison of perception accuracy between pretest and posttest by group



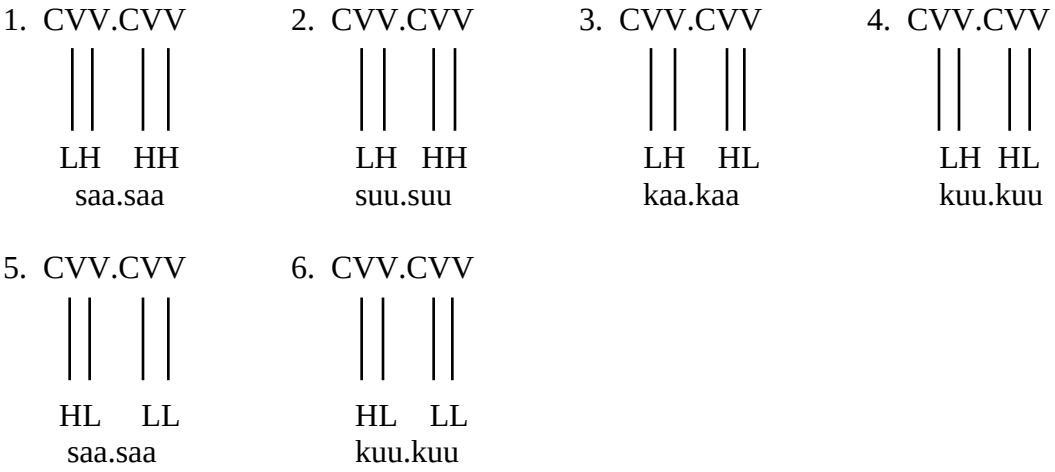
The purpose of having a control group was to determine if L2 learners could improve without training over the same period of time. The participants in the experimental groups spent two weeks receiving perceptual training. They also received regular classroom instruction during the training. Therefore, it was important to have the control group to show that the improvement from the pretest to posttest was due to the training. Since the control group did not improve, it was concluded that the improvement resulted from the training. Therefore, the control group was removed from further analyses.

Influence of Stimulus Variables on Perception Accuracy: In Experiment 1, it was found that there were several interactions involving pitch pattern, preceding consonant, and/or vowel type, which suggested that the combination of these factors affected perception accuracy of vowel length. Based on the results of Experiment 1 which indicated that the position of the long vowel

in the second syllable influenced perception accuracy, and to create a manageable stimulus set, the variables of consonant, vowel type, and pitch pattern were combined into 18 different stimulus types as shown in Figure 34. The pitch pattern groups I – III are the same as those in Experiment 1. No stimuli from Group IV (short vowels only, CV.CV) were used in the testing materials in Experiment 2.

Figure 34: Stimulus type in pretest and posttest in Experiment 2

Group I



Group II

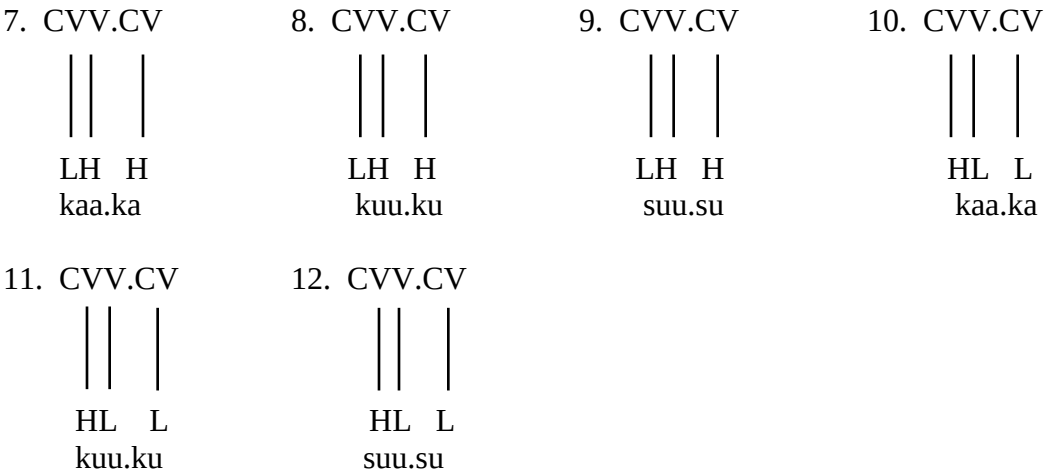
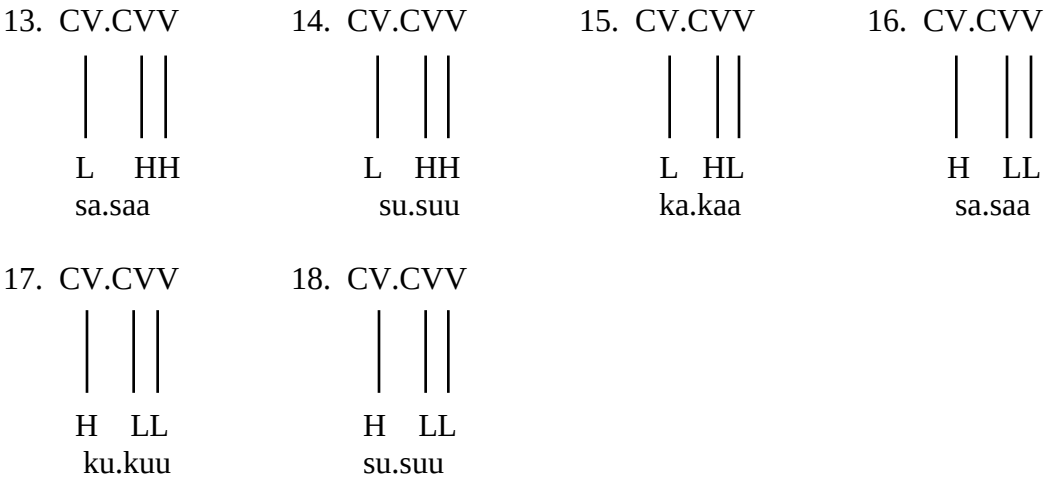


Figure 34 (cont'd)

Group III



First, a mixed ANOVA was used to test whether the effectiveness of the perceptual training varied depending on stimulus type within Group I. Within-subject factors were time (2; pretest and posttest) and stimulus type (6). The between-subject factor was group type (2; AV, A-Only). The dependent variable was perception accuracy. Results indicated significant main effects of time, $F_{\text{Time}}(1, 30) = 44.885$, $p < .001$, $\eta_p^2 = .599$, and stimulus type, $F_{\text{SType}}(5, 150) = 4.241$, $p = .001$, $\eta_p^2 = .124$; however, group type was not significant, $F_{\text{Group}}(1, 30) = .839$, $p = .367$. None of the interactions was significant.

The mean accuracy scores for the tokens at pretest and posttest were .62 and .91 respectively. Therefore, the perception accuracy of the stimuli in Group I improved from pretest to posttest. In addition, it was found that stimulus type had a significant influence. In order to locate where the differences existed among the six stimulus types, pairwise comparisons were

performed using the Bonferroni correction. The mean accuracy scores for each stimulus type are shown in Table 16 below.

Table 16: Mean perception accuracy of the six stimuli in Group I (CVV.CVV) in Experiment 2

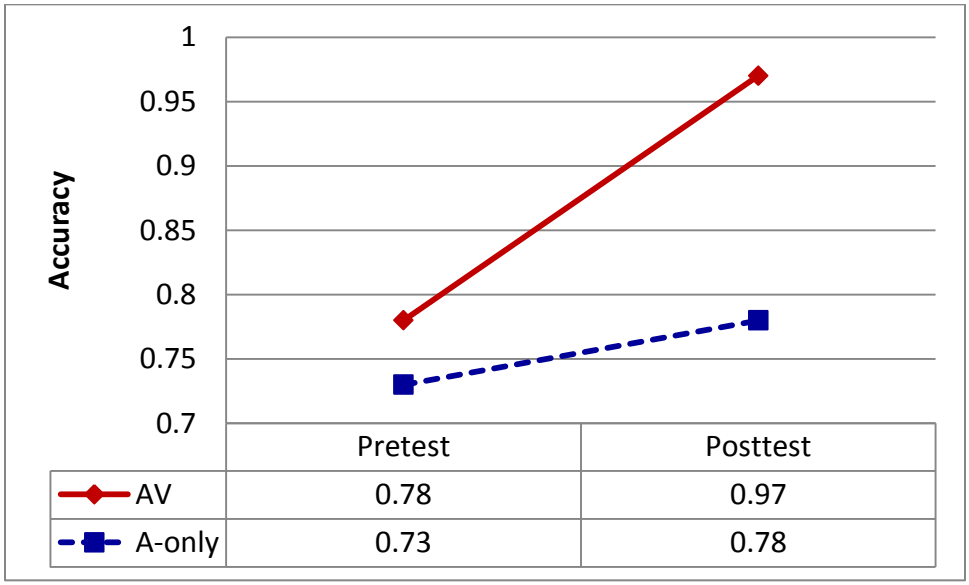
Stimulus Type (ST)	Tokens	Mean Accuracy	
		Pretest	Posttest
1	saa.saa (LH.HH)	.66	.97
2	suu.suu (LH.HH)	.81	.97
3	kaa.kaa (LH.HL)	.44	.81
4	kuu.kuu (LH.HL)	.75	.91
5	saa.saa (HL.LL)	.50	.91
6	kuu.kuu (HL.LL)	.56	.91

Results indicated that ST3 was significantly different from ST4 (p = .004) and ST2 (p = .007). The two tokens ST3 and ST 4 share the same preceding consonant and pitch pattern, but the vowel differs. Also, ST5 was significantly different from ST2 (p = .026). Based on the comparison of these two, the vowel /u/ combined with the consonant /k/ and the LH.HL pitch pattern was perceived more accurately than the vowel /a/ in the same condition.

Next, a mixed ANOVA was used to test whether the effectiveness of the perceptual training varied according to stimulus type within Group II. Within-subject factors were time (2; pretest and posttest) and stimulus type (6). The between-subject factor was group type (2; AV, A-Only). The dependent variable was perception accuracy. Results indicated significant main effects of time, $F_{\text{Time}}(1, 30) = 10.083$, $p = .003$, $\eta_p^2 = .252$, stimulus type, $F_{\text{SType}}(5, 150) = 10.156$, $p < .001$, $\eta_p^2 = .253$, and group type, $F_{\text{Group}}(1, 30) = 6.127$, $p = .019$, $\eta_p^2 = .170$, were all significant. However, none of the interactions was significant.

It was found that all of the factors affected perception accuracy. The mean accuracy scores for the tokens at pretest and posttest were .76 and .88 respectively. Therefore, the perception accuracy of the stimuli in Group II improved from pretest to posttest. In addition, group type had effects on the perception accuracy. Since the AV group had the higher accuracy than the A-only group, it was concluded that the AV training was more effective in developing perception accuracy of the tokens in Group II than the A-only group (Figure 35).

Figure 35: The comparison of perception accuracy of the tokens in Group II (CVV.CV) by training groups in Experiment 2



It was also found that stimulus type had a significant influence on correctly identifying the vowel duration. In order to locate where the differences existed in the six stimulus types, pairwise comparisons were performed using the Bonferroni correction. The mean accuracy scores for each stimulus type are tabulated in Table 17 below.

Table 17: Mean perception accuracy of the six stimulus type in Group II (CVV.CV) in Experiment 2

Stimulus Type (ST)	Tokens	Mean Accuracy	
		Pretest	Posttest
7	kaa.ka (LH.H)	.81	.91
8	kuu.ku (LH.H)	.94	.97
9	suu.su (LH.H)	.81	.81
10	kaa.ka (HL.L)	.91	.96
11	kuu.ku (HL.L)	.63	.88
12	suu.su (HL.L)	.44	.72

Results indicated that ST7 was significantly different from ST12 ($p = .015$); ST8 was different from ST11 ($p = .010$) and ST12 ($p < .001$); and ST10 was significantly different from ST11 ($p = .005$) and ST12 ($p < .001$). The difference between ST8 and ST11 was pitch pattern; therefore, it was concluded that the LH.H pattern was easier for correct perception than the HL.L pitch pattern when the token contains the preceding consonant /k/ and the vowel /u/. In addition, the difference between ST10 and ST11 was vowel type; therefore, it was concluded that the vowel /a/ was easier for correct perception than the vowel /u/ when it followed /k/ in the HL.L pattern.

Finally, a mixed ANOVA was used to test whether the effectiveness of the perceptual training varied according to stimulus type within Group III. Within-subject factors were time (2; pretest and posttest) and stimulus type (6). The between-subject factor was group type (2; AV, A-Only). The dependent variable was perception accuracy. Results indicated significant main effects of time, $F_{\text{Time}}(1, 30) = 24.083$, $p < .001$, $\eta_p^2 = .445$, and stimulus type, $F_{\text{SType}}(5, 150) = 7.358$, $p < .001$, $\eta_p^2 = .197$; however, group type was not significant, $F_{\text{Group}}(1, 30) = .309$, $p = .582$. The mean accuracy scores for the tokens at pretest and posttest were .73 and .91

respectively. Therefore, the perception accuracy of stimuli in Group III improved from pretest to posttest. It was also found that stimulus type had a significant influence on correctly identifying the vowel duration. In order to locate where the differences existed among the six stimulus types, pairwise comparisons were performed using the Bonferroni correction. The mean accuracy scores for each stimulus type are shown in Table 18 below.

Table 18: Mean perception accuracy of the six stimulus type in Group III (CV.CVV) in Experiment 2

Stimulus Type (ST)	Tokens	Mean Accuracy	
		Pretest	Posttest
13	sa.saa (L.HH)	.78	.97
14	su.suu (L.HL)	.81	.94
15	ka.kaa (H.LL)	.56	.88
16	sa.saa (H.LL)	1.00	.97
17	ku.kuu (H.LL)	.65	.88
18	su.suu (H.LL)	.56	.74

Results indicated that ST15 was significantly different from ST16 ($p < .001$); ST16 was different from ST17 ($p = .006$) and ST18 ($p = .001$). The difference between ST16 (sa.saa with H.LL) and ST18 (su.suu with H.LL) was the vowel. Considering the mean scores in Table 18, the vowel /a/ was easier for correct perception than the vowel /u/ when it contains the preceding consonant /s/ in the H.LL pattern. On the other hand, the difference among ST15, ST16 and ST17 was the combination of the consonant and vowel. Considering the mean scores in Table 18, the perception accuracy of /sa/ was higher than /ku/ and /ka/ when the tokens had the H.LL pitch.

In addition to the main effect of time and stimulus type, the Time x Group Type interaction, $F(1, 30) = 5.682$, $p = .024$, $\eta_p^2 = .159$, and the Time x Stimulus Type interaction, $F(5, 150) = 2.538$, $p = .031$, $\eta_p^2 = .159$, were significant for Group III (Figure 36). Based on the result, it was found that the rate of development in the perception accuracy was faster for the learners in the AV group, compared to those in the A-only group for CV.CVV stimuli. Regarding the interaction between the time and stimulus type, the results of the simple effects tests revealed that the differences between ST16 and ST13, ST17, and ST18 were greater in pretest than the posttest. In addition, ST16 revealed the highest accuracy; ST15 and ST18 revealed the lowest accuracy.

Figure 36: The comparison of perception accuracy of the tokens in Group III by training groups in Experiment 2

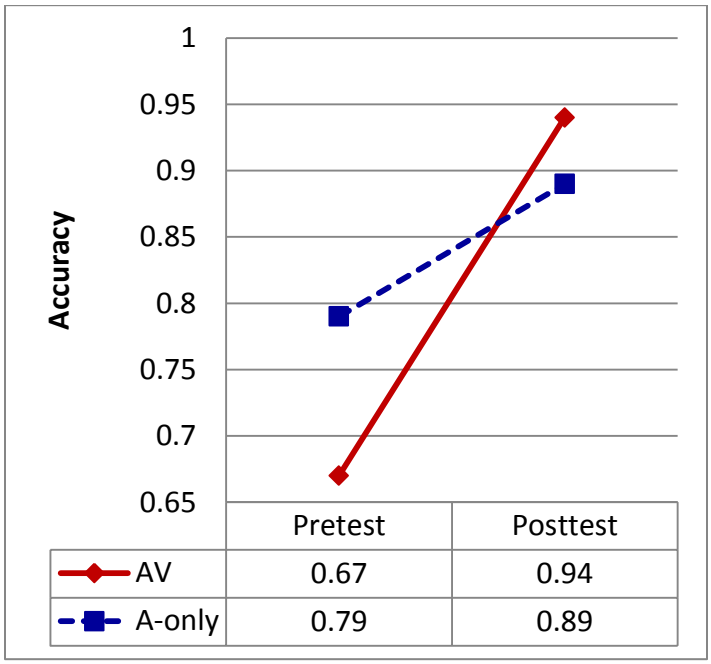
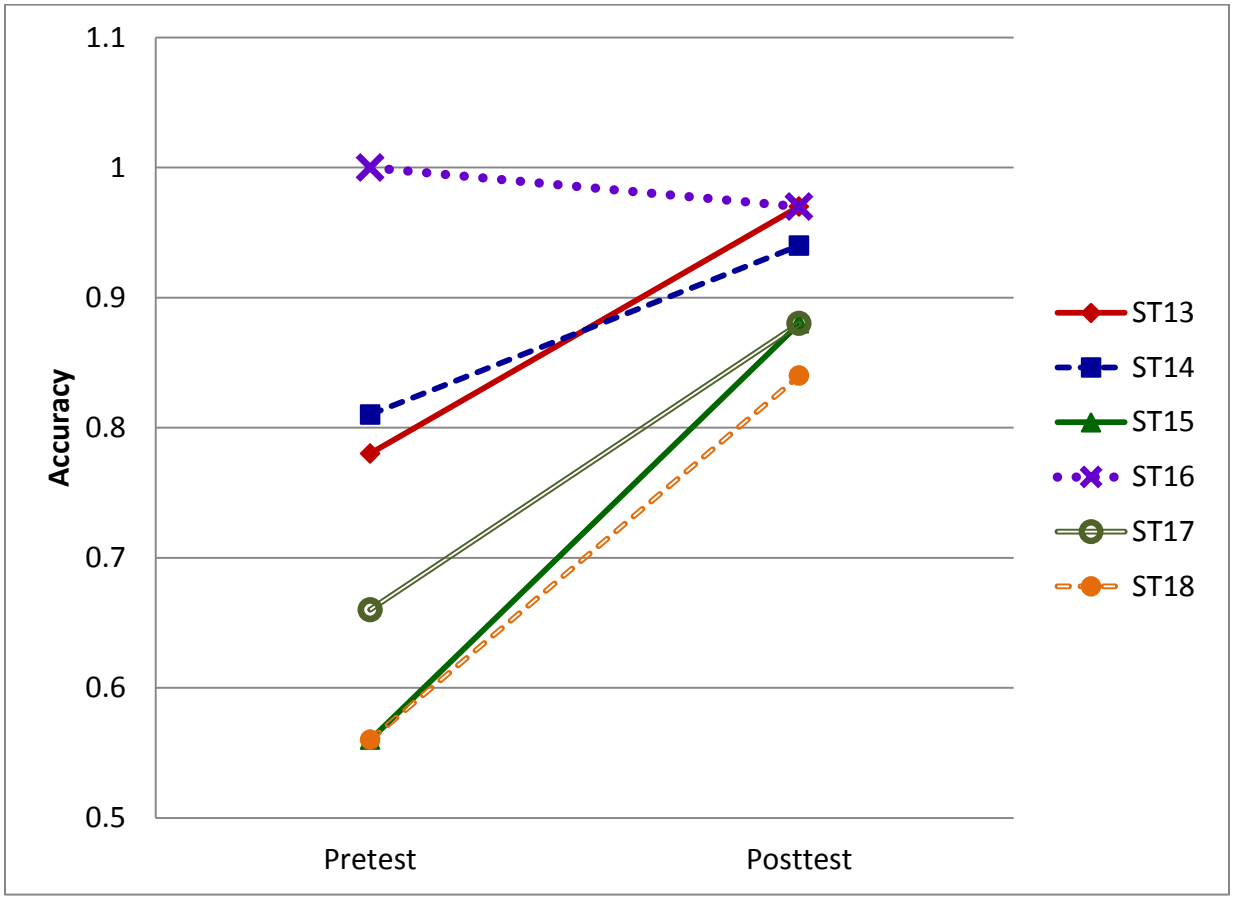


Figure 36 (cont'd)



Effectiveness of Training Type on Perception RT: It was examined whether the effectiveness of the perceptual training on perception RT varied with preceding consonant, vowel type, and pitch pattern. Similar to the analysis of perception accuracy, instead of having the three separate variables as preceding consonant, vowel type, and pitch pattern, they were combined and labeled as stimulus type in Figure 34 in the previous section. The stimulus type was divided into three groups as shown in Figure 34. Prior to the statistical analysis, it was confirmed that the AV and the A-only groups were statistically equivalent at the time of pretest.

First, a mixed ANOVA was used to test whether the effectiveness of the perceptual training on perception RT varied according to stimuli within Group I. Within-subject factors

were time (2; pretest and posttest) and stimulus type (6). The between-subject factor was group type (2; AV, A-Only). The dependent variable was perception RT. Results indicated no significant main effects: time, $F_{\text{Time}}(1, 30) = 3.198$, $p = .084$, and stimulus type, $F_{\text{SType}}(5, 150) = 1.121$, $p = .352$, and group type, $F_{\text{Group}}(1, 30) = 1.104$, $p = .302$. None of the interactions was significant. The mean RTs at pretest and posttest were 2926.07 milliseconds and 3207.77 milliseconds respectively. The pretest revealed faster RT than the posttest; however, the difference was not statistically significant. The mean RTs for each stimulus type are shown in Table 19 below. There were no significant differences among the six tokens.

Table 19: Mean perception RT of the six stimuli in Group I (CVV.CVV) in Experiment 2

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)	
		Pretest	Posttest
1	saa.saa (LH.HH)	2420.03	3301.38
2	suu.suu (LH.HH)	2940.72	3509.88
3	kaa.kaa (LH.HL)	3188.34	3201.00
4	kuu.kuu (LH.HL)	3041.44	3172.80
5	saa.saa (HL.LL)	3010.16	2945.94
6	kuu.kuu (HL.LL)	2955.75	3115.59

Next, a mixed ANOVA was used to test whether the effectiveness of the perceptual training on perception RT varied according to stimulus type within Group II. Within-subject factors were time (2; pretest and posttest) and stimulus type (6). The between-subject factor was group type (2; AV, A-Only). The dependent variable was perception RT. Results indicated significant main effects of time, $F_{\text{Time}}(1, 30) = 7.593$, $p = .010$, $\eta_p^2 = .202$; however, stimulus type, $F_{\text{SType}}(5, 150) = 1.278$, $p = .276$, and group type, $F_{\text{Group}}(1, 30) = 1.469$, $p = .235$, were

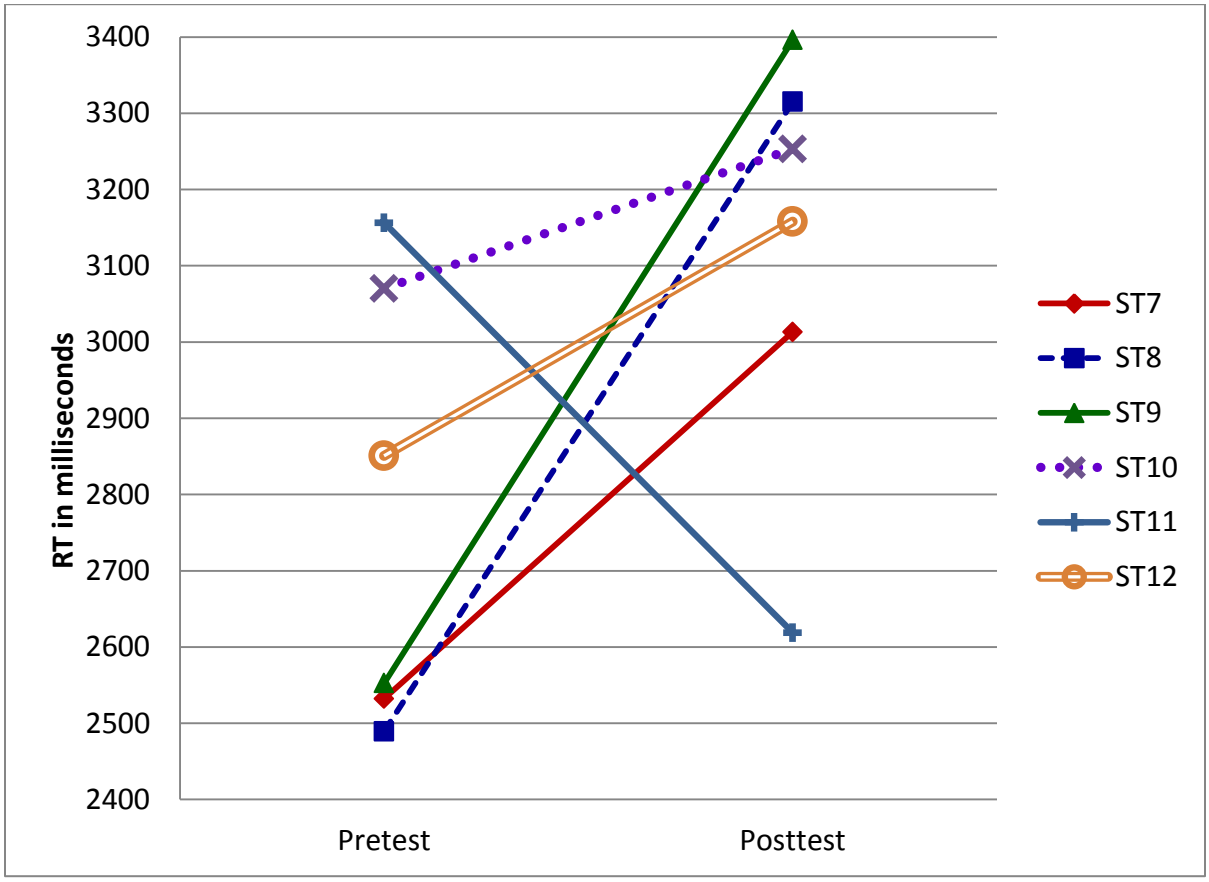
not significant. The mean RTs at pretest and posttest were 2775.23 milliseconds and 3125.75 milliseconds respectively. Therefore, the perception RT of the stimulus type in Group III (CV.CVV) increased at the posttest. The mean RTs for each stimulus type are shown in Table 20 below. There were no significant differences among the six tokens.

Table 20: Mean perception RT of the six stimulus type in Group II (CVV.CV) in Experiment 2

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)	
		Pretest	Posttest
7	kaa.ka (LH.H)	2532.31	3013.28
8	kuu.ku (LH.H)	2489.34	3315.13
9	suu.su (LH.H)	2552.63	3396.16
10	kaa.ka (HL.L)	3070.16	3253.22
11	kuu.ku (HL.L)	3156.28	2618.78
12	suu.su (HL.L)	2850.66	3157.94

Although there were no main effects, the Time x Stimulus Type interaction was found, $F_{\text{Time}}(5, 150) = 5.498$, $p < .001$, $\eta_p^2 = .155$ (Figure 37). In order to examine the interaction, the simple effects tests were conducted and the result revealed that the differences between ST11 and ST7, ST8, and ST11 were greater in the pretest than the posttest. The RT of ST11 was slower than the ST7, ST8, and ST9 in the pretest; however, that of ST11 became faster in the posttest while the RTs of the other three became slower.

Figure 37: The comparison of perception RT of the tokens in Group II by training groups in Experiment 2



Finally, a mixed ANOVA was used to test whether the effectiveness of the perceptual training on perception RT varied according to stimulus type within Group III. Within-subject factors were time (2; pretest and posttest) and stimulus type (6). The between-subject factor was training type (2; AV, A-Only). The dependent variable was perception RT. Results indicated significant main effects of time, $F_{\text{Time}}(1, 30) = 16.515$, $p < .001$, $\eta_p^2 = .355$, and stimulus type, $F_{\text{SType}}(5, 150) = 3.123$, $p = .010$, $\eta_p^2 = .094$; however, group type was not significant, $F_{\text{Group}}(1, 30) = .063$, $p = .804$. The mean perception RT at pretest and posttest were 2624.38 milliseconds

and 3249.11 milliseconds respectively. Therefore, the perception RT of stimuli in Group III increased from pretest to posttest. It was also found that stimulus type had a significant influence on response latency. In order to locate where the differences existed in the six stimulus types, pairwise comparisons were performed using the Bonferroni correction. The mean perception RTs are shown in Table 21 below.

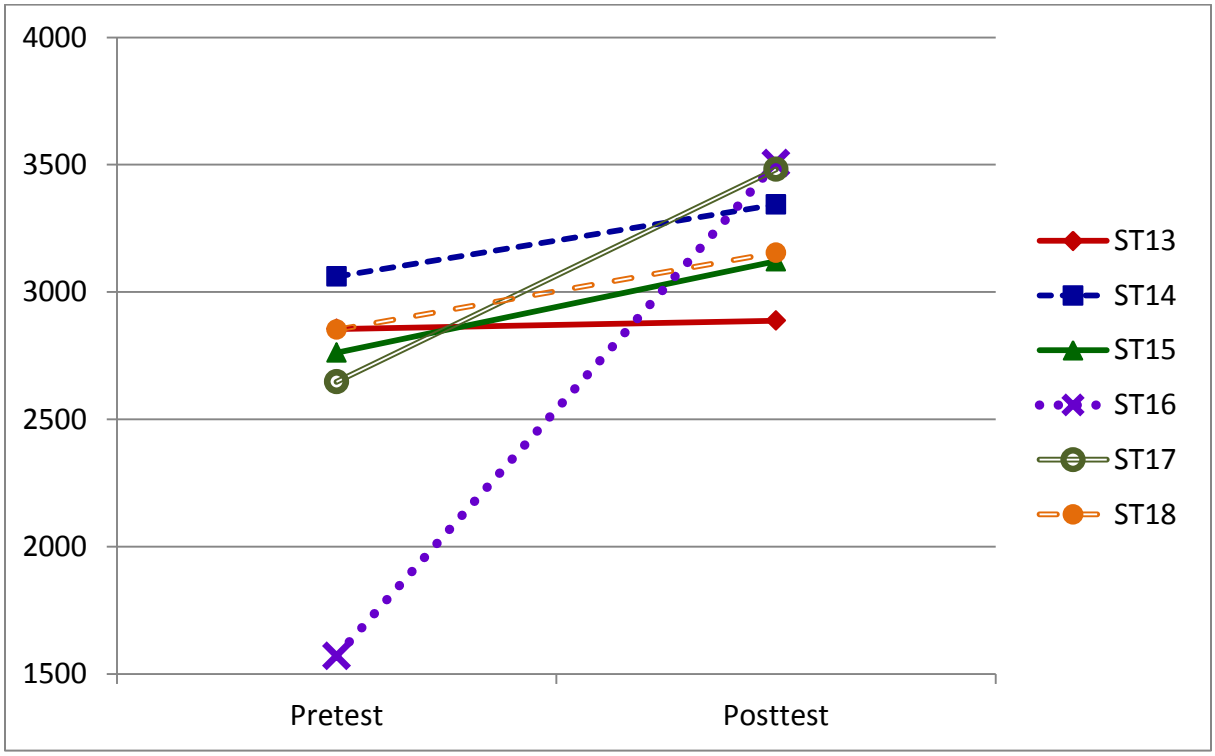
Table 21: Mean perception RT of the six stimulus type in Group III (CV.CVV) in Experiment 2

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)	
		Pretest	Posttest
13	sa.saa (L.HH)	2856.63	2887.28
14	su.suu (L.HL)	3060.66	3344.09
15	ka.kaa (H.LL)	2761.66	3120.56
16	sa.saa (H.LL)	1570.06	3505.66
17	ku.kuu (H.LL)	2647.31	3482.34
18	su.suu (H.LL)	2851.97	3154.72

Results indicated that ST14 was significantly different from ST16 ($p = .008$) and ST16 was significantly different from ST17 ($p = .026$). The difference between ST16 and ST17 was the combination of vowel and consonant; the perception of the long vowel in /sa/ was faster than that in /ku/ when the pitch pattern was H.LL.

In addition to the main effect of time and stimulus type, the Time x Stimulus Type interaction was significant, $F(5, 150) = 7.304$, $p < .001$, $\eta_p^2 = .196$ (Figure 38). In order to examine the interaction, simple effects tests were conducted and the result revealed that the difference between ST16 and ST17 was greater in the pretest than the posttest.

Figure 38: The comparison of perception RT of the tokens in Group III in Experiment 2



In conclusion, the two training groups improved accuracy in identifying vowel duration after the training. There were significant differences between the two types of training (AV vs. A-only); however, the AV group demonstrated grater improvement compared to the A-only group. There were mixed results regarding the influence of preceding consonant, vowel type, and pitch pattern. Depending on the token type (i.e., CVV.CVV, CVV.CV, and CV.CVV), influence of the variables was different. Although perception accuracy showed significant improvement after the training, response latency became slower, which suggested that the learners were processing the input more and thinking more about which choice provided in the identification task was right. In the analysis of training data, it was found that talker's voice affected both perception accuracy and latency.

Analysis of Production Data: The production accuracy before and after the perceptual training was analyzed to examine whether the efficiency of the training on correctly identifying vowel duration would transfer to another skill such as production. The 32 participants in the AV and A-only groups who took the perception training in Experiment 2 took a production pretest before the perception training and posttest after the training. The same raters who rated the pretest data rated the posttest data, using the same procedure. Interrater reliability was checked using Pearson Correlation/Coefficient. There was a significant positive correlation between Rater 1 and Rater 2 ($r = .914$, $p < .001$, $R^2 = .84$), between Rater 1 and Rater 3 ($r = .930$, $p < .001$, $R^2 = .86$), as well as between Rater 2 and Rater 3 ($r = .906$, $p < .001$, $R^2 = .82$); the correlation was strong. The production accuracy scores (i.e., one point for the correct pronunciation of each token) are shown in Table 22. The control group did not show improvement of production accuracy.

Table 22: Descriptive Statistics for production tests in Experiment 2 (Pretest and Posttest) for the AV and A-only groups organized by consonant-vowel combination

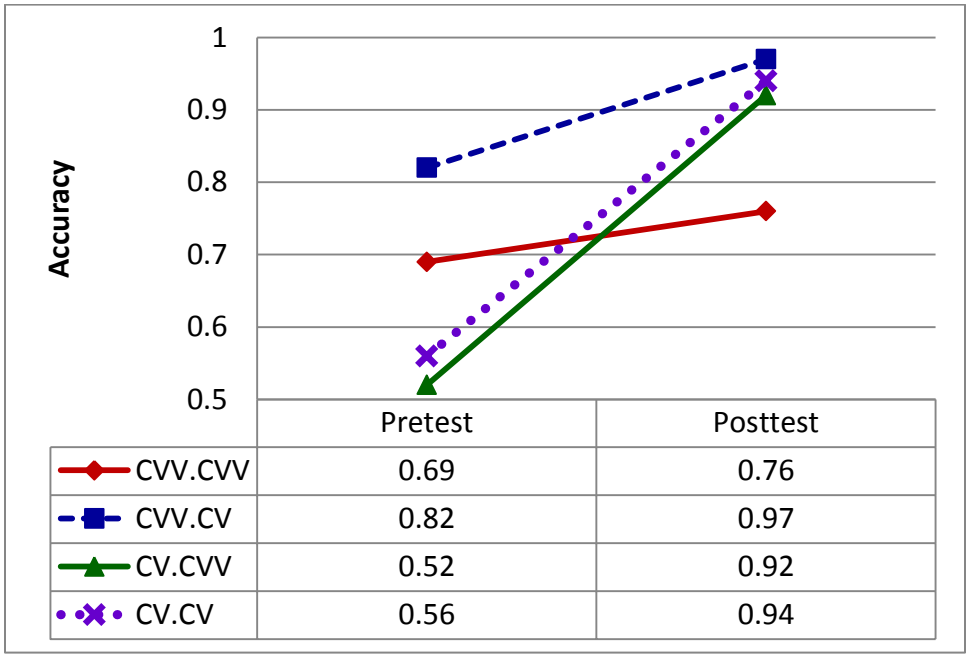
Tokens	AV Group		A-only Group	
	Pretest	Posttest	Pretest	Posttest
	Mean (*s.d.*)	Mean (*s.d.*)	Mean (*s.d.*)	Mean (*s.d.*)
kaa.kaa	.81 (.40)	.75 (.45)	.75 (.45)	.88 (.34)
kaa.ka	.75 (.45)	.88 (.94)	.93 (.25)	1.00 (.00)
ka.kaa	.43 (.51)	1.00 (.00)	.63 (.50)	.94 (.25)
ka.ka	.50 (.52)	.81 (.40)	.68 (.48)	1.00 (.00)
kuu.kuu	.56 (.51)	.75 (.45)	.62 (.50)	.63 (.50)
kuu.ku	.94 (.25)	.93 (.25)	.94 (.25)	1.00 (.00)
ku.kuu	.50 (.52)	.88 (.34)	.38 (.50)	.88 (.34)
ku.ku	.43 (.51)	1.00 (.00)	.63 (.50)	1.00 (.00)
saa.saa	.75 (.45)	.81 (.40)	.68 (.48)	.81 (.40)
saa.sa	.75 (.45)	1.00 (.00)	.68 (.48)	1.00 (.00)
sa.saa	.50 (.52)	.94 (.25)	.63 (.50)	.88 (.34)
sa.sa	.43 (.51)	.94 (.25)	.68 (.48)	.94 (.25)
suu.suu	.63 (.50)	.75 (.45)	.69 (.48)	.69 (.48)
suu.su	.75 (.45)	.94 (.25)	.81 (.40)	1.00 (.00)
su.suu	.50 (.52)	1.00 (.00)	.63 (.50)	.88 (.34)
su.su	.44 (.51)	.94 (.25)	.69 (.48)	.88 (.34)

A repeated-measure ANOVA was used to test whether the effects of perceptual training transfer to correct production of the vowel duration. Within-subject factors were time (2; pretest and posttest), vowel type (2: high /u/ and low /a/ vowel), preceding consonant (2: /k/ and /s/), token type (4: CVV.CVV, CVV.CV, CV.CVV, CV.CV); the between-subject factor was group type (2; AV, A-Only). The dependent variable was production accuracy. Results indicated significant main effects of time, $F_{\text{Time}}(1, 30) = 67.148$, $p < .001$, $\eta_p^2 = .691$, and token type, $F_{\text{TType}}(3, 90) = 5.392$, $p = .002$, $\eta_p^2 = .152$; however, vowel type, $F_{\text{Vowel}}(1, 30) = 1.815$, p

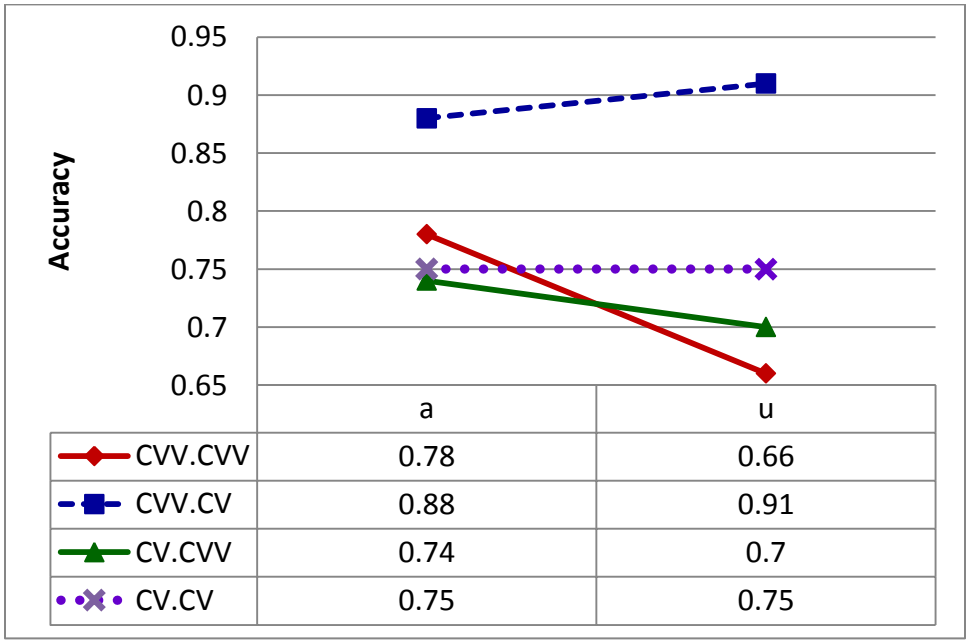
= .188, and group type, $F_{Group}(1, 30) = 1.600$, $p = .216$, and preceding consonant, $F_{PreC}(1, 30) = .062$, $p = .806$, were not significant. It was found that the token types significantly affected the accuracy of participant's production of vowel duration. The mean accuracy for the CVV.CVV was .72, CVV.CV was .90, CV.CVV was .72, and CV.CV was .75. In order to locate where the differences existed in the four token types, pairwise comparisons were performed using the Bonferroni correction. The results indicated that CVV.CV was significantly different from CVV.CVV ($p = .003$), CV.CVV ($p = .004$), and CV.CV ($p = .012$) and showed more accurate production. The findings suggest that learners found it easier to produce a long vowel when there was only one and it occurred in the first syllable.

In addition to the main effects of token type, the Time x Token Type interaction, $F(3, 90) = 7.977$, $p < .001$, $\eta_p^2 = .210$, and the Vowel Type x Token Type interaction, $F(3, 90) = 2.929$, $p = .038$, $\eta_p^2 = .089$, were also significant (Figure 39). To analyze the interactions in detail, simple effects tests were conducted. Regarding the Time x Token Type interaction, results revealed that CVV.CV was better at pretest, CV.CVV and CV.CV showed parallel improvement, and CVV.CVV barely improved. Regarding the Vowel Type x Token Type interaction, the accuracy of the CVV.CVV token type was higher when the vowel was /a/ compared to /u/.

Figure 39: The comparison of production accuracy by vowel and token type in Experiment 2



	Pretest	Posttest
CVV.CVV	0.69	0.76
CVV.CV	0.82	0.97
CV.CVV	0.52	0.92
CV.CV	0.56	0.94



	a	u
CVV.CVV	0.78	0.66
CVV.CV	0.88	0.91
CV.CVV	0.74	0.7
CV.CV	0.75	0.75

The production errors that the learners made during the production test are summarized in Table 23 below. The learners made more errors when they pronounced the CVV.CVV tokens; they tended to shorten the vowel in the second syllable. Also, the errors of the CV.CVV and

CV.CV types showed that the short vowels on the first syllable were harder to correctly pronounce because they were generally lengthened.

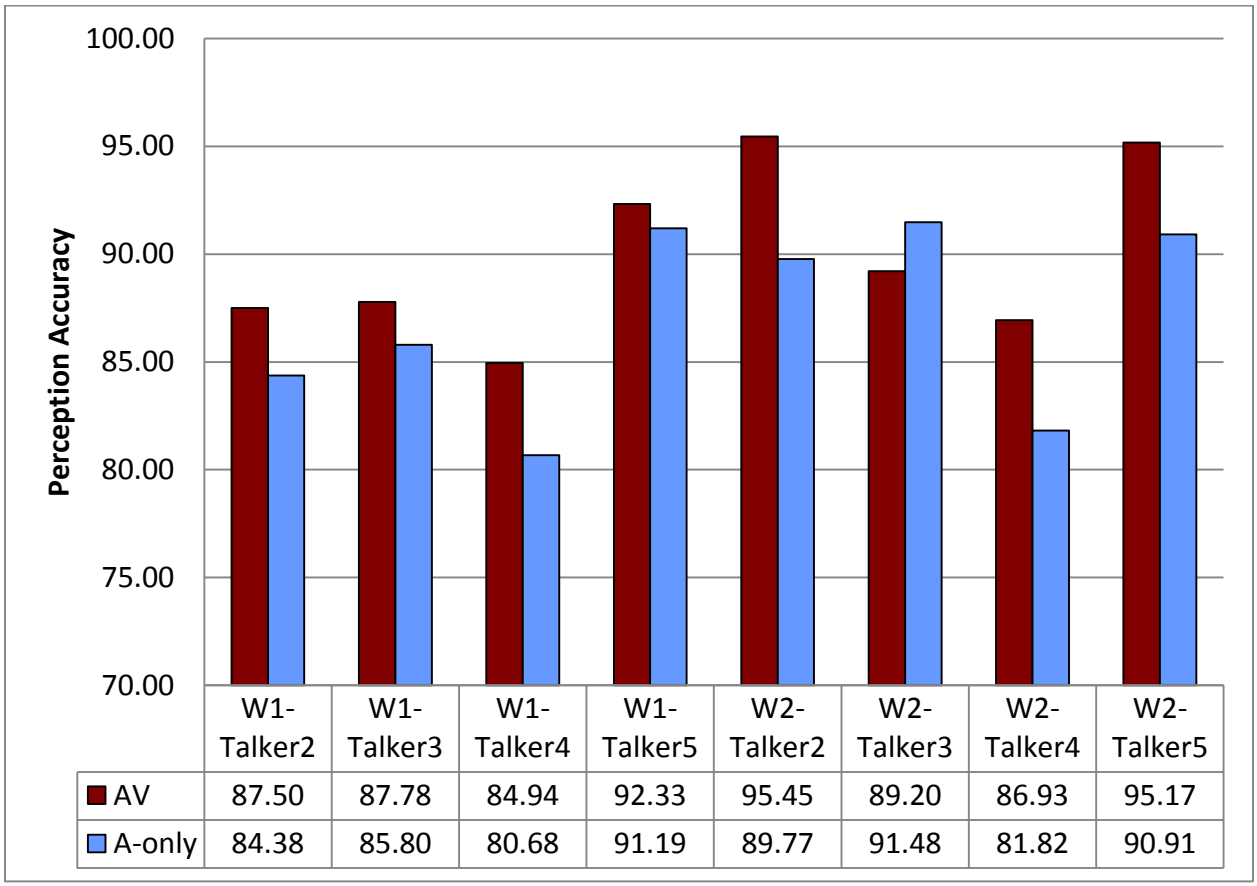
Table 23: Errors observed in the production posttest in Experiment 2

Token with /a/	Errors	Number	Token with /u/	Errors	Number
CaaCaa	CaaCa	11	CuuCuu	CuuCu	17
	CaCaa	1		CuCuu	1
CaaCa	CaCaa	1	CuuCu	CuCu	1
	CaCa	1		CuCu	4
CaCaa	CaaCa	2	CuCuu	CuuCu	1
	CaaCaa	1		CuuCuu	3
	CaCCaa	1		CuCu	2
CaCa	CaCaa	1	CuCu	CuCuu	1
	CaaCa	5		CuuCu	1

In conclusion, production accuracy improved from pretest to the posttest while there was no statistically significant difference between the two training groups. Thus, since the learners did not receive any specific production training or practice, it was considered that the positive effect of the focused perceptual training on the L2 vowel duration was transferred to production. The interaction between time (i.e., pretest and posttest) and token type as well as vowel type and token type was found. The three token types, CVV.CV, CV.CVV, and CV.CV, significantly improved after the training, but not the CVV.CVV type. Also, there was a tendency for the CVV.CVV tokens to be more accurately pronounced if they contained the vowel /a/, compared to the vowel /u/.

Analysis of Effectiveness of Training per Group: In order to examine the development of accuracy and response latency as well as effects of talker and other factors (i.e., pitch pattern, vowel type, and preceding consonant) during the training, the perception accuracy and RT in the training sessions were analyzed by training groups. Figure 40 illustrates the identification accuracy in each training session (total of 8) by the AV and A-only groups. For both groups, perception accuracy starting the end of the first week (i.e., Session 4) became higher; however, accuracy in the third session in the second week (i.e., Session 7) was lower than the other sessions in the weeks. In addition, AV groups showed higher accuracy than A-only groups, except for Session 6.

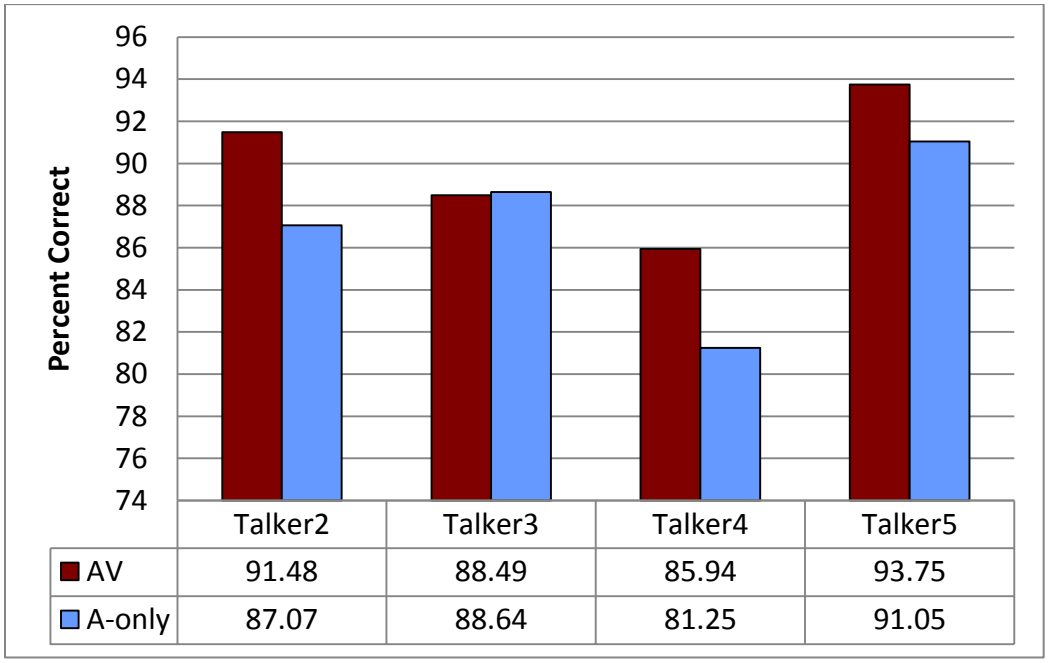
Figure 40: Perception accuracy in each week and talker by AV and A-only groups



	W1-Talker2	W1-Talker3	W1-Talker4	W1-Talker5	W2-Talker2	W2-Talker3	W2-Talker4	W2-Talker5
AV	87.50	87.78	84.94	92.33	95.45	89.20	86.93	95.17
A-only	84.38	85.80	80.68	91.19	89.77	91.48	81.82	90.91

Figure 41 shows the perception accuracy by the four talkers used in the training; Talker2 (F) was assigned for the first and the fourth sessions; Talker3 (M) was assigned for the second and sixth sessions; Talker 4 (M) was assigned for the third and seventh sessions; and Talker 5 (F) was assigned for the fourth and eighth sessions. Accuracy for tokens produced by Talker 3 was comparable for both groups; in other cases, the AV training group showed higher scores.

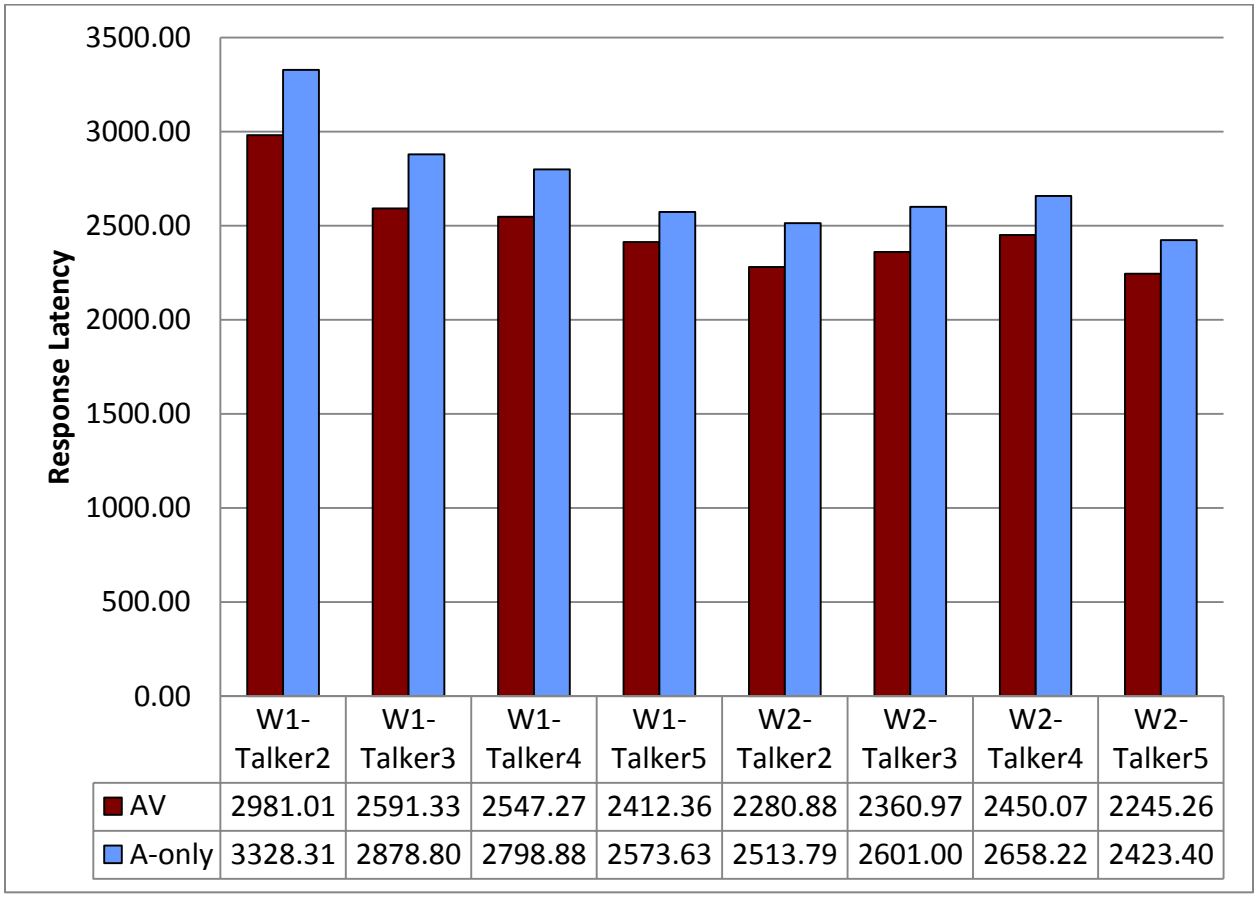
Figure 41: Perception accuracy by talker in perceptual training



	Talker2	Talker3	Talker4	Talker5
AV	91.48	88.49	85.94	93.75
A-only	87.07	88.64	81.25	91.05

Figure 42 illustrates the RT in each training session by the AV and A-only groups. Across the sessions, RTs were faster for the AV groups.

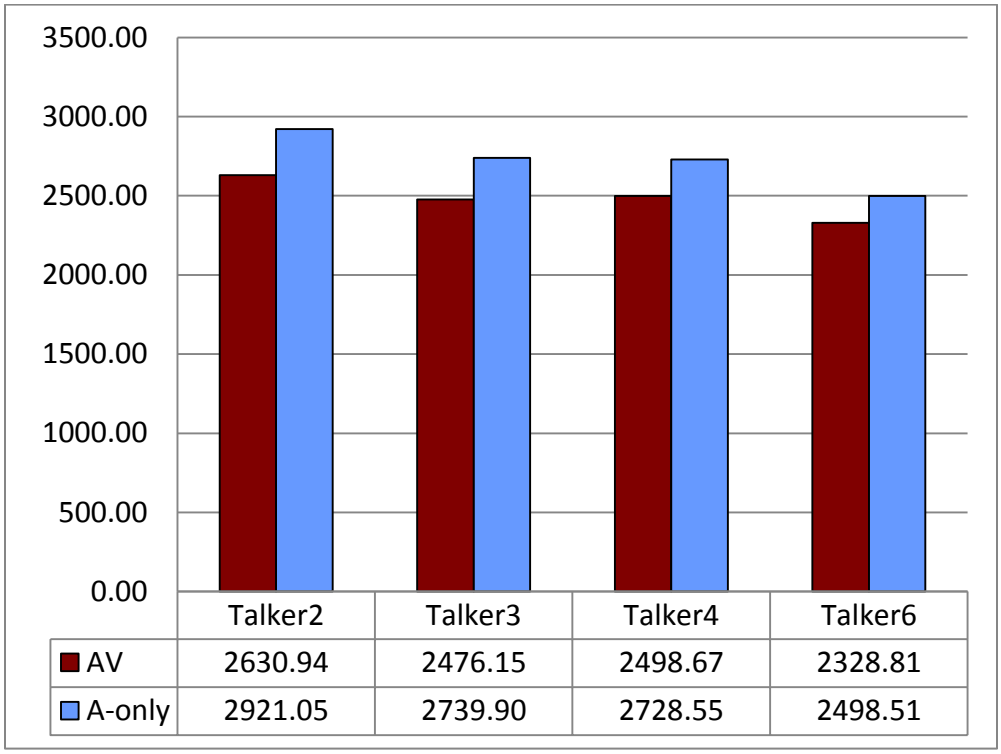
Figure 42: The RT for each week and talker by AV and A-only groups



	W1-Talker2	W1-Talker3	W1-Talker4	W1-Talker5	W2-Talker2	W2-Talker3	W2-Talker4	W2-Talker5
AV	2981.01	2591.33	2547.27	2412.36	2280.88	2360.97	2450.07	2245.26
A-only	3328.31	2878.80	2798.88	2573.63	2513.79	2601.00	2658.22	2423.40

Figure 43 shows the RT to tokens produced by the four talkers used in the training. As described earlier, Talker2 (F) was assigned for the first and the fourth sessions; Talker3 (M) was assigned for the second and sixth sessions; Talker 4 (M) was assigned for the third and seventh sessions; and Talker 5 (F) was assigned for the fourth and eighth sessions. The AV groups showed faster RTs to all talkers.

Figure 43: The RT in the training grouped by the four talkers



In order to examine the development of correct identification, response latency, and the influence of other factors such as talker in training sessions, pitch pattern, vowel type, and preceding consonant, the 22 tokens used in the training (Appendix F) were divided into four groups depending on the pitch pattern as shown in Figure 44; vowel type, preceding consonant, and pitch pattern were combined as stimulus type following the earlier analysis of the pretest and posttest data.

Figure 44: Tokens in the training sessions by stimulus type

Group I

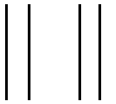
1. CVV.CVV

LH HH
kaa.kaa

2. CVV.CVV

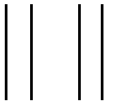
HL LL
kaa.kaa

3. CVV.CVV



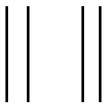
LH HL
saa.saa

4. CVV.CVV



LH HH
kuu.kuu

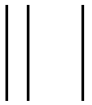
5. CVV.CVV



LH HL
suu.suu

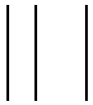
Group II

6. CVV.CV



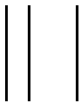
LH H
kaa.ka

7. CVV.CV



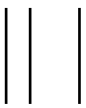
LH H
saa.sa

8. CVV.CV



HL H
saa.sa

9. CVV.CV



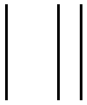
LH H
suu.su

Group III

10. CV.CVV

L HH
ka.kaa

11. CV.CVV



L HL
sa.saa

12. CV.CVV

L HL
ku.kuu

13. CV.CVV

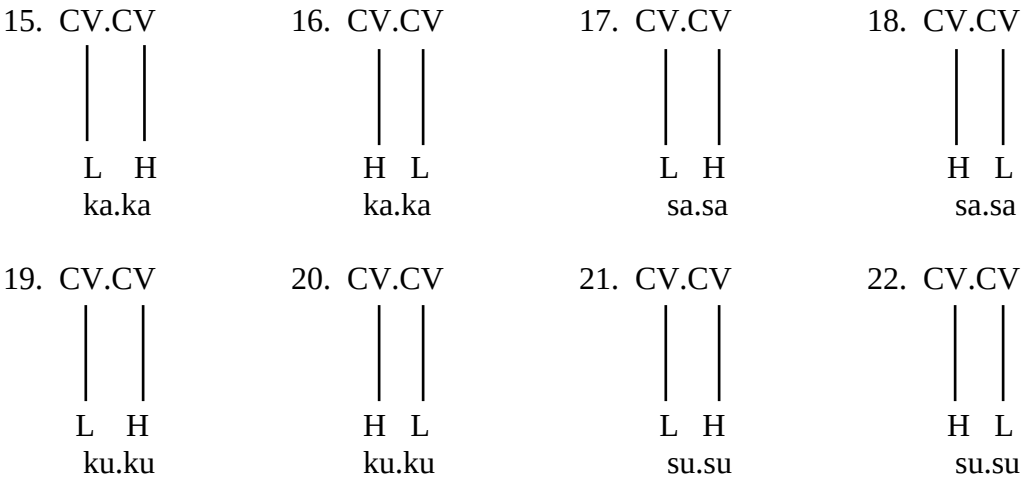
H LL
ku.kuu

14. CV.CVV

L HH
su.suu

Figure 44 (cont'd)

Group IV



Perception Accuracy in Training - AV group: A three-way ANOVA was performed to examine the development of perception accuracy and effects of the factors for the AV group. The independent variables were week (2: Week1, Week2), talker (4: Talker2, 3, 4, 5), and stimulus type. The dependent variable was perception accuracy in the eight training sessions.

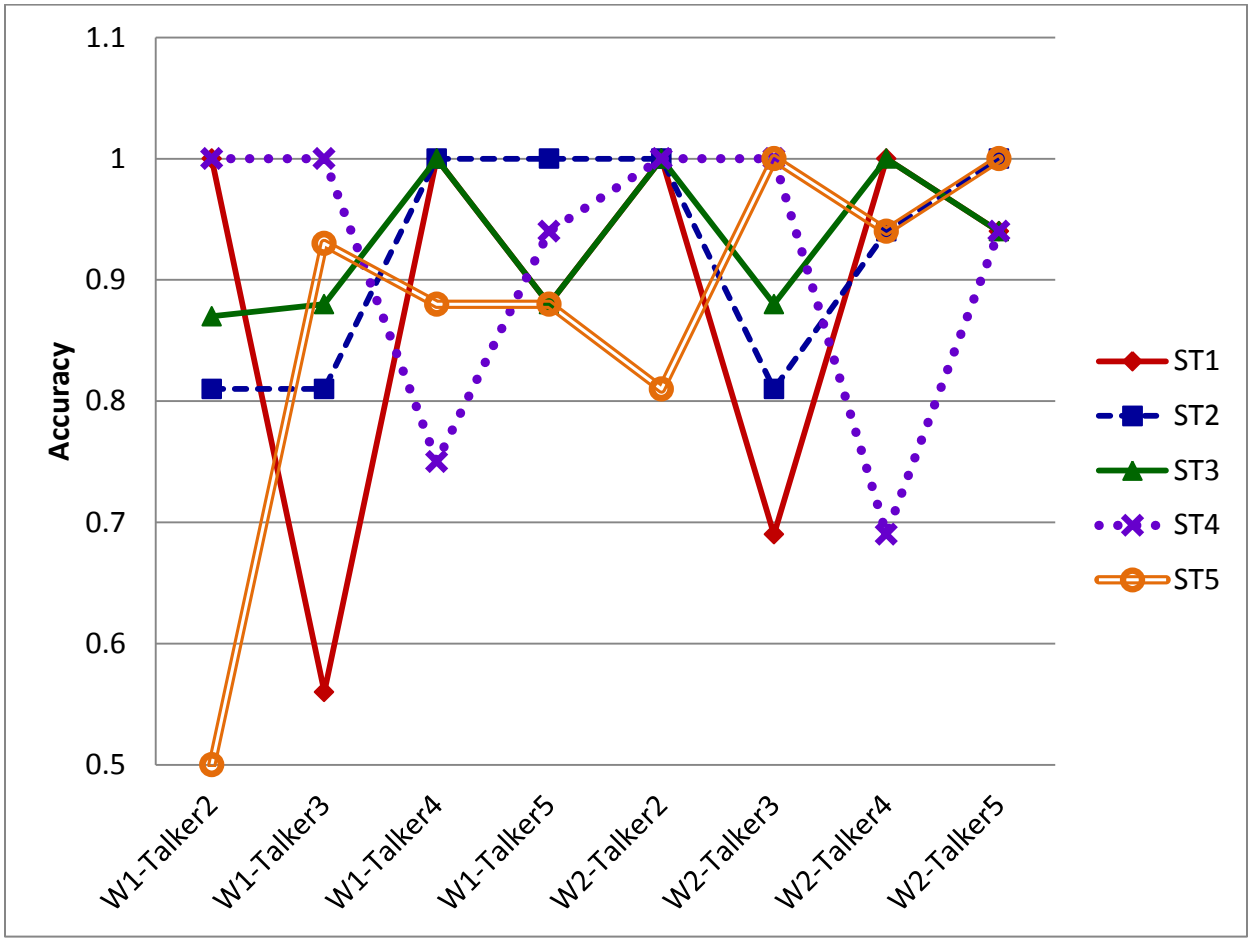
Regarding the tokens in Group I (CVV.CVV), results indicated no significant main effects: $F_{\text{Week}}(1, 15) = 4.444$, $p = .052$; talker, $F_{\text{Talker}}(3, 45) = 2.042$, $p = .121$; and stimulus type, $F_{\text{Type}}(4, 60) = 1.113$, $p = .350$. The mean accuracy scores for the first week and second week were .88 and .93 respectively. The difference was marginally significant. The mean accuracy scores for each talker were .90 (Talker 2), .86 (Talker 3), .92 (Talker 4), and .94 (Talker 5), and there were no significant differences among them. Table 24 shows mean accuracy scores for each stimulus type in Group I. There were no statistically significant differences among the five tokens.

Table 24: Mean accuracy scores of the five tokens in Group I (CVV.CV) (AV group)

Stimulus Type (ST)	Tokens	Mean Accuracy Scores
1	kaakaa (LH.HH)	.88
2	kaa.kaa (HL.LL)	.92
3	saa.saa (LH.HL)	.93
4	kuu.kuu (LH.HH)	.91
5	suu.suu (LH.HL)	.87

Although there were no significant main effects, the Talker x Stimulus Type interaction was significant: $F(12, 180) = 5.835$, $p < .001$, $\eta_p^2 = .280$ (Figure 45). The Week x Voice interaction was marginally significant, $F(3, 45) = 2.818$, $p = .050$. Results of simple effects tests revealed that perception accuracy of ST5 produced by Talker2 and that of ST1 produced by Talker3 were significantly lower in the first week than in the second week.

Figure 45: The comparison of perception accuracy of tokens in Group I for AV training group



Regarding the tokens in Group II (CVV.CV), results indicated significant main effects of talker, $F_{\text{Talker}}(3, 45) = 19.056$, $p < .001$, $\eta_p^2 = .560$, and stimulus type, $F_{\text{Type}}(3, 45) = 17.328$, $p < .001$, $\eta_p^2 = .536$; however, week was not significant, $F_{\text{Week}}(1, 15) = .958$, $p = .343$. The mean accuracy scores for the first week and second week were .79 and .83 respectively. The second week had higher accuracy; however, the difference was not significant. The mean accuracy scores for each talker were .93 (Talker 2), .82 (Talker 3), .63 (Talker 4), and .89 (Talker 5). Results of the pairwise comparisons with Bonferroni correction indicated that Talker4 was different from Talker2 ($p < .001$), Talker3 ($p = .002$), as well as Talker5 ($p < .001$). Thus,

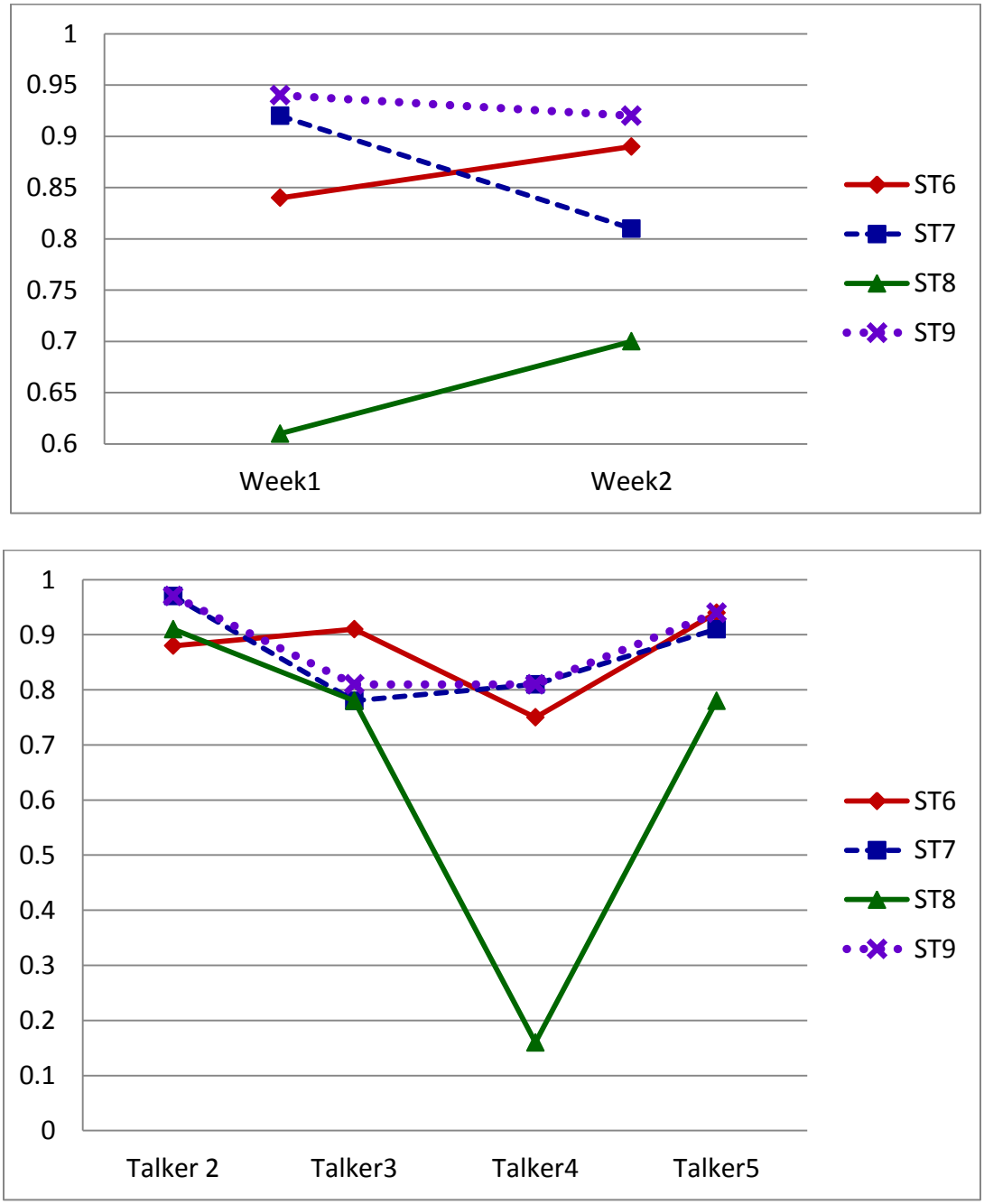
Talker4, a male talker, was the most difficult for L2 learners to correctly perceive vowel duration. Table 25 shows mean accuracy scores for each stimulus type in Group II. Results of pairwise comparisons with Bonferroni correction indicated that ST8 was different from ST6 ($p < .001$), ST7 ($p < .001$), and ST9 ($p < .001$). The ST6 had the lowest accuracy among the four tokens.

Table 25: Mean accuracy scores of the four tokens in Group II (CVV.CV) (AV group)

Stimulus Type (ST)	Tokens	Mean Accuracy Scores
6	kaa.ka (HL.L)	.87
7	saa.sa (LH.H)	.87
8	saa.sa (HL.L)	.66
9	suu.su (LH.H)	.88

In addition to the main effects, the Week x Stimulus Type interaction was significant: $F(3, 45) = 3.169$, $p = .033$, $\eta_p^2 = .174$ (Figure 46). Results of the simple effects tests indicated that the difference between ST7 and ST8 was greater in the second week than in the first week; the accuracy of ST8 improved in the second week although that of ST7 decreased. In addition, the Talker x Stimulus Type interaction was significant: $F(9, 135) = 5.326$, $p < .001$, $\eta_p^2 = .262$. Results of the simple effects tests indicated that the effects of the talker were greater for ST8. The accuracy of ST8 was higher with Talker2, Talker3, and Talker5; however, Talker4 revealed significantly lower accuracy as shown in the graph.

Figure 46: The comparison of perception accuracy of tokens in Group II for AV training group



Regarding the tokens in Group III (CV.CVV), results indicated significant main effects of talker, $F_{\text{Talker}}(3, 45) = 4.470$, $p = .008$, $\eta_p^2 = .230$, and stimulus type, $F_{\text{Type}}(4, 60) = 3.982$, p

= .006, η_p^2 = .210; however, week was not significant, $F_{Week}(1, 15) = .283$, $p = .603$. None of the interactions was significant. The mean accuracy scores for the first week and second week were .91 and .92 respectively; there was no significant difference between the two weeks. The mean accuracy scores for each talker were .89 (Talker 2), .91 (Talker 3), .89 (Talker 4), and .96 (Talker 5). Results of the pairwise comparisons with Bonferroni correction indicated that Talker5 was different from Talker2 ($p = .019$) and Talker3 ($p = .045$). Thus, Talker5, a female talker, was easier for L2 learners to correctly perceive vowel duration than Talker2, another female talker, and Talker3, a male talker. Table 26 shows mean accuracy scores for each token in Group III. Results of pairwise comparisons with Bonferroni correction did not indicate any significant differences among the five tokens; however, ST11 had relatively lower accuracy than the other four tokens.

Table 26: Mean accuracy scores of the five tokens in Group III (CV.CVV) (AV group)

Stimulus Type (ST)	Tokens	Mean Accuracy Scores
10	ka.kaa (L.HH)	.94
11	sa.saa (L.HL)	.84
12	ku.kuu (L.HL)	.92
13	ku.kuu (H.LL)	.95
14	su.suu (L.HH)	.91

Regarding the tokens in Group IV (CV.CV), results indicated significant main effects of stimulus type, $F_{Type}(7, 105) = 2.717$, $p = .012$, η_p^2 = .153, and week, $F_{Week}(1, 15) = 6.363$, p = .023, η_p^2 = .298; however, talker was not significant, $F_{Talker}(3, 45) = .884$, $p = .456$. None of the interactions was significant. The mean accuracy scores for the first week and second week

were .91 and .95 respectively; perception accuracy for the second week was significantly higher than the first week. Thus, it was concluded that there was a significant development of accuracy in the second week. The mean accuracy scores for each talker were .93 (Talker2), .92 (Talker3), .91 (Talker4), and .95 (Talker5); there were no significant differences among the four talkers. Table 27 shows mean accuracy scores for each token in Group IV. Although the significant differences among the 8 tokens were found, results of pairwise comparisons with Bonferroni correction did not indicate any significant differences among the eight tokens. However, ST18 and ST22 revealed relatively lower accuracy than the other six tokens.

Table 27: Mean accuracy scores of the eight tokens in Group IV (CV.CV) (AV group)

Stimulus Type (ST)	Tokens	Mean Accuracy Scores
15	ka.ka (L.H)	.96
16	ka.ka (H.L)	.91
17	sa.sa (L.H)	.97
18	sa.sa (H.L)	.88
19	ku.ku (L.H)	.96
20	ku.ku (H.L)	.88
21	su.su (L.H)	.95
22	su.su (H.L)	.93

Perception Accuracy in Training – A-only Group: A three-way ANOVA was performed to examine the development of perception accuracy and effects of the factors for the A-only group. The independent variables were week (2: Week1, Week2), talker (4: Voice2, 3, 4, 5), and stimulus type. The dependent variable was perception accuracy in the eight training sessions.

Regarding the tokens in Group I (CVV.CVV), results indicated significant main effects of week, $F_{\text{Week}}(1, 15) = 6.310$, $p = .024$, $\eta_p^2 = .296$; however, talker, $F_{\text{Talker}}(3, 45) = .823$, p

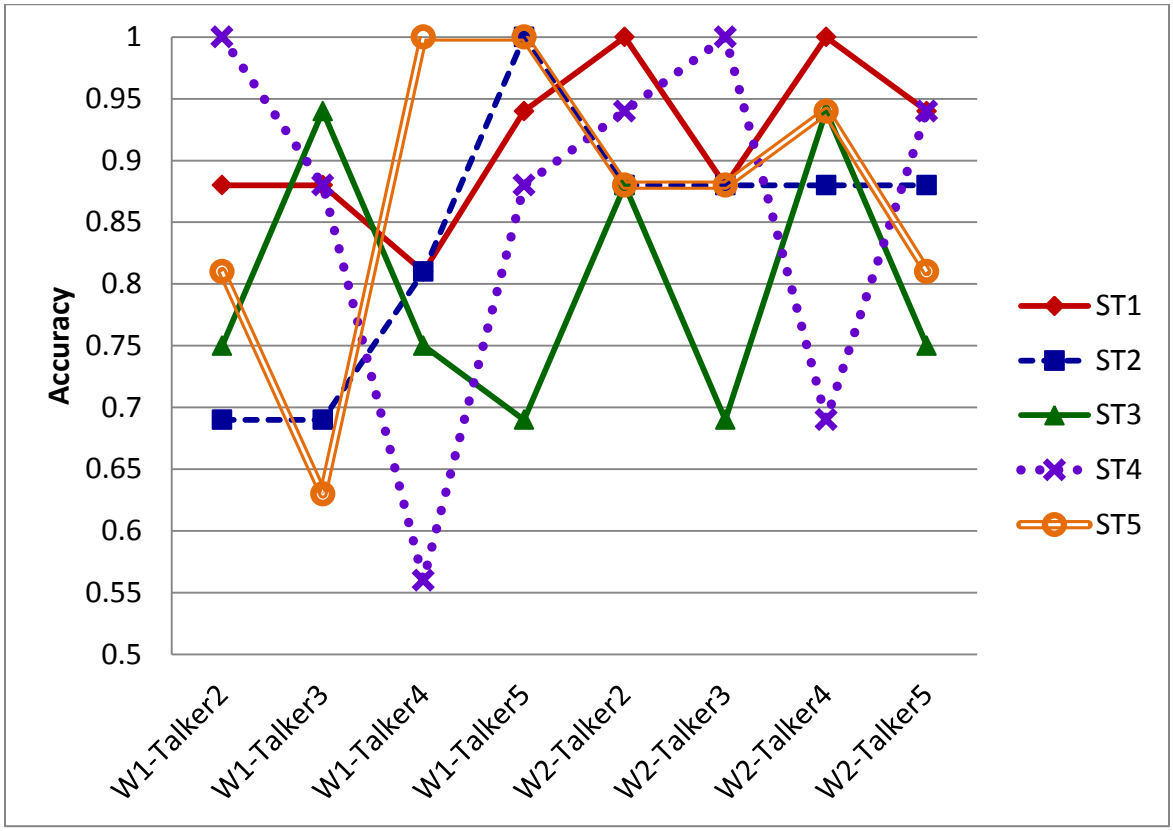
= .488, and stimulus type, $F_{Type}(4, 60) = 1.919$, $p = .119$, were not significant. The mean accuracy scores for the first week and second week were .83 and .88 respectively; there was significant development of accuracy from the first week to the second week. The mean accuracy scores for each talker were .87 (Talker2), .83 (Talker3), .84 (Talker4), and .88 (Talker5), and there were no significant differences among them. Table 28 shows mean accuracy scores for each stimulus type in Group I. ST1 had relatively higher accuracy and ST3 had relatively lower accuracy; however, there were no significant differences among the five tokens.

Table 28: Mean accuracy scores of the five tokens in Group I (CVV.CVV) (A-only group)

Stimulus Type (ST)	Tokens	Mean Accuracy Scores
1	kaa.kaa (LH.HH)	.91
2	kaa.kaa (HL.LL)	.84
3	saa.saa (LH.HL)	.80
4	kuu.kuu (LH.HH)	.86
5	suu.suu (LH.HL)	.87

In addition to the significant main effects of week, the Talker x Stimulus Type interaction was significant: $F(12, 180) = 2.834$, $p = .001$, $\eta_p^2 = .159$. In addition, the Week x Talker x Stimulus Type interaction was significant: $F(12, 180) = 1.815$, $p = .049$, $\eta_p^2 = .108$ (Figure 47). Simple effects tests were performed to analyze the three-way interaction, and results revealed that perception accuracy of ST4 produced by Talker4 in the first week was significantly lower. In addition, perception accuracy of ST5 produced by Talker3 in the first week was lower; however, it improved in the second week.

Figure 47: The comparison of perception accuracy of tokens in Group I (CVV.CVV) for A-only training group



Regarding the tokens in Group II (CVV.CV), results indicated significant main effects of talker, $F_{\text{Talker}}(3, 45) = 15.527$, $p < .001$, $\eta_p^2 = .509$, and stimulus type, $F_{\text{Type}}(3, 45) = 7.242$, $p < .001$, $\eta_p^2 = .326$; however, week was not significant, $F_{\text{Week}}(1, 15) = 3.412$, $p = .085$. The mean accuracy scores for the first week and second week were .77 and .82 respectively. The second week had higher accuracy; however, the difference was not significant. The mean accuracy scores for each voice were .91 (Talker2), .84 (Talker3), .58 (Talker4), and .88 (Talker5). Results of the pairwise comparisons with Bonferroni correction indicated that Talker4 was different from Talker2 ($p = .001$), Talker3 ($p = .008$), and Talker5 ($p < .002$). Thus, Talker4 was

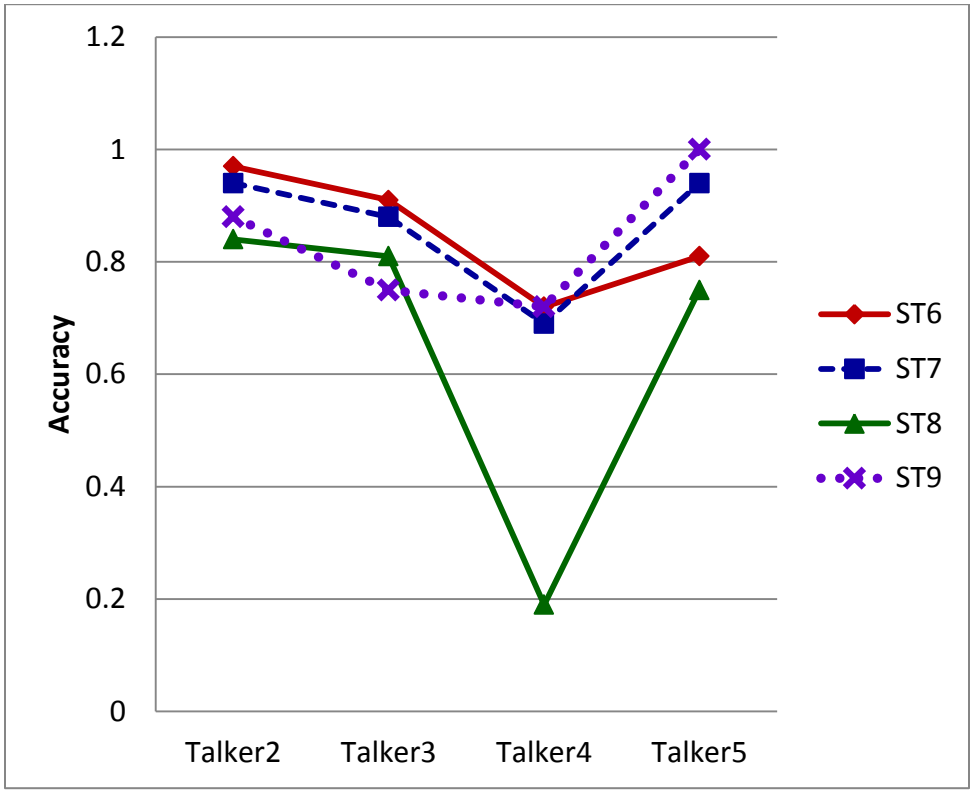
the most difficult for L2 learners to correctly perceive vowel duration. Table 29 shows mean accuracy scores for each stimulus type in Group II. Results of pairwise comparisons with Bonferroni correction indicated that ST8 was different from ST6 ($p < .009$), ST7 ($p < .011$), and ST9 ($p < .002$). The ST 8 had the lowest accuracy among the four tokens.

Table 29: Mean accuracy scores of the four tokens in Group II (CVV.CV) (A-only group)

Stimulus Type (ST)	Tokens	Mean Accuracy Scores
6	kaa.ka (HL.L)	.85
7	saa.sa (LH.H)	.86
8	saa.sa (HL.L)	.65
9	suu.su (LH.H)	.84

In addition to the main effects, the Talker x Stimulus Type interaction was significant: $F(9, 135) = 4.659$, $p < .001$, $\eta_p^2 = .237$ (Figure 48). Results of the simple effects tests indicated that the effects of the talker were greater for ST8. The accuracy of ST8 was higher with Talker2, Talker3, and Talker5; however, Talker4 revealed significantly lower accuracy.

Figure 48: The comparison of perception accuracy of tokens in Group II (CVV.CV) for A-only training group



Regarding the tokens in Group III (CV.CVV), results indicated significant main effects of talker, $F_{Talker}(3, 45) = 3.425$, $p = .025$, $\eta_p^2 = .186$, and stimulus type, $F_{Type}(4, 60) = 8.788$, $p < .001$, $\eta_p^2 = .369$; however, week was not significant, $F_{Week}(1, 15) = .516$, $p = .484$. The mean accuracy scores for the first week and second week were .90 and .91 respectively; there was no difference between the two weeks. The mean accuracy scores for each talker were .86 (Talker2), .92 (Talker3), .91 (Talker4), and .95 (Talker5). Results of the pairwise comparisons with Bonferroni correction did not find significant differences among the four talkers; however, the difference between Talker2 and Talker5 was approaching significance ($p = .081$). Table 30 shows mean accuracy scores for each token in Group III. Results of pairwise comparisons with

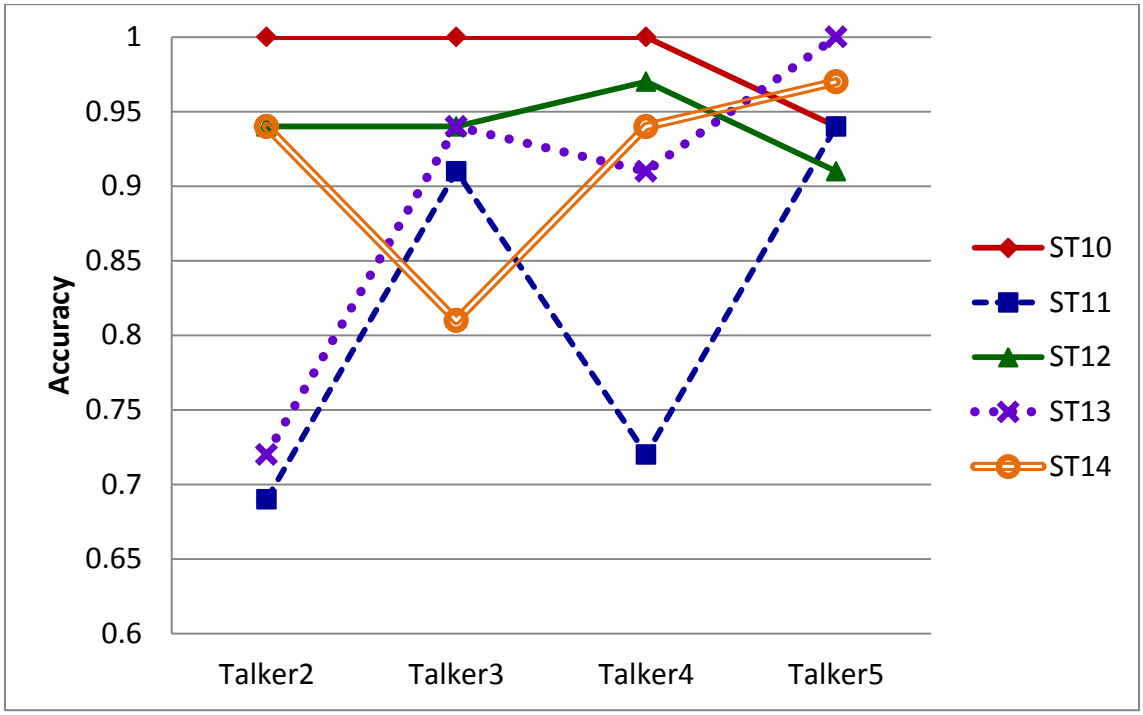
Bonferroni correction indicated that ST11 was significantly different from ST10 ($p < .001$), ST12 ($p = .015$), and ST14 ($p = .047$). Thus, the token with L.HL pitch and the combination of consonant /s/ and a vowel /a/ was more difficult for the learners to perceive correctly than the tokens with the L.HH pitch and /ka/ or /su/ as well as one with the L.HL pitch and /ku/.

Table 30: Mean accuracy scores of the four tokens in Group III (CV.CVV) (A-only group)

Stimulus Type (ST)	Tokens	Mean Accuracy Scores
10	ka.kaa (L.HH)	.98
11	sa.saa (L.HL)	.81
12	ku.kuu (L.HL)	.94
13	ku.kuu (H.LL)	.89
14	su.suu (L.HH)	.91

In addition to the main effects above, the Talker x Stimulus Type interaction was significant: $F(12, 180) = 2.792$, $p = .002$, $\eta_p^2 = .157$ (Figure 49). Results of simple effects tests revealed that the differences between ST11 and ST13 were greater with Talker4 than with the other talkers. The learners demonstrated significantly lower accuracy for ST11 when it was produced by Talker4. In general, Figure 49 shows that accuracy for ST10 and ST12 were much less variable across talkers compared to ST11, ST13, and ST14.

Figure 49: The comparison of perception accuracy of tokens in Group III (CV.CVV) for A-only training group



Regarding the tokens in Group IV (CV.CV), results indicated significant main effects of stimulus type, $F_{Type}(7, 105) = 5.770$, $p < .001$, $\eta_p^2 = .278$, and talker, $F_{Talker}(3, 45) = 3.431$, $p = .025$, $\eta_p^2 = .186$, were significant; however, week was not significant, $F_{Week}(1, 15) = .460$, $p = .508$. The mean accuracy scores for the first week and second week were .89 and .90 respectively. Accuracy in the second week was slightly higher than in the first week; however, the difference was not significant. The mean accuracy scores for each talker were .86 (Talker2), .93 (Talker3), .86 (Talker4), and .92 (Talker5). Results of pairwise comparisons with Bonferroni correction did not indicate any significant differences among the eight tokens; however, Talker3 and Talker5 revealed relatively higher accuracy than the other two talkers. Table 31 shows mean accuracy scores for each token in Group IV. Results of pairwise

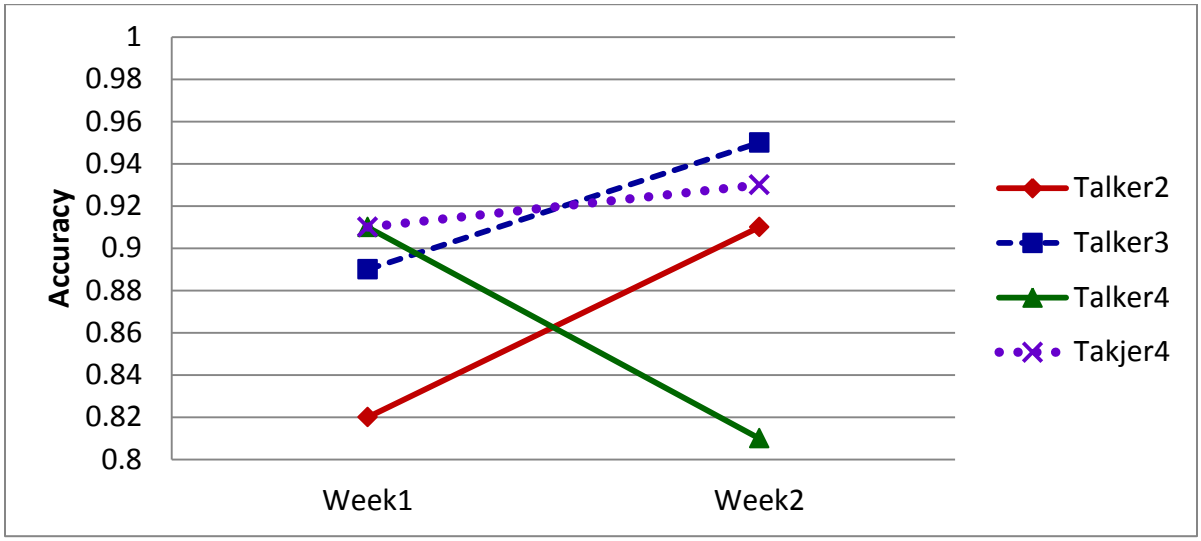
comparisons with Bonferroni correction indicated there were significant differences between ST16 and ST21 (p = .028). Thus, ST21 (L.H) was significantly easier for L2 learners to correctly perceive than ST16 (H.L).

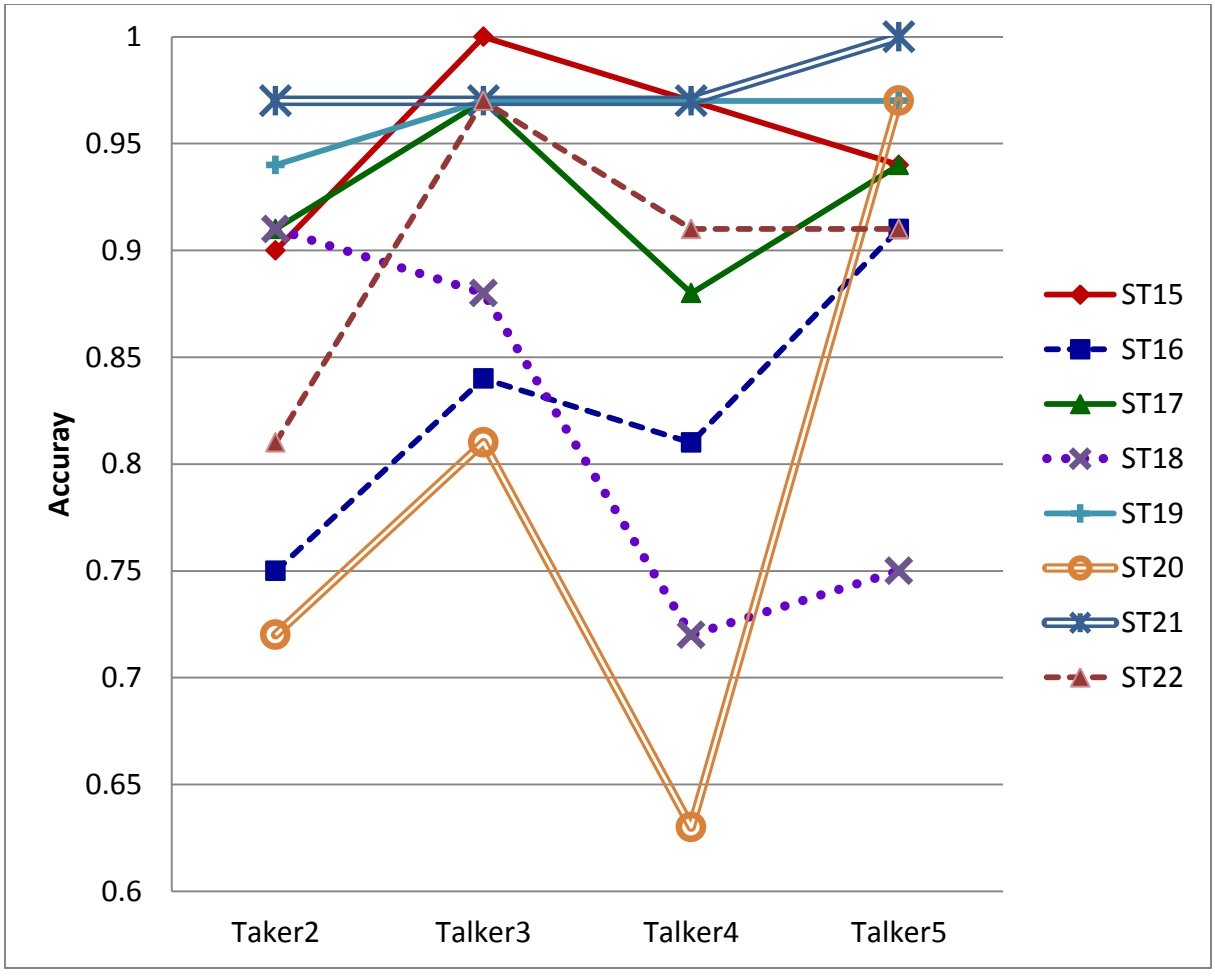
Table 31: Mean accuracy scores of the eight tokens in Group IV (CV.CV) (A-only group)

Stimulus Type (ST)	Tokens	Mean Accuracy Scores
15	ka.ka (L.H)	.95
16	ka.ka (H.L)	.83
17	sa.sa (L.H)	.92
18	sa.sa (H.L)	.81
19	ku.ku (L.H)	.96
20	ku.ku (H.L)	.78
21	su.su (L.H)	.98
22	su.su (H.L)	.90

In addition to the main effects, the Week x Talker interaction was significant: F (3, 45) = 5.293, p = .003, η_p^2 = .261 (Figure 50). Results of the simple effects tests indicated that the difference between Talker 3 and Talker 4 was greater in the second week, compared to the first week. The learners demonstrated lower accuracy in correctly identifying the vowel duration of tokens produced by Talker 3 in the second week. The Talker x Stimulus Type interaction was also significant: F (21, 315) = 1.843, p = .014, η_p^2 = .109. Results of the simple effects tests indicated that the perception accuracy for ST20 was highest with Talker 4 and lowest with Talker4.

Figure 50: The comparisons of perception accuracy of tokens in Group IV (CV.CV) for A-only training group





Perception RT in Training - AV Group: A three-way ANOVA was performed to examine the development of perception RT and effects of the factors for the AV group. The independent variables were week (2: Week1, Week2), talker (4: Talker2, 3, 4, 5), and stimulus type. The dependent variable was perception RT in the eight training sessions.

Regarding the tokens in Group I (CVV.CVV), results indicated significant main effects of week, $F_{Week}(1, 15) = 19.363$, $p = .001$, $\eta_p^2 = .563$, and stimulus type, $F_{Type}(4, 60) = 7.395$, $p < .001$, $\eta_p^2 = .330$; however, talker was not significant, $F_{Talker}(3, 45) = 2.340$, $p = .086$. The mean RT scores for the first week and second week were 2733.71 milliseconds and 2399.91 milliseconds respectively. The RT in the second week was significantly faster than the one in the first week. The mean RT scores for each voice were 2687.28 milliseconds (Talker2), 2598.64 milliseconds (Talker3), 2576.61 milliseconds (Talker4), and 2404.71 milliseconds (Talker5), and there were no significant differences among them. Table 32 shows mean RT scores for each stimulus type in Group I.

Table 32: Mean RT scores of the five tokens in Group I (CVV.CVV) (AV group)

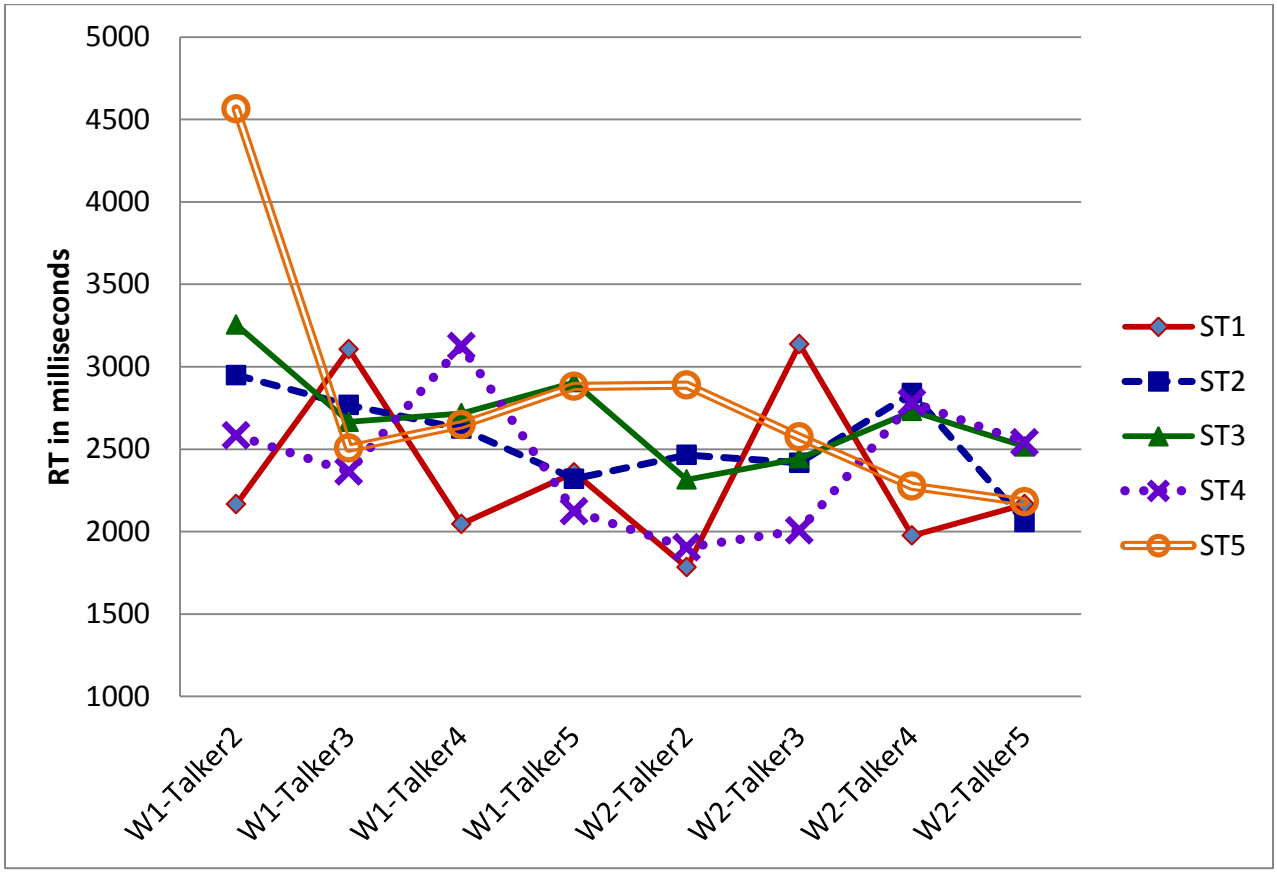
Stimulus Type (ST)	Tokens	Mean RT (milliseconds)
1	kaa.kaa (LH.HH)	2342.13
2	kaa.kaa (HL.LL)	2555.11
3	saa.saa (LH.HL)	2693.26
4	kuu.kuu (LH.HH)	2429.35
5	suu.suu (LH.HL)	2814.21

Results of pairwise comparisons with Bonferroni correction indicated that ST5 was significantly different from ST1 ($p = .002$) as well as ST4 ($p = .001$). In addition, ST4 was significantly different from ST3 ($p = .038$). The learners' response latency for ST5 was significantly slower

than ST1 and ST4. ST5 shares the same pitch pattern as ST3 but involved /su/ versus /sa/. Also, the response latency of ST3 was slower than ST4.

In addition to the main effects, the Week x Talker interaction, $F(3, 45) = 4.985$, $p = .005$, $\eta_p^2 = .249$, the Week x Stimulus Type interaction, $F(12, 180) = 9.305$, $p < .001$, $\eta_p^2 = .383$, and the Week x Talker x Stimulus Type interaction, $F(12, 180) = 1.911$, $p = .036$, $\eta_p^2 = .113$, were significant (Figure 51). Simple effects tests were performed in order to analyze the three-way interaction, and results indicated that the RT difference between ST1 and ST5 was greater for Talker2, compared to the other talkers, in the first week. Thus, the learners demonstrated slower RTs for ST5 produced by Talker2 than ST1 in the first week.

Figure 51: The comparison of perception RT of tokens in Group I (CVV.CVV) for AV training group



Regarding the tokens in Group II (CVV.CV), results indicated significant main effects of week, $F_{Week}(1, 15) = 5.105$, $p = .039$, $\eta_p^2 = .254$, and stimulus type, $F_{Type}(3, 45) = 3.139$, $p = .034$, $\eta_p^2 = .173$; however, talker was not significant, $F_{Talker}(3, 45) = 2.400$, $p = .080$. The mean RTs for the first week and second week were 2916.06 milliseconds and 2690.68 milliseconds respectively; the second week had significantly faster RTs. The mean RTs for each talker were 2696.95 milliseconds (Talker2), 2915.54 milliseconds (Talker3), 2934.62 milliseconds (Talker4), and 2666.37 milliseconds (Talker5). Two female talkers (Talker2 and

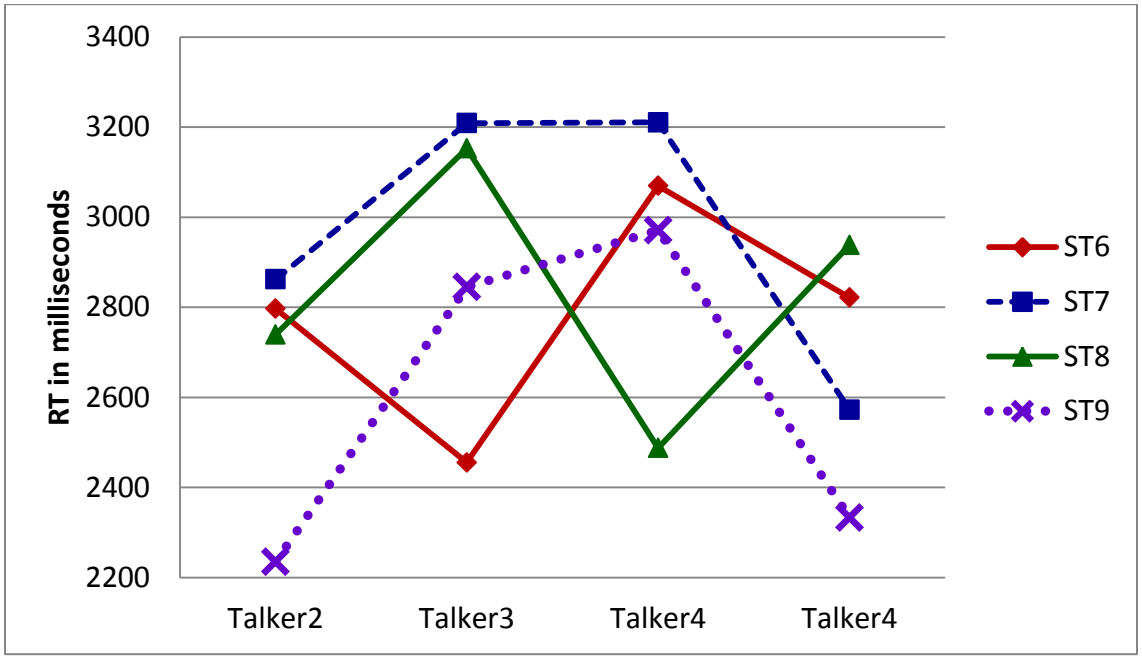
Talker5) had relatively faster RT than the two male talkers (Talker3 and Talker4); however, the difference was not significant. Table 33 shows mean RTs for stimuli in Group II. Results of pairwise comparisons with Bonferroni correction indicated that ST7 was significantly different from ST9 (p = .034); RT of ST9 was significantly faster than that of ST7.

Table 33: Mean RT scores of the four tokens in Group II (AV group)

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)
6	kaa.ka (HL.L)	2824.40
7	saa.sa (LH.H)	2963.51
8	saa.sa (HL.L)	2829.56
9	suu.su (LH.H)	2596.02

In addition to the main effects, the Talker x Stimulus Type interaction was significant: F (9, 135) = 3.786, p < .001, η_p^2 = .202 (Figure 52). Results of the simple effects tests indicated that the differences between ST6 and ST8 were greater for Talker3 and Talker4, compared to Talker5. The learners demonstrated significantly slower RT for ST8 produced by Talker3 compared to ST6. On the other hand, the learners showed significantly longer RT for ST6 produced by Talker4 compared to ST8.

Figure 52: The comparison of perception RT of tokens in Group II (CVV.CV) for AV training group



Regarding the tokens in Group III (CV.CVV), results indicated significant main effects of week, F_{Week} (1, 15) = 5.525, p = .033, η_p^2 = .269, and talker, F_{Talker}(3, 45) = 6.417, p = .001, η_p^2 = .300; however, stimulus type was not significant, F_{Type}(4, 60) = 2.319, p = .067. None of the interactions was significant. The mean RT scores for the first week and second week were 2819.68 milliseconds and 2610.72 milliseconds respectively; the RT of the second week was significantly faster than the first week. The mean RTs for each talker were 2957.73 milliseconds (Talker2), 2644.70 milliseconds (Talker3), 2691.86 milliseconds (Talker4), and 2566.53 milliseconds (Talker5). Results of the pairwise comparisons with Bonferroni correction indicated that Talker3 was different from Talker3 (p = .021) and Talker5 (p = .004). Thus, the learners had longer response latency for Talker2, a female talker, compared to Talker3, a male

talker, and Talker5, another female talker. Table 34 shows mean RTs for each token in Group III. ST10 revealed relatively faster RT than other four tokens; however, the difference was not significant.

Table 34: Mean RT scores of the five tokens in Group III (CV.CVV) (AV group)

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)
10	ka.kaa (L.HH)	2530.57
11	sa.saa (L.HL)	2790.12
12	ku.kuu (L.HL)	2722.75
13	ku.kuu (H.LL)	2850.44
14	su.suu (L.HH)	2682.14

Regarding the tokens in Group IV (CV.CV), results indicated significant main effects of week, $F_{\text{Week}}(1, 15) = 14.181$, $p = .002$, $\eta_p^2 = .486$, talker, $F_{\text{Talker}}(3, 45) = 5.452$, $p = .003$, $\eta_p^2 = .267$, and stimulus type, $F_{\text{Type}}(7, 105) = 6.041$, $p < .001$, $\eta_p^2 = .287$. The mean accuracy scores for the first week and second week were 2311.82 milliseconds and 1942.32 milliseconds respectively; RT of the second week was significantly faster than the first week. The mean RT scores for each talker were 2358.48 milliseconds (Talker2), 2074.56 milliseconds (Talker3), 2111.23 milliseconds (Talker4), and 1964.01 milliseconds (Talker5). The results of pairwise comparisons with Bonferroni correction revealed that Talker2 was significantly different from Talker5 ($p = .003$). The learners demonstrated faster RTs for tokens produced by Talker5 than Talker2. Table 35 shows mean RTs for each token in Group IV. Results of pairwise comparisons with Bonferroni correction indicated that (1) ST20 was significantly different from ST15 ($p = .004$), ST17 ($p < .001$), and ST19 ($p = .002$); and (2) ST18 was significantly different from ST15 ($p = .029$), ST17 ($p = .001$), and ST19 ($p = .003$). Thus, RTs for both ST18 and

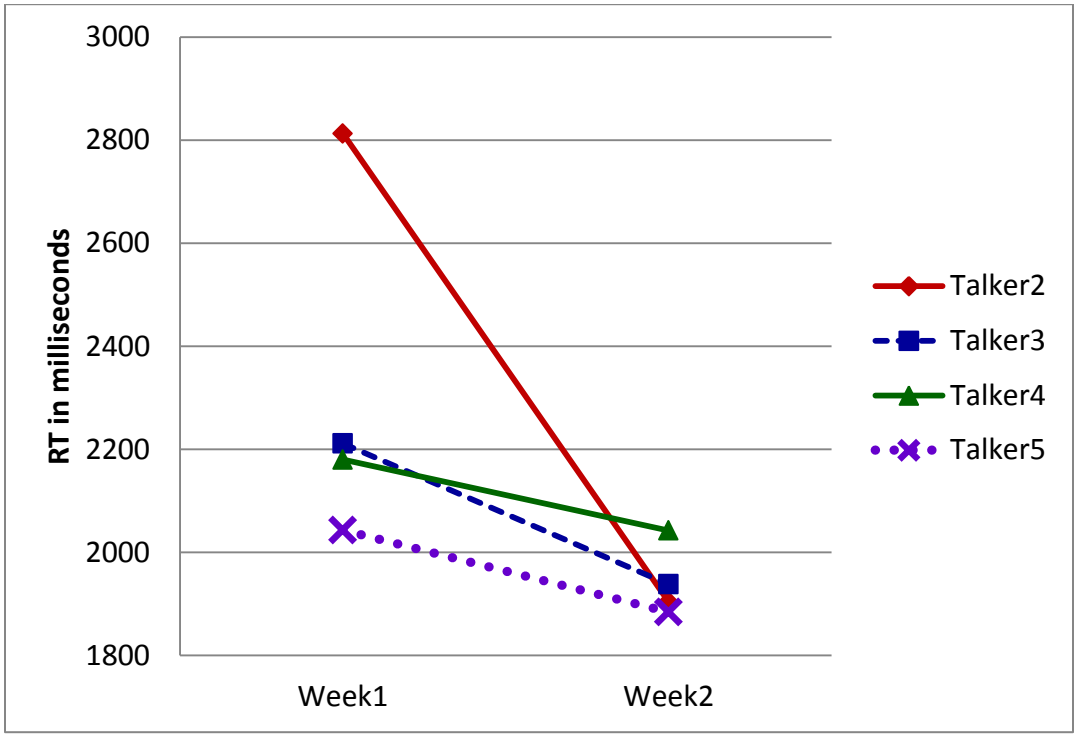
ST20 were significantly slower than ST15, ST17, and ST19; the tokens with the H.L pitch pattern had a tendency to have longer RTs than ones with the L.H pitch pattern.

Table 35: Mean RT scores of the eight tokens in Group IV (CV.CV) (AV group)

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)
15	ka.ka (L.H)	1947.84
16	ka.ka (H.L)	2167.38
17	sa.sa (L.H)	1885.06
18	sa.sa (H.L)	2394.39
19	ku.ku (L.H)	2003.44
20	ku.ku (H.L)	2437.48
21	su.su (L.H)	2047.84
22	su.su (H.L)	2133.16

In addition to the main effects above, the Week x Talker interaction was significant: F (3, 45) = 6.672, p = .001, η_p^2 = .308 (Figure 53). Results of simple effects tests indicated that the difference in RT between Talker2 and other talkers was significant in the first week, compared to the second week. The learners demonstrated slower RTs for tokens produced by Talker2 in the first week; however it was shortened significantly in the second week.

Figure 53: The comparison of perception RT of tokens in Group IV (CV.CV) for AV training group



Perception RT in Training – A-only Group: A three-way ANOVA was performed to examine the development of perception RT and effects of the factors for the A-only group. The independent variables were week (2: Week1, Week2), talker (4: Talker2, 3, 4,5), and stimulus type. The dependent variable was perception RT in the eight training sessions.

Regarding the tokens in Group I (CVV.CVV), results indicated significant main effects of week, $F_{\text{Week}}(1, 15) = 8.683$, $p = .010$, $\eta_p^2 = .367$, and stimulus type, $F_{\text{Type}}(4, 60) = 6.661$, $p < .001$, $\eta_p^2 = .308$; however, talker was not significant, $F_{\text{Talker}}(3, 45) = 1.720$, $p = .176$. The mean RT scores for the first week and second week were 3004.08 milliseconds and 2645.49 milliseconds respectively. The RT in the second week was significantly faster than one in the

first week. The mean RT scores for each talker were 2967.48 milliseconds (Talker2), 2927.73 milliseconds (Talker3), 2787.69 milliseconds (Talker4), and 2616.23 milliseconds (Talker5), and there were no significant differences among them. Table 36 shows mean RT scores for each stimulus type in Group I. Results of pairwise comparisons with Bonferroni correction indicated that ST4 was significantly different from ST3 (p = .010) and ST5 (p = .046). The learners' response latency for ST4 was significantly faster than for ST3 and ST5. Thus, the learners responded more quickly to the token with the LH.HH pitch, the consonant /k/, and the vowel /u/ than the token with the LH.HL pitch, the consonant /k/ or /s/, and the vowel /a/ or /u/.

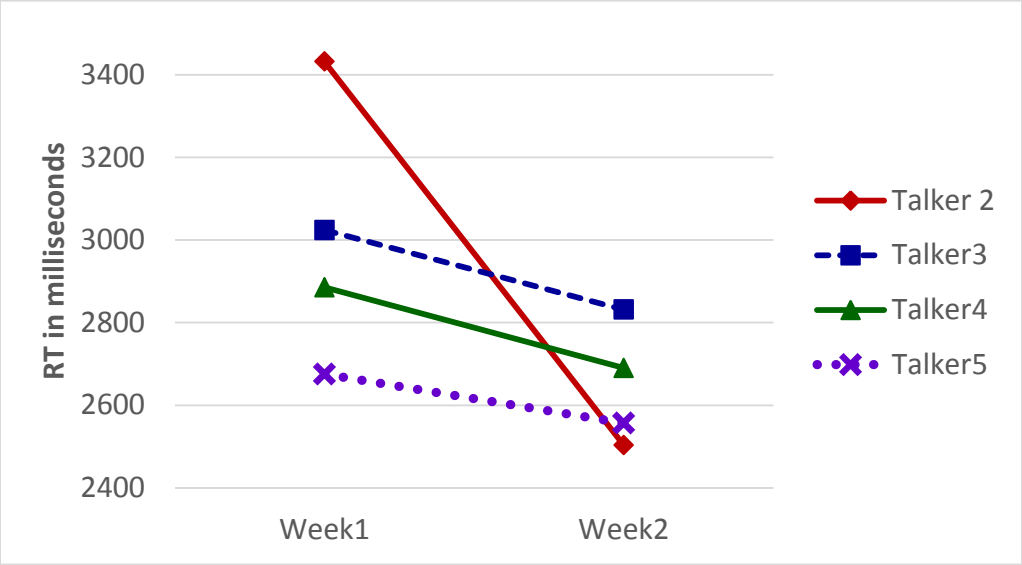
Table 36: Mean RT scores of the five tokens in Group I (CVV.CVV) (A-only group)

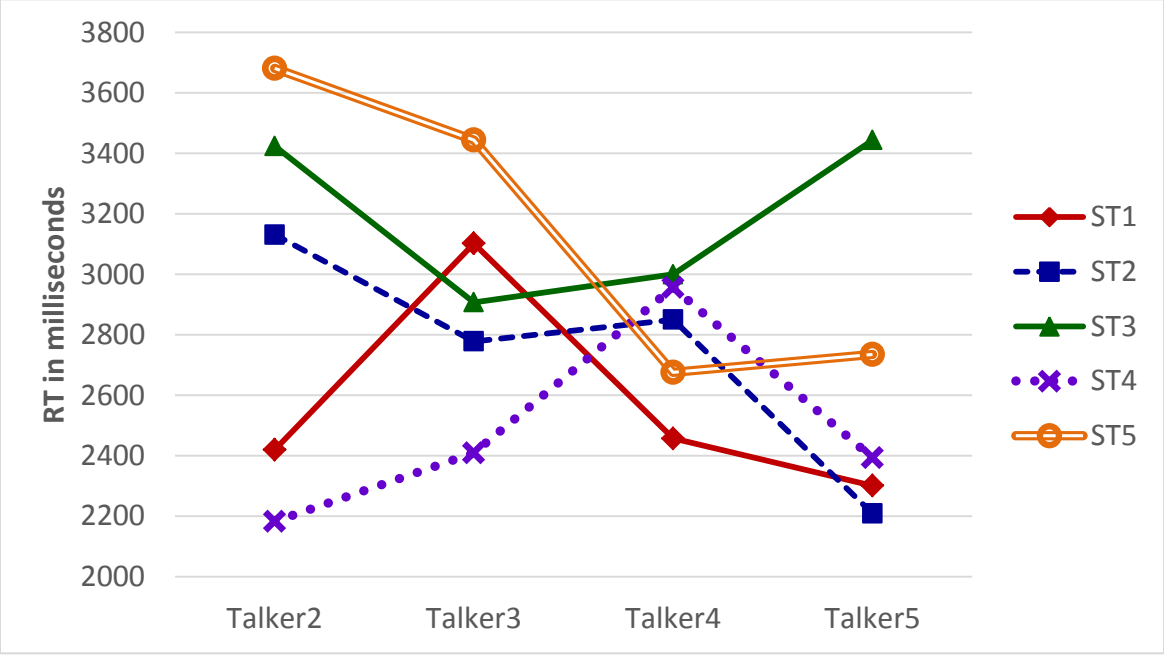
Stimulus Type (ST)	Tokens	Mean RT Scores (milliseconds)
1	kaakaa (LH.HH)	2342.13
2	kaa.kaa (HL.LL)	2555.11
3	saa.saa (LH.HL)	2693.26
4	kuu.kuu (LH.HH)	2429.35
5	suu.suu (LH.HL)	2814.21

In addition to the main effects, the Week x Talker interaction was significant, F (3, 45) = 4.312, p = .011, η_p^2 = .216 (Figure 54). The results of simple effects tests revealed that the difference between Talker2 and the other talkers was significant in the first week compared to the second week. The learners showed longer RTs for tokens produced by Talker2 than the other three talkers in the first week; however, the difference was not significant in the second week because the RT of the Talker2 was significantly shortened in the second week. In addition, the Talker x Stimulus Type interaction, F (12, 180) = 4.387, p < .001, η_p^2 = .226, was significant. Results of simple effects tests indicated that (1) the differences between ST5 and ST1 as well as ST4 were

greater with Talker2, compared to the other three talkers; and (2) the differences between ST3 and ST1 as well as ST4 were greater with Talker5. Thus, the learners showed slower RTs with ST5 produced by Talker2 and with ST3 produced by Talker5.

Figure 54: The comparisons of perception RT of tokens in Group I (CVV.CVV) for A-only training group





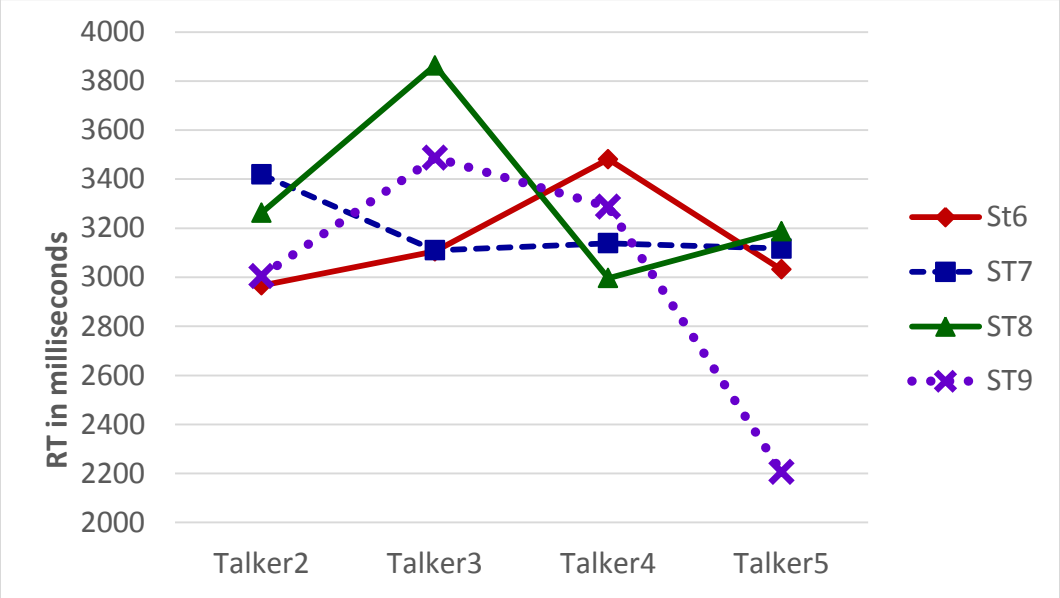
Regarding the tokens in Group II (CVV.CV), results indicated significant main effects of talker, $F_{\text{Talker}}(3, 45) = 3.410$, $p = .025$, $\eta_p^2 = .185$; however, week, $F_{\text{Week}}(1, 15) = 3{,}970$, $p = .065$, and stimulus type, $F_{\text{Type}}(3, 45) = 1.644$, $p = .193$, were not significant. The mean RT scores for the first week and second week were 3311.05 milliseconds and 3021.92 milliseconds respectively. The second week revealed faster RTs than the first week; however, the difference was not significant. The mean RT scores for each voice were 3163.13 milliseconds (Talker2), 3391.39 milliseconds (Talker3), 3226.11 milliseconds (Talker4), and 2885.31 milliseconds (Talker5). Results of pairwise comparisons with Bonferroni correction did not detect significant differences among the four talkers; however, the difference between Talker3 and Talker5 was approaching significance ($p = .075$). Table 37 shows mean RT scores for each stimulus type in Group II. There were no significant differences among the four stimulus types.

Table 37: Mean RT scores of the four tokens in Group II (CVV.CV) (A-only group)

Stimulus Type (ST)	Tokens	Mean RT Scores (milliseconds)
6	kaa.ka (HL.L)	3146,13
7	saa.sa (LH.H)	3196.09
8	saa.sa (HL.L)	3327.49
9	suu.su (LH.H)	2996.23

In addition to the main effects, the Talker x Stimulus Type interaction was significant: $F(9, 135) = 2.908$, $p = .004$, $\eta_p^2 = .162$ (Figure 55). Results of the simple effects tests indicated that the difference of ST 8 and ST9 was greatest for Talker5.

Figure 55: The comparison of perception RT of tokens in Group II (CVV.CV) for A-only training group



Regarding the tokens in Group III (CV.CVV), results indicated significant main effects of talker, $F_{Talker}(3, 45) = 7.610$, $p < .001$, $\eta_p^2 = .337$, and stimulus type, $F_{Type}(4, 60) = 8.414$, $p < .001$, week, $\eta_p^2 = .359$; however, week was not significant, $F_{Week}(1, 15) = 3.185$, $p = .095$. None of the interactions was significant. The mean RT scores for the first week and second week were 2945.29 milliseconds and 2736.69 milliseconds respectively. RT for the second week was faster than the first week; however, the difference was not significant. The mean RT scores for each voice were 3047.54 milliseconds (Talker2), 2919.75 milliseconds (Talker3), 2797.11 milliseconds (Talker4), and 2599.56 milliseconds (Talker5). Results of the pairwise comparisons with Bonferroni correction indicated that Talker5 was different from Talker2 ($p = .008$) and Talker3 ($p = .038$). Thus, the learners had faster response latency for Talker5, a female talker, compared to Talker2, another female talker, and Talke3, a male talker. Table 38

shows the mean RT scores for each token in Group III. Results of pairwise comparison with Bonferroni correction revealed that (1) ST10 was significantly different from ST11 (p = .001), ST12 (p = .003), and ST13 (p = .003); (2) ST11 was significantly different from ST12 (p = .024). The differences between ST12 and ST13 (p = .051) as well as ST10 and ST14 (p = .055) were marginally significant. Thus, the learners demonstrated faster RTs for tokens with the L.HH pitch than the L.HL or H.LL pitch patterns. Also, with the L.HL pitch pattern, the learners demonstrated faster RTs when the combination of consonant and vowel was /ku/ than /sa/.

Table 38: Mean RT scores of the five tokens in Group III (CV.CVV) (A-only group)

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)
10	ka.kaa (L.HH)	2464.07
11	sa.saa (L.HL)	3051.66
12	ku.kuu (L.HL)	2730.14
13	ku.kuu (H.LL)	3061.66
14	su.suu (L.HH)	2897.41

Regarding the tokens in Group IV (CV.CV), results indicated significant main effects of week, $F_{\text{Week}}(1, 15) = 29.426$, $p < .001$, $\eta_p^2 = .662$, talker, $F_{\text{Talker}}(3, 45) = 6.095$, $p = .001$, $\eta_p^2 = .289$, and stimulus type, $F_{\text{Type}}(7, 105) = 5.372$, $p < .001$, $\eta_p^2 = .264$. The mean RT scores for the first week and second week were 2587.11 milliseconds and 2135.21 milliseconds respectively; RTs for the second week were significantly faster than the first week. The mean RT scores for each talker were 2691.94 milliseconds (Talker2), 2184.36 milliseconds (Talker3), 2399.96 milliseconds (Talker4), and 2168.39 milliseconds (Talker5). The results of pairwise comparisons with Bonferroni correction revealed that Talker2 was significantly different from Talker3 (p = .018). The difference between Talker2 and Talker5 was marginally significant (p

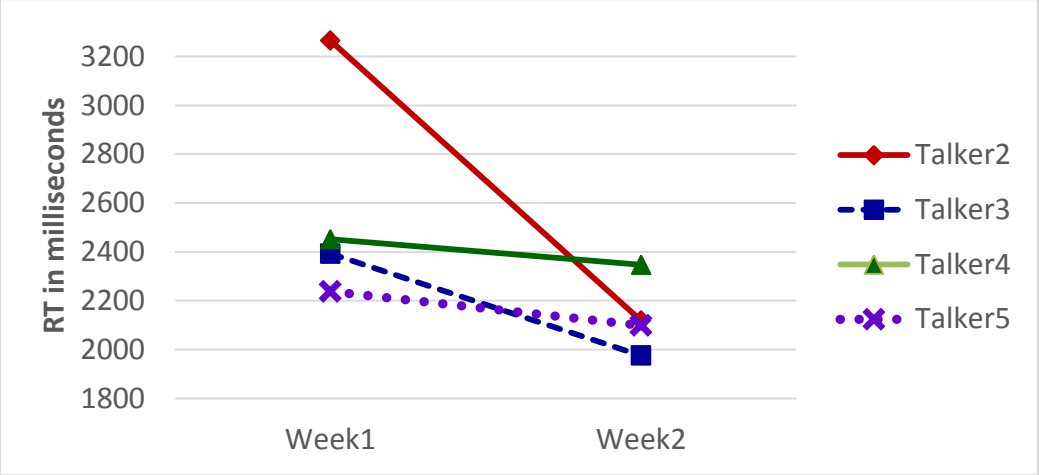
= .051). The learners demonstrated faster RTs for tokens produced by Talker3 than Talker2. Table 39 shows mean RT scores for each token in Group IV. Results of pairwise comparisons with Bonferroni correction indicated that the difference between ST17 and ST18 was significant (p = .050). Thus, the learners demonstrated faster RTs for the token with L.H pitch with /sa/ than one with H.L pitch with the same combination of consonant and vowel.

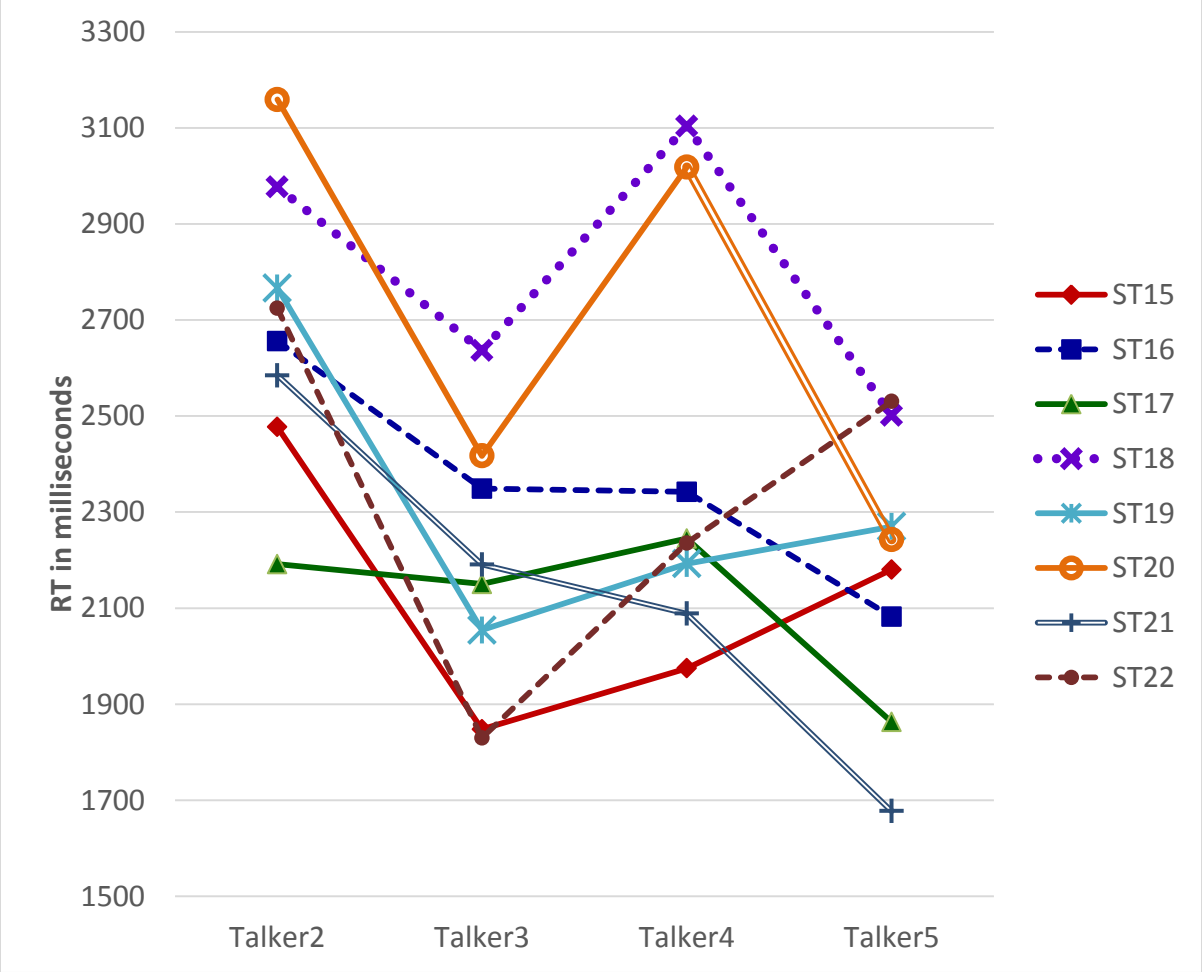
Table 39: Mean RT scores of the eight tokens in Group IV (CV.CV) (A-only group)

Stimulus Type (ST)	Token	Mean RT Scores (milliseconds)
15	ka.ka (L.H)	2119.91
16	ka.ka (H.L)	2357.30
17	sa.sa (L.H)	2112.29
18	sa.sa (H.L)	2804.39
19	ku.ku (L.H)	2320.82
20	ku.ku (H.L)	2709.23
21	su.su (L.H)	2135.37
22	su.su (H.L)	2329.98

In addition to the main effects above, the Week x Talker interaction was significant: F (3, 45) = 12.816, p < .001, η_p^2 = .461 (Figure 56). Results of simple effects tests indicated that the differences in RTs between Talker2 and other talkers were significant in the first week, compared to the second week. The learners demonstrated slower RTs for tokens produced by Talker2 in the first week; however it was shortened significantly in the second week. The Talker x Stimulus Type interaction was also significant: F (21, 315) = 1.715, p = .027, η_p^2 = .103. Results of the simple effects tests indicated that the differences between ST20 and ST15, ST19, ST21, and ST22 were greater with Talker4 than with the other three talkers. Thus, the learners demonstrated slower RTs when they identified ST20 produced by Talker4.

Figure 56: The comparisons of perception RT of tokens in Group IV (CV.CV) for A-only training group





TG with novel tokens – Comparison of Production Accuracy: A production TG was also given in order to assess whether the effects of perceptual training that had transferred to production could be generalized to the production of novel tokens. The three raters who rated the pretest and the posttest rated the TG, using the same procedures. Interrater reliability was checked using Pearson Correlation/Coefficient. There was a significant positive correlation between Rater 1 and Rater 2 (r = .915, p = .001, R^2 = .84), between Rater 1 and Rater 3 (r = .920, p = .001, R^2 = .85), as well as between Rater 2 and Rater 3 (r = .961, p = .001, R^2 = .92); the correlation was strong. Table 40 shows descriptive statistics for production accuracy scores in the pre-/post-tests and in the TG for each training group; Table 41 below shows production errors that the learners made during the TG.

Table 40: Descriptive Statistics (mean, SD) of the production accuracy in the pretest, posttest, and TG

Stimulus Type	Pretest		Posttest		TG	
	AV	A-only	AV	A-only	AV	A-only
CVV.CVV	70.31% (27.72)	78.13% (25.62)	75.56% (23.21)	75.00% (24.15)	87.50% (20.64)	85.42% (27.13)
CVV.CV	87.50% (22.36)	82.81% (23.66)	93.75% (19.37)	100.00% (.00)	97.92% (8.33)	100.00% (.00)
CV.CVV	60.94% (37.60)	51.56% (37.05)	95.31% (10.08)	89.06% (30.23)	93.75% (18.13)	85.42% (32.13)
CV.CV	59.38% (42.70)	64.06% (35.32)	92.19% (17.60)	95.31% (10.08)	91.67% (19.24)	97.92% (8.33)

Table 41: Errors observed in the production data in Experiment 2 (TG)

Token with /e/	Errors	Number	Token with /a/	Errors	Number
seesee	seese	8	taataa	taata	3
seese	sese	1			
sesee	seesee	1	tataa	tattaa	2
	sessee	6		tuutuu	3
sese	sesee	3	tata	tataa	2

First, the pretest scores were compared to the TG scores using a mixed ANOVA in order to examine whether there were any improvements in correctly producing vowel duration for the novel tokens. Independent variables were test (2; Pretest, TG), token type (4: CVV.CVV, CVV.CV, CV.CVV, CV.CV), and group type (2; AV, A-only); dependent variables were production accuracy in pretest and TG. First, the tokens with /ka/ in the pretest and /ta/ in the TG (a new consonant and a familiar vowel) were compared. The results of a mixed ANOVA indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 22.845$, $p < .001$, $\eta_p^2 = .432$, and token type, $F_{\text{Test}}(3, 90) = 3.913$, $p = .011$, $\eta_p^2 = .115$; however, group type was not significant, $F_{\text{Group}}(1, 30) = 2.028$, $p = .165$. None of the interactions was significant. Since the mean accuracy of TG was higher (.95) than that of the pretest (.65), there was improvement. In addition, among the four token types, there was a significant difference between CVV.CV and CV.CVV. The CVV.CV type had a higher mean accuracy (.86) than the CV.CVV type (.72); therefore, CVV.CV was easier to produce.

Second, the tokens with /sa/ in the pretest and /se/ in the TG (a familiar consonant and a new vowel) were compared. The results of a mixed ANOVA indicated significant main effects of test, $F_{Test}(1, 30) = 40.814$, $p < .001$, $\eta_p^2 = .576$; however, token type, $F_{Type}(3, 90) = 1.412$, $p = .245$, and group type, $F_{Group}(1, 30) = .028$, $p = .864$, were not significant. None of the interactions was significant. Since the mean accuracy of TG was higher (.95) than that of the pretest (.64), there was improvement.

Next, the production accuracy in the posttest and the TG were compared to examine whether the two tests were comparable. Independent variables were test (2; Posttest, TG), token type (4: CVV.CVV, CVV.CV, CV.CVV, CV.CV), and group type (2; AV and A-only); dependent variable were production accuracy in the posttest and TG. First, the tokens with /ka/ in the pretest and /ta/ in the TG (a new consonant and a familiar vowel) were compared. The results of a mixed ANOVA indicated no significant main effects: test, $F_{Test}(1, 30) = .717$, $p = .407$, token type, $F_{Type}(3, 90) = 1.725$, $p = .168$, and group type, $F_{Group}(1, 30) = 1.788$, $p = .191$. None of the interactions was significant. Since there was no significant difference between the two tests, it was concluded that they were comparable.

Second, the tokens with /sa/ in the pretest and /se/ in the TG (a familiar consonant and a new vowel) were compared. The results of a mixed ANOVA indicated a significant main effect of token type, $F_{Type}(1, 30) = 3.533$, $p = .018$, $\eta_p^2 = .105$; however, test, $F_{Test}(1, 30) = 1.364$, $p = .252$, and group type, $F_{Group}(1, 30) = 2.647$, $p = .114$, were not significant. None of the interactions was significant. Among the four token types, there was a significant difference between CVV.CVV and CVV.CV ($p = .029$); the CVV.CV was easier to produce than

CVV.CVV. Since there was no significant difference between the two tests, it was concluded that they were comparable.

Overall Effects of TG (familiar and novel tokens) – Perception Accuracy: Tests of generalizations (TGs) were given to the two experimental groups, in order to assess whether the effects of perceptual training on correctly identifying duration of vowels could be generalized to novel tokens (Appendix I) spoken by a familiar talker (TG1) and familiar tokens (i.e., tokens used in testing; Appendix E) spoken by a novel talker (TG2). Table 42 shows descriptive statistics of perception accuracy in the pre-/post-tests as well as in the TGs for each experimental group.

Table 42: Descriptive Statistics for the perception accuracy in pretest, posttest, and two TGs

Group	Sample Size	Pretest		Posttest		TG1 (novel tokens)		TG2 (novel voice)	
		Mean %	(*SD*)	Mean %	(*SD*)	Mean %	*SD*	Mean %	(*SD*)
AV	16	68.75	(16.21)	96.53	(8.58)	93.36	(7.02)	92.71	(9.01)
A-only	16	71.18	(14.94)	87.50	(12.91)	89.06	(12.40)	88.89	(8.84)

The two TGs were compared with the pretest in order to examine whether there were any improvements in correctly identifying vowel duration from the pretest to TGs. In order to examine the overall effects of pretest to TG1 (novel tokens), a mixed ANOVA was performed. Independent variables were test (2: pretest, TG1) and group type (2: AV, A-only); the dependent variable was perception accuracy. Results indicated significant main effects of test, $F_{\text{Test}}(1, 30)$

= 108.167, $p < .001$, $\eta_p^2 = .783$; however, group type was not significant, $F_{Group}(1, 30) = .050$, $p = .824$. Perception accuracy of novel tokens in TG1 exceeded that in the pretest. The Test x Training Modality interaction was not significant, $F(1, 30) = 2.711$, $p = .110$.

In order to examine the overall effects of pretest to TG2 (novel talker), a mixed ANOVA was performed. Independent variables were test (2: pretest, TG2) and training type (2: AV, A-only); the dependent variable was perception accuracy. Results indicated significant main effects of test, $F_{Test}(1, 30) = 88.889$, $p < .001$, $\eta_p^2 = .748$; perception accuracy also increased for stimuli produced by a new voice. However, group type was not significant, $F_{Group}(1, 30) = .032$, $p = .860$. The Test x Training Modality interaction was also not significant, $F(1, 30) = 2.000$, $p = .168$.

In addition, the two TGs were compared with the posttest in order to examine whether the posttest improvement following training could be generalized to novel tokens and a new talker. In order to examine whether the posttest and TG1 were comparable, a mixed ANOVA was performed. Independent variables were test (2: posttest, TG1) and group type (2: AV, A-only); the dependent variable was perception accuracy. Results indicated no significant main effects of test, $F_{Test}(1, 30) = .438$, $p = .513$, or group type, $F_{Group}(1, 30) = 3.586$, $p = .068$. The Test x Training Modality interaction was also not significant, $F(1, 30) = 3.800$, $p = .061$.

In order to examine whether the posttest and TG2 were comparable, a mixed ANOVA was performed. Independent variables were test (2: posttest, TG2) and group type (2: AV, A-only); the dependent variable was perception accuracy. Results indicated no significant main effect of test, $F_{Test}(1, 30) = .786$, $p = .382$; however, group type was marginally significant,

$F_{Group}(1, 30) = 3.890$, $p = .058$. The Test x Training Modality interaction was approaching significance, $F(1, 30) = 3.610$, $p = .067$.

Thus, overall, there was accuracy development from the pretest to the TG1 (novel tokens) and TG2 (novel voice). In addition, the two TGs were comparable to the posttest; therefore, the training effects were generalized to novel tokens and a novel talker. In order to examine the effects of pitch pattern, preceding consonant, and vowel type, tokens in TG1 were divided into three groups used earlier (see Table 43). Each token in the TG contained a /s/ (familiar) + /e/ (novel) or /t/ (novel) + /b/ (familiar) consonant/vowel combination. The tokens in the TG1 were compared with ones in the pretest/posttest in the following way.

Table 43: List of stimulus type in TG1

TG1 Stimuli				Pretest and Posttest Token	Group	
Stimulus Type (ST)	Token	Novel Segment	Familiar Segment			
ST1	taa.taa (LH.HL)	t	a	kaa.kaa (LH.HL)	I	CVV.CVV
ST2	see.see (LH.HH)	e	s	saa.saa (LH.HH)	I	
ST3	see.see (HL.LL)	e	s	saa.saa (HL.LL)	I	
ST4	taa.ta (LH.H)	t	a	kaa.ka (LH.H)	II	CVV.CV
ST5	taa.ta (HL.L)	t	a	kaa.ka (HL.L)	II	
ST6	see.se (LH.H)	e	s	suu.su (LH.H)	II	
ST7	see.se (HL.L)	e	s	suu.su (HL.L)	II	
ST8	ta.taa (H.LL)	t	a	ka.kaa (H.LL)	III	CV.CVV
ST9	se.see (L.HH)	e	s	sa.saa (L.HH)	III	
ST10	se.see (H.LL)	e	s	sa.saa (H.LL)	III	

Comparing Accuracy in Pretest and TG1 (Novel Tokens): Perception accuracy in the pretest and TG1 was compared using a mixed ANOVA in order to examine whether there were any developments in identifying vowel duration for the novel tokens spoken by the familiar talker

(i.e., the talker in the training sessions). In the comparison between pretest and TG1, independent variables were test (2; pretest, TG1), group type (2; AV and A-only), and stimulus type (3 or 4 depending the group); the dependent variable was perception accuracy. Regarding the tokens in Group I (CVV.CVV), the results of a mixed ANOVA indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 65.574$, $p < .001$, $\eta_p^2 = .686$; however, stimulus type, $F_{\text{Type}}(2, 60) = 2.391$, $p = .100$, and group type, $F_{\text{Group}}(1, 30) = .000$, $p = 1.00$, were not significant. The mean accuracy scores of the pretest and TG1 were .53 and .95 respectively. Thus, there was development of perception accuracy for the tokens in Group I. Table 44 below shows mean accuracy scores for each token in Group I; there were no differences among them.

Table 44: Mean accuracy scores of tokens in Group I (CVV.CVV) in the comparison between pretest and TG1

Stimulus Type (ST)	Pretest		TG1	
	Token	Mean Accuracy	Token	Mean Accuracy
ST1	kaa.kaa (LH.HL)	.44	taa.taa (LH.HL)	.97
ST2	saa.saa (LH.HH)	.66	see.see (LH.HH)	1.00
ST3	saa.saa (HL.LL)	.50	see.see (HL.LL)	.88

Regarding the tokens in Group II (CVV.CV), the results of a mixed ANOVA indicated significant main effects of stimulus type, $F_{\text{Type}}(3, 90) = 16.858$, $p < .001$, $\eta_p^2 = .360$; however, test, $F_{\text{Test}}(1, 30) = 2.301$, $p = .140$, and group type, $F_{\text{Group}}(1, 30) = .303$, $p = .586$, were not significant. None of the interactions was significant. The mean accuracy scores for the pretest and TG1 were .74 and .82 respectively. Perception accuracy scores were higher in TG1;

however, the difference between pretest and TG1 was not significant. In order to locate where differences existed among the four stimulus types, pairwise comparisons with Bonferroni correction were performed. Table 45 below shows mean accuracy scores for each token in Group II.

Table 45: Mean accuracy scores of tokens in Group II (CVV.CV) in the comparison between pretest and TG1

Stimulus Type(ST)	Pretest		TG1	
	Token	Mean Accuracy	Token	Mean Accuracy
ST4	kaa.ka (LH.H)	.81	taa.ta (LH.H)	.97
ST5	kaa.ka (HL.L)	.90	taa.ta (HL.L)	.91
ST6	suu.su (LH.H)	.81	see.se (LH.H)	.88
ST7	suu.su (HL.L)	.43	see.se (HL.L)	.53

As the mean perception scores of each token in Table 45 show, ST7 was significantly lower than ST4 ($p < .001$), ST5 ($p < .001$), and ST6 ($p = .001$). Thus, ST7 was the most difficult token to correctly perceive among the four types. ST7 was significantly more difficult to correctly identify than ST6, although they involved the same consonant and vowel but differed in pitch pattern. Also, ST5, which contained the novel consonant but had the same pitch pattern as ST7, had a higher accuracy. Therefore, the novel vowel with the HL.L pitch pattern appears to have caused the difficulty.

Regarding the tokens in Group III (CV.CVV), the results of a mixed ANOVA indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 13.364$, $p = .001$, $\eta_p^2 = .308$, and stimulus type, $F_{\text{Type}}(2, 60) = 5.955$, $p = .004$, $\eta_p^2 = .166$; however, group type was not significant, $F_{\text{Group}}(1, 30) = .000$, $p = 1.000$. The mean accuracy scores for Group III for the pretest and TG1 were .78

and .93 respectively. Thus, there was development of perception accuracy for the tokens in Group III. Stimulus type was also significant; therefore, pairwise comparisons with Bonferroni correction were performed in order to locate where differences existed among the three stimulus types. Table 46 below shows mean accuracy scores for each token in Group III.

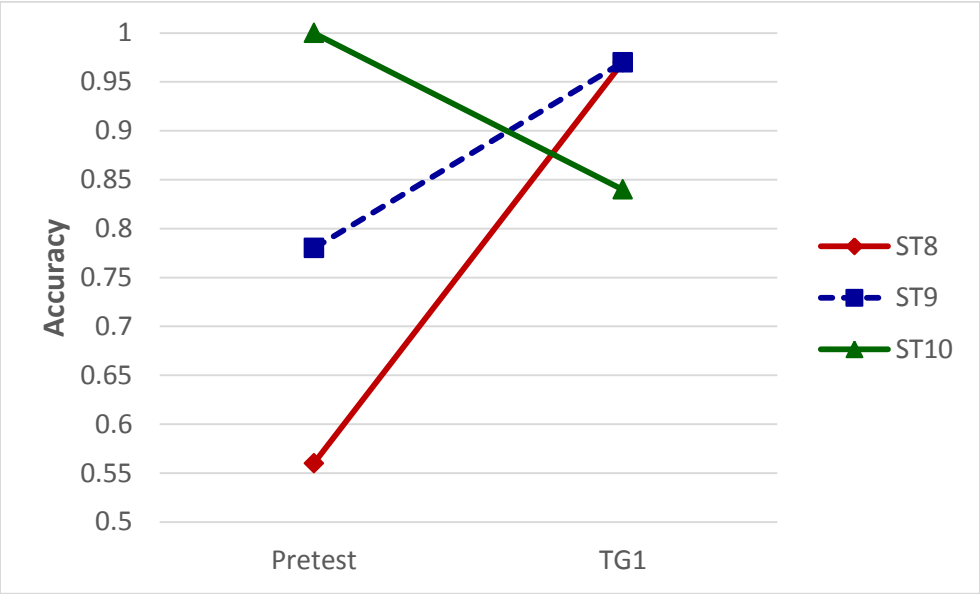
Table 46: Mean accuracy scores for tokens in Group III (CV.CVV) in the comparison between pretest and TG1

Stimulus	Pretest		TG1	
Type(ST)	Token	Mean Accuracy	Token	Mean Accuracy
ST8	ka.kaa (H.LL)	.56	ta.taa (H.LL)	97
ST9	sa.saa (L.HH)	.78	se.see (L.HH)	.97
ST10	sa.saa (H.LL)	1.00	se.see (H.LL)	.84

The results showed the difference between ST8 and ST10 was significant (p = .008), but the pitch pattern between ST8 and ST10 was identical. ST8 contained a novel preceding consonant /t/and familiar vowel /a/; ST10 contained a familiar preceding consonant /s/ and a novel vowel /e/. Thus, the learners had more difficulty identifying the vowel duration with a novel consonant.

In addition to the main effects above, the Test x Stimulus Type interaction was significant, $F(2, 60) = 10.994$, $p < .001$, $\eta_p^2 = .268$ (Figure 57). Results of simple effects tests revealed that the difference between ST8 and ST10 was significantly greater in the pretest than in TG1; the accuracy of ST10 was higher and that of ST8 was lower in the pretest. The vowel /e/ in the H.LL pitch in TG1 revealed lower accuracy than the vowel /a/ in the same pitch pattern in pretest. In contrast, the accuracy of /t/ in the H.LL pitch revealed higher accuracy than the consonant /k/ in the same pitch.

Figure 57: The comparison of perception accuracy of tokens in Group III (CV.CVV) between the pretest and TG1



Comparing Accuracy in Pretest and TG2 (Familiar Tokens by Novel Talker): Perception accuracy scores of the pretest and the TG2 were compared using a mixed ANOVA. Following the analysis of the pretest and posttest comparison, the tokens used in the TG2 were divided into three categories (Group I, II and III) as shown in Figure 34 in the previous part. Independent variables were test (2; pretest, TG2), group type (2; AV and A-only), and stimulus type (6); the dependent variable was perception accuracy. Regarding stimulus type in Group I (CVV.CVV), results of a mixed ANOVA indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 49.681$, $p < .001$, $\eta_p^2 = .623$, and stimulus type, $F_{\text{Type}}(5, 150) = 3.844$, $p = .003$, $\eta_p^2 = .114$; however, group type was not significant, $F_{\text{Group}}(1, 30) = .464$, $p = .501$. None of the interactions was significant. TG2 had a higher mean accuracy (.93) than the pretest (.62). Therefore, it was

concluded that there was a development of perception accuracy from pretest to TG2 for the tokens in Group I. In addition, stimulus type had significant effects; therefore, pairwise comparisons were performed using the Bonferroni correction in order to locate the differences. Table 47 shows the mean perception accuracy for each token in Group I. The results revealed that perception accuracy of ST3 was significantly different from ST2 (p = .007). ST2 had higher accuracy than ST3; therefore, the former was easier to identify correctly than the latter.

Table 47: Mean perception accuracy of the six stimulus type in Group I (CVV.CVV) in pretest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean Accuracy	
		Pretest	TG2
1	saa.saa (LH.HH)	.66	1.00
2	suu.suu (LH.HH)	.81	.97
3	kaa.kaa (LH.HL)	.44	.91
4	kuu.kuu (LH.HL)	.75	.91
5	saa.saa (HL.LL)	.50	.97
6	kuu.kuu (HL.LL)	.56	.88

Regarding stimulus type in Group II (CVV.CV), the results of a mixed ANOVA indicated significant main effects of stimulus test, $F_{\text{Test}}(1, 30) = 4.156$, $p = .050$, $\eta_p^2 = .122$, and stimulus type, $F_{\text{Type}}(5, 150) = 6.235$, $p < .001$, $\eta_p^2 = .172$; however, group type was not significant, $F_{\text{Group}}(1, 30) = .385$, $p = .540$. The perception accuracy significantly increased from the pretest (.76) to the TG2 (.84). In order to locate where the differences existed among the six tokens, pairwise comparisons were performed with Bonferroni correction. Table 48 shows the mean perception accuracy for each token in Group II. The results revealed that ST7 was

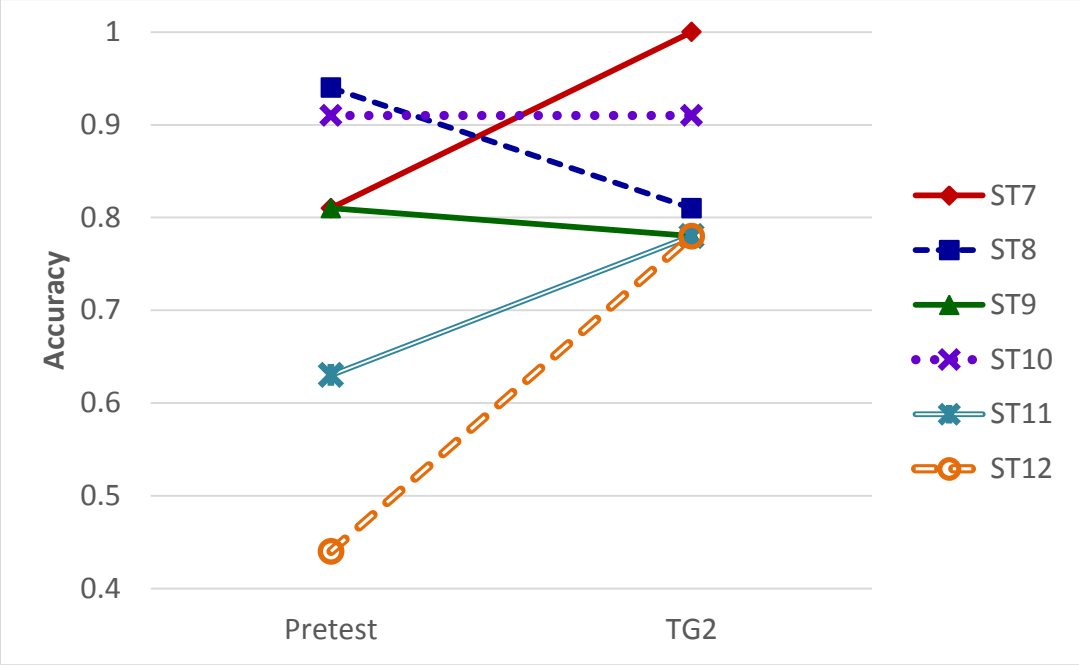
different from ST11 (p = .029) and ST12 (p = .006). In addition, ST10 was different from ST11 (p = .048) and ST12 (p = .009). The accuracy differences across the three tokens (ST10, ST11, and ST12) demonstrate that the issue is not only pitch pattern as these have the same pattern and it is not solely the consonant or vowel but on interaction of all factors.

Table 48: Mean perception accuracy of the six tokens in Group II (CVV.CV) in pretest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean Accuracy	
		Pretest	TG2
7	kaa.ka (LH.H)	.81	1.00
8	kuu.ku (LH.H)	.94	.81
9	suu.su (LH.H)	.81	.78
10	kaa.ka (HL.L)	.91	.91
11	kuu.ku (HL.L)	.63	.78
12	suu.su (HL.L)	.44	.78

In addition to the main effects, the Test x Stimulus Type interaction was significant, $F(5, 150) = 3.573$, p = .004, η_p^2 = .106 (Figure 58). Results of the simple effects tests revealed that the differences between ST7 and ST12 were greater in the pretest than TG1.

Figure 58: The comparison of perception accuracy of tokens in Group II (CVV.CV) between the pretest and TG2



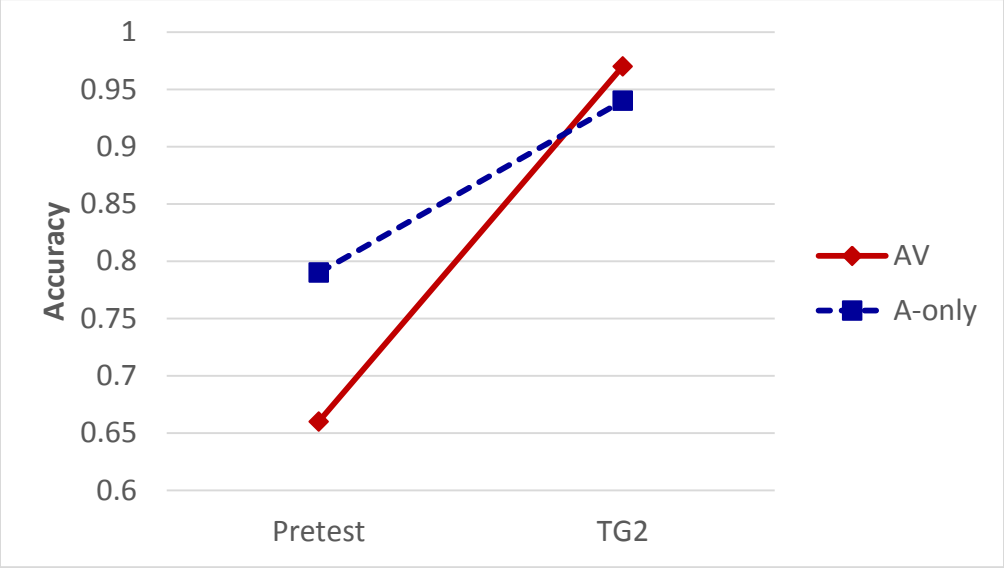
Regarding stimulus type in Group III (CV.CVV), the results of a mixed ANOVA indicated significant main effects of test, $F_{Test}(1, 30) = 34.539$, $p < .001$, $\eta_p^2 = .535$, and stimulus type, $F_{Type}(5, 150) = 3.622$, $p = .004$, $\eta_p^2 = .108$; however, group type was not significant, $F_{Group}(1, 30) = .758$, $p = .391$. The perception accuracy significantly increased from the pretest (.73) to the TG2 (.95). In order to locate differences among the six tokens, pairwise comparisons with Bonferroni correction were performed. Table 49 shows the mean accuracy of each token in Group III. The results revealed that ST16 was significantly different from ST15 ($p = .021$) and ST18 ($p = .003$).

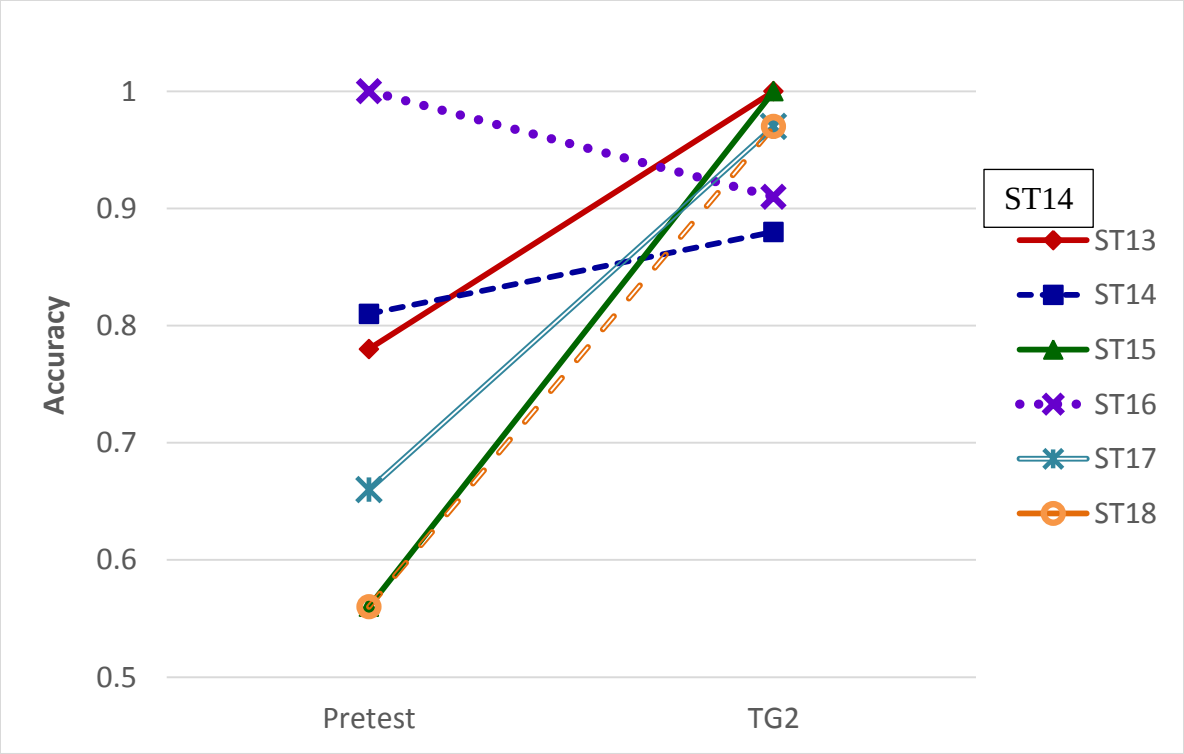
Table 49: Mean perception accuracy of the six tokens in Group III (CV.CVV) in pretest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean Accuracy	
		Pretest	TG2
13	sa.saa (L.HH)	.78	1.00
14	su.suu (L.HL)	.81	.88
15	ka.kaa (H.LL)	.56	1.00
16	sa.saa (H.LL)	1.00	.91
17	ku.kuu (H.LL)	.65	.97
18	su.suu (H.LL)	.56	.97

In addition to the main effects, the Test x Group Type interaction was significant, $F(1, 30) = 4.203$, $p = .049$, $\eta_p^2 = .123$. As shown in Figure 59, the improvement for the AV group was greater than that of the A-only group. The Test x Stimulus Type interaction was also significant, $F(5, 150) = 7.276$, $p < .001$, $\eta_p^2 = .195$. Results of the simple effects tests revealed that the accuracy of ST15, ST17, and ST18 improved the most from pretest to TG2.

Figure 59: The comparison of perception accuracy of tokens in Group II (CVV.CV) between the pretest and TG2





Comparing Accuracy in Posttest and TG1 (Novel Tokens): Perception accuracy in the posttest and TG1 was compared using a mixed ANOVA in order to examine whether the two tests were comparable (i.e., training effects were generalized to correctly identifying vowel duration of novel tokens). Independent variables were test (2; pretest, TG1), group type (2; AV and A-only), and stimulus type (3 or 4 depending the group); the dependent variable was perception accuracy in posttest and TG1. Regarding the tokens in Group I (CVV.CVV) in Table 43 in the previous part, the results of a mixed ANOVA indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 10.090$, $p = .002$, $\eta_p^2 = .216$, and group type, $F_{\text{Group}}(1, 30) = 8.710$, $p = .006$, $\eta_p^2 = .225$; however, stimulus type was not significant, $F_{\text{Type}}(2, 60) = 2.547$, $p = .087$. The mean accuracy scores of the posttest and TG1 were .90 and .98 respectively; therefore, there was development from posttest to TG1. The difference between the two training groups was significant; however, this difference was probably due to the difference in the posttest (the two groups were not homogeneous before the comparison). Table 50 below shows mean accuracy scores for each token; however, the differences were not significant.

Table 50: Mean perception accuracy of the six tokens in Group I (CVV.CVV) in posttest and TG1 comparison

Stimulus Type (ST)	Posttest		TG1	
	Token	Mean Accuracy	Token	Mean Accuracy
ST1	kaa.kaa (LH.HL)	.81	taa.taa (LH.HL)	.97
ST2	saa.saa (LH.HH)	.97	see.see (LH.HH)	1.00
ST3	saa.saa (HL.LL)	.91	see.see (HL.LL)	.88

Regarding the tokens in Group II (CVV.CV), the results of a mixed ANOVA indicated significant main effects of stimulus type, $F_{Type}(3, 90) = 14.670$, $p < .001$, $\eta_p^2 = .328$, and group type, $F_{Group}(1, 30) = 6.788$, $p = .014$, $\eta_p^2 = .328$; however, test was not significant: $F_{Test}(1, 30) = 714$, $p = .405$. The mean accuracy of the posttest was .85; that of TG1 was .82. The difference between posttest and TG1 was not significant; therefore, the tokens in Group II in the two tests were comparable. Stimulus type was significant; therefore, pairwise comparisons with Bonferroni correction were performed in order to locate where differences existed among the four stimulus types. Table 51 below shows mean accuracy scores for each token. ST7 was significantly lower than ST4 ($p < .001$), ST5 ($p < .001$), and ST6 ($p = .011$). Thus, ST7 was the most difficult token to correctly perceive among the four types.

Table 51: Mean perception accuracy of the six tokens in Group II (CVV.CV) in posttest and TG1 comparison

Stimulus	Posttest		TG1	
Type(ST)	Token	Mean Accuracy	Token	Mean Accuracy
ST4	kaa.ka (LH.H)	.91	taa.ta (LH.H)	.97
ST5	kaa.ka (HL.L)	.97	taa.ta (HL.L)	.91
ST6	suu.su (LH.H)	.81	see.se (LH.H)	.88
ST7	suu.su (HL.L)	.72	see.se (HL.L)	.53

In addition to the main effects above, the Time x Group Type interaction was significant: $F(1, 30) = 6.429$, $p = .017$, $\eta_p^2 = .176$. Among the two groups, the differences in perception accuracy of the two groups were significantly greater in TG1 than in the posttest. The AV group had significantly higher accuracy in the posttest.

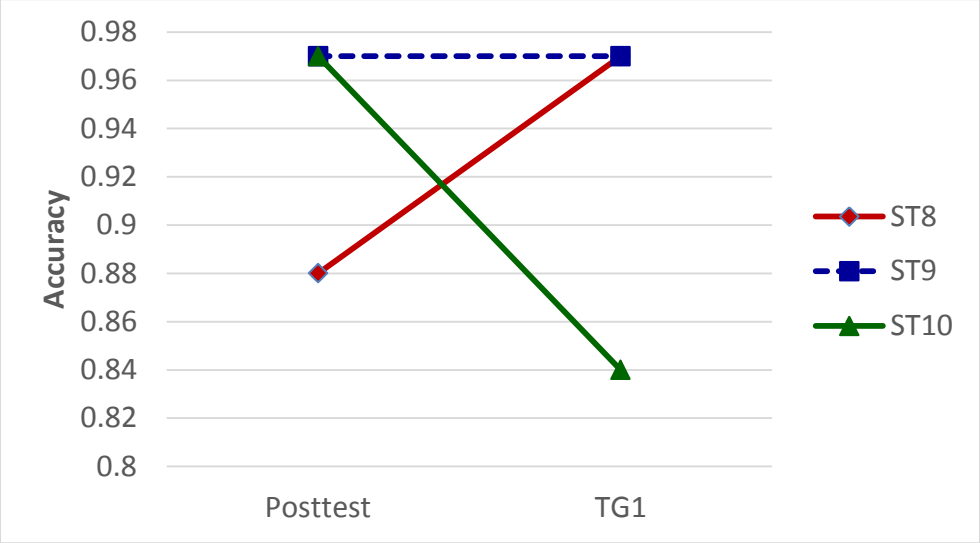
Regarding the tokens in Group III (CV.CVV), the results of a mixed ANOVA did not indicate any significant main effects: test, $F_{Test}(1, 30) = .105$, $p = .748$, stimulus type, $F_{Type}(2, 60) = 1.455$, $p = .241$, and group type, $F_{Group}(1, 30) = .034$, $p = .858$. The mean accuracy scores of the posttest and TG1 were quite higher: .94 and .93 respectively. Table 52 shows the mean accuracy of each stimulus type; there were no statistical differences among them.

Table 52: Mean perception accuracy of the six tokens in Group III (CV.CVV) in posttest and TG1 comparison

Stimulus	Posttest		TG1	
Type(ST)	Token	Mean Accuracy	Token	Mean Accuracy
ST8	ka.kaa (H.LL)	.88	ta.taa (H.LL)	.97
ST9	sa.saa (L.HH)	.97	se.see (L.HH)	.97
ST10	sa.saa (H.LL)	.97	se.see (H.LL)	.84

Although there were no significant main effects, the Test x Stimulus Type interaction was significant, $F(2, 60) = 4.549$, $p = .014$, $\eta_p^2 = .132$ (Figure 60). Results of simple effects tests revealed that the differences between ST8 and ST10 were significantly greater in TG1 than in the posttest. The accuracy of ST8 significantly developed while that of ST10 significantly decreased from the posttest to TG1.

Figure 60: The comparison of perception accuracy of the tokens in Group III (CV.CVV) between the posttest and TG1



Comparing Accuracy in Posttest and TG2 (Familiar Tokens by Novel Talker): The tokens used in the TG2 were divided into three categories (Group I, II and III) as shown in Figure 34 in the previous part. Independent variables were test (2; posttest, TG2), group type (2; AV and A-only), and stimulus type (6); the dependent variable was perception accuracy. Regarding the tokens in Group I (CVV.CVV), the results of a mixed ANOVA indicated significant main effects of stimulus type, $F_{\text{Type}}(5, 150) = 2.839$, $p = .018$, $\eta_p^2 = .086$, and group type, $F_{\text{Group}}(1, 30) = 5.867$, $p = .022$, $\eta_p^2 = .164$; however, test was not significant, $F_{\text{Test}}(1, 30) = .808$, $p = .376$. None of the interactions was significant. Since there was no difference between the two tests, it was considered that the two tests, posttest and TG2, were comparable for the tokens in Group I. Group type was significant; however, it was significant because the perception accuracy of the AV and A-only group in the posttest was significantly different: $F(1, 30) = 5.428$, $p = .027$. Pairwise comparison was performed to locate where the differences existed among the six tokens

in Group I. Table 53 shows the mean accuracy of each stimulus type in Group I. Regarding stimulus type, perception accuracy of ST1 was higher than that of ST3; however, the difference was not significant.

Table 53: Mean perception accuracy of the six stimulus type in Group I (CVV.CVV) in posttest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean Accuracy	
		Posttest	TG2
1	saa.saa (LH.HH)	.97	1.00
2	suu.suu (LH.HH)	.97	.97
3	kaa.kaa (LH.HL)	.81	.91
4	kuu.kuu (LH.HL)	.91	.91
5	saa.saa (HL.LL)	.91	.97
6	kuu.kuu (HL.LL)	.91	.88

Regarding the tokens in Group II (CVV.CV), the results of a mixed ANOVA revealed significant main effects of stimulus type, $F_{\text{Type}}(5, 150) = 3.225$, $p = .009$, $\eta_p^2 = .097$, and group type, $F_{\text{Group}}(1, 30) = 5.758$, $p = .023$, $\eta_p^2 = .161$; however, test was not significant, $F_{\text{Test}}(1, 30) = .871$, $p = .358$. The mean accuracy scores of the posttest (.88) were not significantly different from that of TG2 (.84). There were significant differences between the two groups; however, the two groups were not homogeneous at the time of posttest. Regarding token type, Table 54 below shows the mean accuracy of tokens in Group II. The results of the pairwise comparisons did not reveal any significant differences among the six token types; however, the difference between ST7 and ST12 approached significance ($p = .070$).

Table 54: Mean perception accuracy of the six tokens in Group II (CVV.CV) in posttest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean Accuracy	
		Posttest	TG2
7	kaa.ka (LH.H)	.91	1.00
8	kuu.ku (LH.H)	.97	.81
9	suu.su (LH.H)	.81	.78
10	kaa.ka (HL.L)	.97	.91
11	kuu.ku (HL.L)	.88	.78
12	suu.su (HL.L)	.72	.78

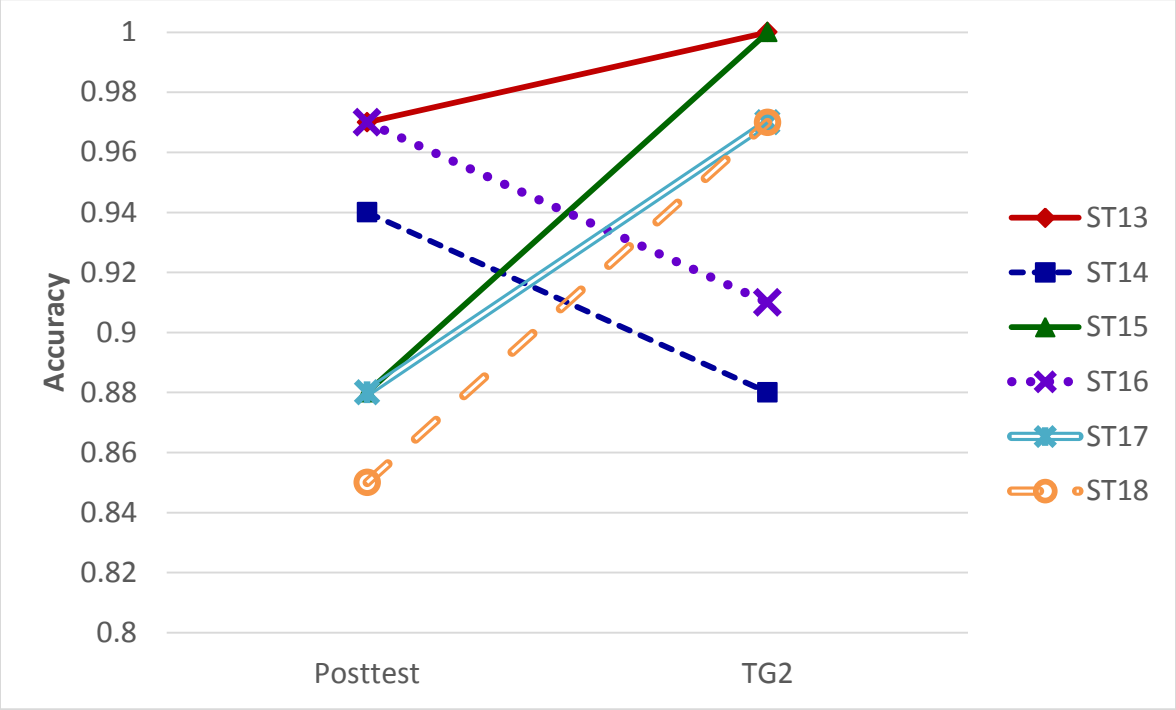
Regarding the tokens in Group III (CV.CVV), the results of a mixed ANOVA did not indicate any significant main effects: test $F_{\text{Test}}(1, 30) = 3.357$, $p = .077$, stimulus type, $F_{\text{Type}}(5, 150) = 1.447$, $p = .211$, and group type, $F_{\text{Group}}(1, 30) = .551$, $p = .464$. Table 55 shows mean accuracy scores for each token in Group III; there were no statistically significant differences among them.

Table 55: Mean perception accuracy of the six tokens in Group III (CV.CVV) in posttest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean Accuracy	
		Posttest	TG2
13	sa.saa (L.HH)	.97	1.00
14	su.suu (L.HL)	.94	.88
15	ka.kaa (L.HL)	.88	1.00
16	sa.saa (H.LL)	.97	.91
17	ku.kuu (H.LL)	.88	.97
18	su.suu (H.LL)	.85	.97

On the other hand, the Test x Stimulus Type interaction was significant, $F(5.\ 150) = 2.805$, $p < .019$, $\eta_p^2 = .86$ (Figure 61). The results of the simple effects tests revealed that (1) the difference between ST13 and ST15 was greater in the posttest than in the TG2; and (2) the difference between ST 13 and ST 14 was greater in TG2 than in posttest. The accuracy of ST15 significantly improved in the TG2; however, that of ST14 significantly decreased in the TG2.

Figure 61: The comparison of perception accuracy of tokens in Group III (CV.CVV) between the posttest and TG2



In conclusion, two tests of generalization were conducted one with novel tokens produced by a familiar voice (TG1) and one with familiar tokens produced by a new voice (TG2). First, the pretest and the two TGs were compared. Overall, it was found that accuracy improved from pretest to TG1 and TG2; however, there were some tokens that failed to generalize. In TG1,

the CVV.CV type did not demonstrate higher accuracy. In addition, it was found that generalization to a new vowel was more difficult than to a new consonant. Second, the posttest and two TGs were compared. Overall, it was found that the learners demonstrated comparable performance while there were some cases which failed the generalization. Thus, it was considered that the training effects were generalized to new tokens and a new talker. Regarding effects of the training modality on perception accuracy, there were no statistically significant differences between the two training types. However, it was found that the AV training was more effective for the development of accuracy for the most difficult one for the learners.

Test of Generalization (Familiar and Novel Tokens) – Comparison of RT: Tests of generalization were given to the two experimental groups, in order to assess whether the effect of perceptual training on the response speed to identify vowel duration could be generalized to novel tokens (Appendix I) spoken by a familiar talker (TG1) and familiar tokens (i.e., tokens used in testing; Appendix E) spoken by a novel talker (TG2). Table 56 shows descriptive statistics for the perception RT in the pre-/post-tests and two TGs.

Table 56: Descriptive Statistics of the perception RT in the pre-/post-tests, and two TGs

Group	Sample Size	Pretest		Posttest		TG1 (novel tokens)		TG2 (novel voice)	
		Mean %	(*SD*)	Mean %	(*SD*)	Mean %	*SD*	Mean %	(*SD*)
AV	16	2782.15	(557.66)	3155.17	(532.95)	2435.90	(528.33)	2392.59	(571.46)
A-only	16	2893.53	(516.01)	3241.33	(492.71)	2675.71	(477.38)	2685.66	(764.26)

The two TGs were compared with the pretest in order to examine whether there were any developments in perception RT to identify vowel duration from the pretest to TGs. In order to examine the overall effects of pretest to TG1, a mixed ANOVA was performed. Independent variables were test (2: pretest, TG1) and group type (2: AV, A-only); the dependent variable was perception RT. Results indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 6.263$, $p = .018$, $\eta_p^2 = .173$; RTs in TG1 was faster than in the pretest. However, group type was not significant, $F_{\text{Group}}(1, 30) = .394$, $p = .535$. The Test x Group Type interaction was not significant, $F(1, 30) = 1.757$, $p = .195$.

In order to examine the changes in perception RT from pretest to TG2, a mixed ANOVA was performed. Independent variables were test (2: pretest, TG2) and group type (2: AV, A-only); the dependent variable was perception RT. Results indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 5.446$, $p = .027$, $\eta_p^2 = .154$; RTs in TG2 were faster than in the pretest. However, group type was not significant, $F_{\text{Group}}(1, 30) = .492$, $p = .489$. The Test x Group Type interaction was also not significant, $F(1, 30) = 1.897$, $p = .179$.

In addition, the two TGs were compared with the posttest in order to examine whether the posttest and each TG was comparable. To compare the posttest and TG, a mixed ANOVA was performed. Independent variables were test (2: posttest, TG1) and group type (2: AV, A-only); the dependent variable was perception RT. Results indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 92.711$, $p < .001$, $\eta_p^2 = .756$; RT in TG1 were faster. However, group type was not significant, $F_{\text{Group}}(1, 30) = .796$, $p = .379$. The Test x Group Type interaction was not significant, $F(1, 30) = 1.873$, $p = .181$.

In order to examine whether the posttest and TG2 were comparable, a mixed ANOVA was performed. Independent variables were test (2: posttest, TG2) and group type (2: AV, A-only); the dependent variable was perception accuracy. Results indicated significant main effects of test, $F_{Test}(1, 30) = 28.422$, $p < .001$, $\eta_p^2 = .486$; RTs in TG2 were faster. However, group type was not significant, $F_{Group}(1, 30) = 1.038$, $p = .316$. The Test x Group Type interaction was also not significant, $F(1, 30) = .941$, $p = .340$.

Thus, overall, RT scores of from the TGs (TG1: 2555.57 milliseconds; TG2: 2539.13 milliseconds) were faster compared to the pretest (2830.94 milliseconds) and posttest (3143.38 milliseconds). In order to examine the effects of pitch pattern, preceding consonant, and vowel type, tokens in TG1 were categorized into the three groups as shown in Table 47 in the earlier part.

Comparing Perception RT in Pretest and TG1 (Novel Tokens): Perception RT in pretest and TG1 was compared using a mixed ANOVA in order to examine whether there were any developments in response speed in identifying vowel duration for the novel tokens spoken by the familiar talker (i.e., the talker in the training sessions). In the comparison between pretest and TG1, independent variables were test (2; pretest, TG1), group type (2; AV and A-only), and stimulus type (3 or 4 depending on the structural pitch pattern group); the dependent variable was perception RT. Regarding the tokens in Group I (CVV.CVV), the results of a mixed ANOVA indicated significant main effects of stimulus type, $F_{Type}(2, 60) = 19.992$, $p < .001$, $\eta_p^2 = .400$; however, test, $F_{Test}(1, 30) = 2.085$, $p = .159$, and group type, $F_{Group}(1, 30) = .349$, $p = .559$, were not significant. The mean RTs for the pretest and TG1 were 2872.84 milliseconds

and 2612.03 milliseconds. In order to locate where differences existed among the three stimulus types in Group I, pairwise comparisons with Bonferroni correction were performed. Table 57 below shows mean RT scores for each token.

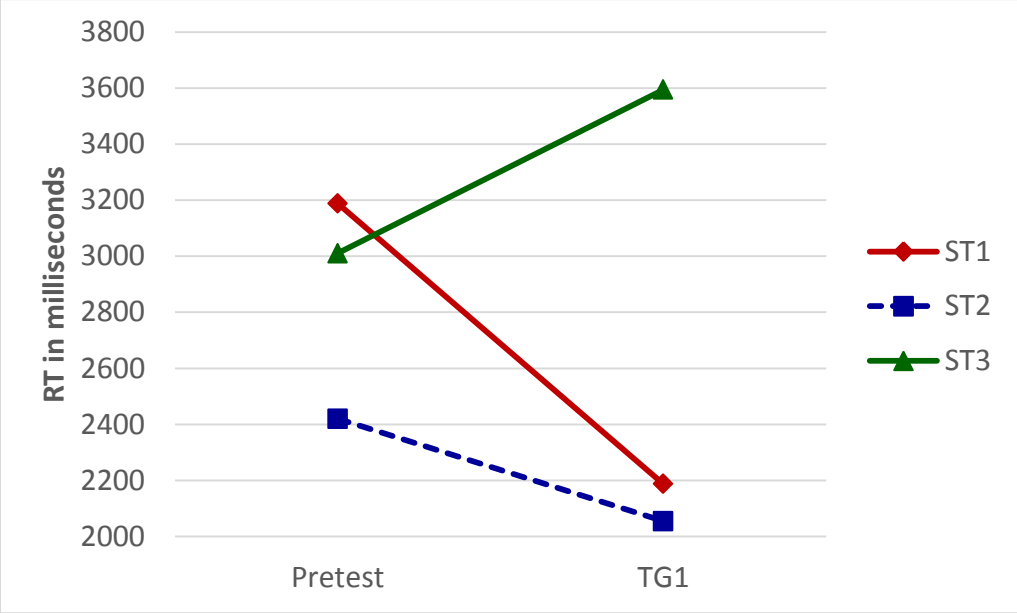
Table 57: Mean RT scores of the tokens in Group I (CVV.CVV) in the comparison between pretest and TG1

Stimulus Type (ST)	Pretest		TG1	
	Token	Mean RT (milliseconds)	Token	Mean RT (milliseconds)
ST1	kaa.kaa (LH.HL)	3188.34	taa.taa (LH.HL)	2187.50
ST2	saa.saa (LH.HH)	2420.03	see.see (LH.HH)	2054.34
ST3	saa.saa (HL.LL)	3010.16	see.see (HL.LL)	3594.25

The results showed ST3 were significantly different from ST1 ($p = .009$) and ST2 ($p < .001$). ST3 had the longest RT compared to the other two tokens. The source of the difficulty for ST3 appears to be the pitch pattern.

In addition to the main effects above, the Test x Stimulus Type interaction was significant, $F(2, 60) = 8.272$, $p = .001$, $\eta_p^2 = .216$ (Figure 62). Results of simple effects tests revealed that the differences between ST1 and ST2 were significantly greater in pretest than in TG1. The RT of ST1 as well as ST2 decreased from pretest to TG1; however, the rate of decrease was greater for ST1.

Figure 62: The comparison of perception RT for the tokens in Group I (CVV.CVV) between the pretest and TG1



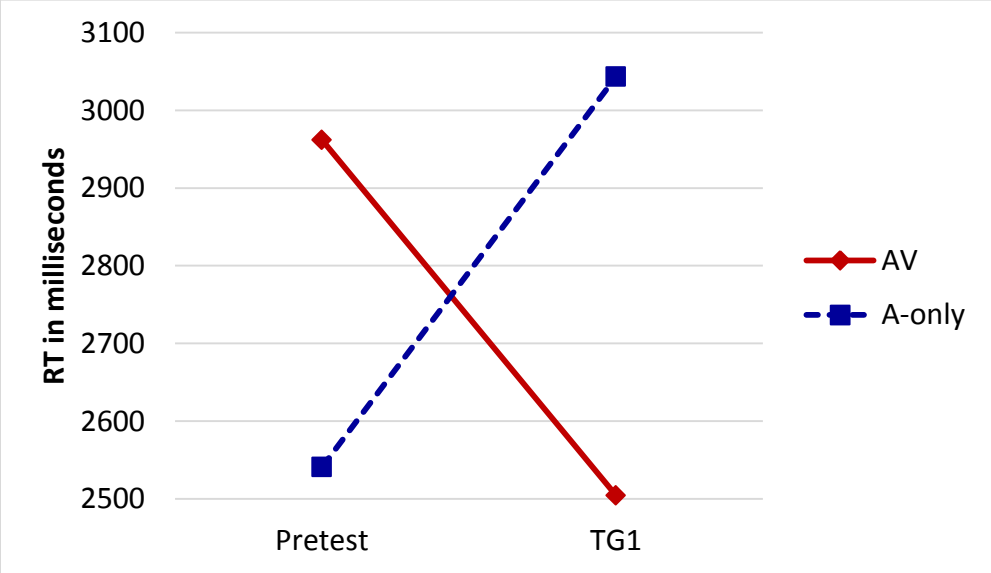
Regarding the tokens in Group II (CVV.CV), the results of a mixed ANOVA indicated significant main effects of stimulus type, $F_{\text{Type}}(3, 90) = 3.155$, $p = .029$, $\eta_p^2 = .095$; however, test, $F_{\text{Test}}(1, 30) = .031$, $p = .860$, and group type, $F_{\text{Group}}(1, 30) = .112$, $p = .740$, were not significant. The mean RTs of the pretest and TG1 were 2751.44 milliseconds and 2773.82 milliseconds respectively. The difference between pretest and TG1 was not significant. In order to locate where differences existed among the four stimulus types in Group II, pairwise comparisons with Bonferroni correction were performed. Table 58 below shows mean accuracy scores for each token. The difference between ST4 and ST5 was marginally significant ($p = .053$).

Table 58: Mean RT scores of the tokens in Group II (CVV.CV) in the comparison between pretest and TG1

Stimulus Type(ST)	Pretest		TG1	
	Token	Mean RT (milliseconds)	Token	Mean RT (milliseconds)
ST4	kaa.ka (LH.H)	2532.31	taa.ta (LH.H)	2570.75
ST5	kaa.ka (HL.L)	3070.16	taa.ta (HL.L)	2740.41
ST6	suu.su (LH.H)	2552.63	see.se (LH.H)	2847.53
ST7	suu.su (HL.L)	2850.66	see.se (HL.L)	2936.59

In addition to the significant main effects, the Test x Group Type interaction was significant: $F(1, 30) = 14.441$, $p = .001$, $\eta_p^2 = .325$ (Figure 63). The two groups had greater RT difference in TG1 compared to the pretest, and the RT of the AV group decreased while that of the A-only group increased.

Figure 63: The comparison of perception RT for the tokens in Group II (CVV.CV) between the pretest and TG1



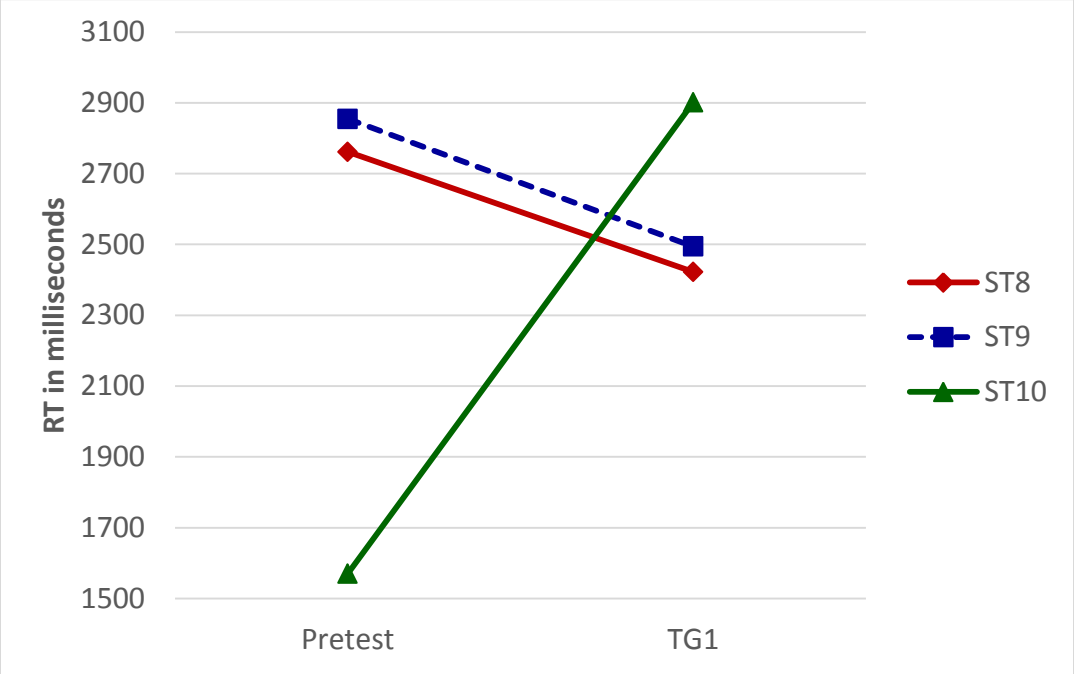
Regarding the tokens in Group III (CV.CVV), the results of a mixed ANOVA indicated significant main effects of stimulus type, F_{Type} (2, 60) = 3.296, p = .044, η_p^2 = .099; however, test, F_{Test} (1, 30) = 1.393, p = .247, and group type, F_{Group}(1, 30) = .001, p = 970, were not significant. The mean RT of the pretest was 2395.45 milliseconds; that of TG1 was 2605.95 milliseconds. The RT looks like it lengthened in the TG1; however, the difference was not significant. Stimulus type was also significant; therefore, pairwise comparisons with Bonferroni correction were performed in order to locate where differences existed among the three stimulus types. Table 59 below shows mean RT scores for each token. The results showed the difference between ST9 and ST10 was significant (p = .005), suggesting that the source was the pitch pattern. The token with the novel vowel with the H.LL pitch had significantly faster RT than the one with the L.HH pitch.

Table 59: Mean RT scores of the tokens in Group III (CV.CVV) in the comparison between pretest and TG1

Stimulus Type(ST)	Pretest		TG1	
	Token	Mean RT (milliseconds)	Token	Mean RT (milliseconds)
ST8	ka.kaa (H.LL)	2761.66	ta.taa (H.LL)	2422.38
ST9	sa.saa (L.HH)	2854.63	se.see (L.HH)	2494.38
ST10	sa.saa (H.LL)	1570.06	se.see (H.LL)	2901.09

In addition to the main effects above, the Test x Stimulus Type interaction was significant, F(2, 60) = 10.994, p < .001, η_p^2 = .268 (Figure 64).

Figure 64: The comparison of perception RT of the tokens in Group III (CV.CVV) between the pretest and TG1



Results of simple effects tests revealed that the differences between ST10 and ST8, and ST10 and ST9 were greater in pretest than in TG1. The RT of ST10 was significantly lengthened in TG1, compared to pretest, while that of ST8 and ST9 were shortened in TG1.

Comparing RT in Pretest and TG2 (Novel Talker): Perception RT scores for the pretest and the TG2 were compared using a mixed ANOVA. Following the previous analyses, the tokens used in the TG2 were also divided into three categories (Group I, II and III) as shown in Figure 34. Independent variables were test (2; pretest, TG2), group type (2; AV and A-only), and stimulus type (6); the dependent variable was perception RT. Regarding stimulus type in Group I (CVV.CVV), the results of a mixed ANOVA indicated significant main effects of test, $F_{\text{Test}}(1,$

30) = 16.465, $p < .001$, $\eta_p^2 = .345$; however, stimulus type, $F_{Type}(5, 150) = 1.661$, $p = .147$, and group type, $F_{Group}(1, 30) = 2.663$, $p = .113$, were not significant. None of the interactions was significant. The mean RT of pretest was 2926.07 milliseconds; the mean RT of TG2 was 2393.98 milliseconds. Therefore, the RT was shortened from the pretest to TG2. Table 60 shows mean RT scores for each stimulus type in Group I.

Table 60: Mean perception RT of the six stimulus type in Group I (CVV.CVV) in pretest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean RT	
		Pretest	TG2
1	saa.saa (LH.HH)	2420.03	2261.97
2	suu.suu (LH.HH)	2940.72	2323.59
3	kaa.kaa (LH.HL)	3188.34	2277.03
4	kuu.kuu (LH.HL)	3041.44	2484.28
5	saa.saa (HL.LL)	3010.16	2530.31
6	kuu.kuu (HL.LL)	2955.75	2486.72

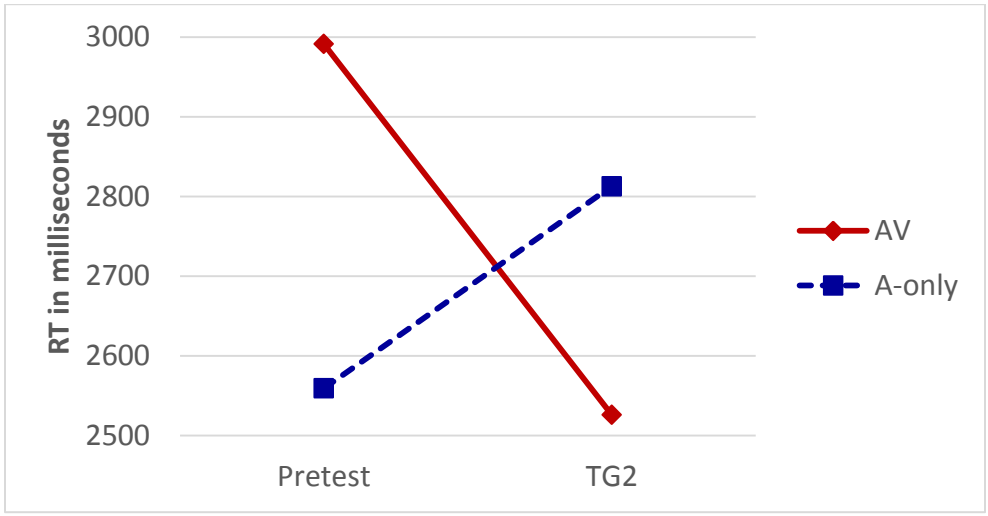
Regarding stimulus type in Group II (CVV.CV), the results of a mixed ANOVA did not indicate any significant main effects: test, $F_{Test}(1, 30) = .447$, p .509, stimulus type, $F_{Type}(5, 150) = 1.348$, $p = .223$, and group type, $F_{Group}(1, 30) = .113$, $p = .739$. The mean RT scores decreased from 2775.23 milliseconds (pretest) to 2664.20 milliseconds (TG2); however, the difference was not significant. Table 61 shows mean RT scores for each stimulus type in Group II.

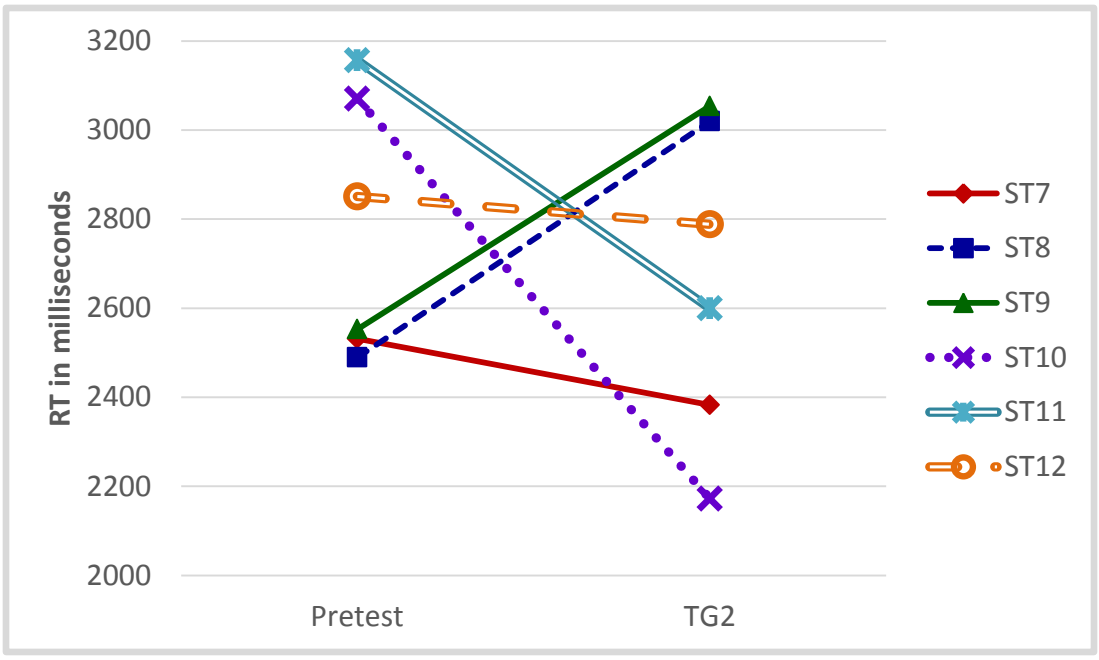
Table 61: Mean perception RT of the six tokens in Group II (CVV.CV) in pretest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean RT	
		Pretest	TG2
7	kaa.ka (LH.H)	2531.31	2382.78
8	kuu.ku (LH.H)	2489.34	3019.44
9	suu.su (LH.H)	2552.63	3053.59
10	kaa.ka (HL.L)	3070.16	2171.75
11	kuu.ku (HL.L)	3156.28	2599.19
12	suu.su (HL.L)	2850.66	2788.44

The Test x Group Type interaction was significant: $F(1, 30) = 5.132$, $p = .031$, $\eta_p^2 = .146$. As shown in Figure 65, the two training groups had greater RT difference at pretest, compared to TG2; the RT of the AV group shortened whereas that of the A-only group lengthened in TG2. The Test x Stimulus Type interaction was also significant: $F(5, 150) = 5.249$, $p < .001$, $\eta_p^2 = .149$. The results of the simple effects tests revealed that (1) the differences between ST10 and ST12 were significantly greater in TG2 than in pretest; and (2) the differences between ST10 and ST7 were significantly greater in pretest than in TG2. RT of ST10 significantly shortened from the pretest to TG2.

Figure 65: The comparison of perception RT of tokens in Group II (CVV.CV) between the pretest and TG2





Regarding stimulus type in Group III (CV.CVV), the results of a mixed ANOVA indicated significant main effects of stimulus type, $F_{Type}(5, 150) = 7.355$, $p < .001$, $\eta_p^2 = .197$; however, test, $F_{Test}(1, 30) = .218$, $p = .644$, and group type, $F_{Group}(1, 30) = .001$, $p = .973$, were not

significant. The mean RT decreased from 2624.38 milliseconds (pretest) to 2554.20 (TG2); however, the change was not significant. To locate where the differences existed among the 6 stimulus types, pairwise comparisons were performed with Bonferroni correction. Table 62 shows mean RT scores for each stimulus type in Group III.

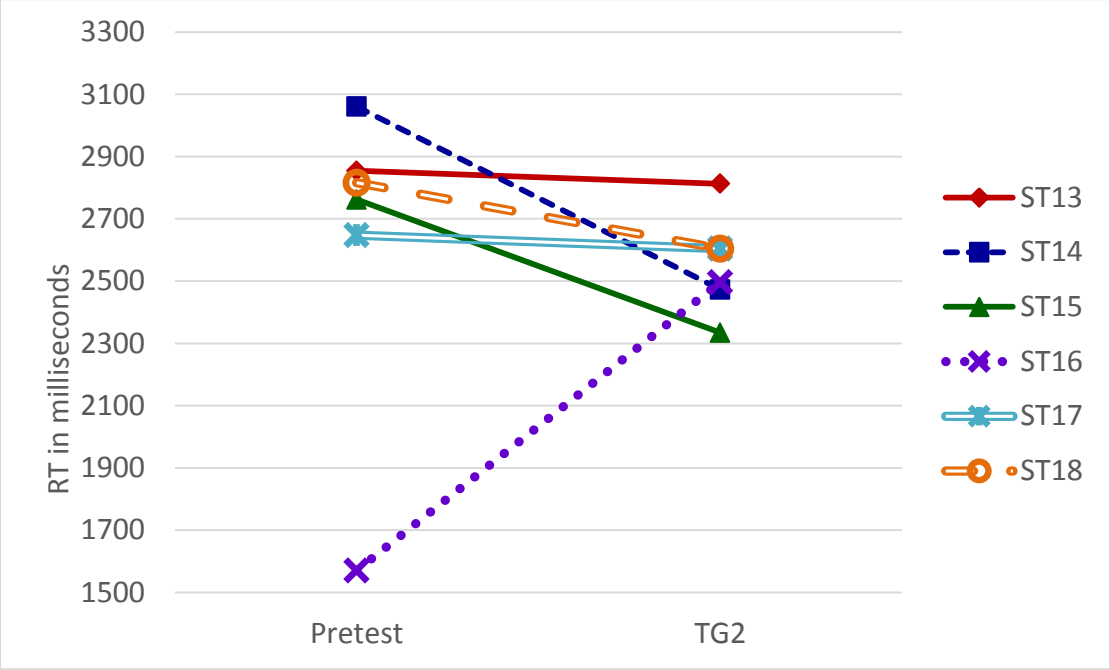
Table 62: Mean perception RT of the six tokens in Group III (CV.CVV) in pretest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)	
		Pretest	TG2
13	sa.saa (L.HH)	2854.63	2812.97
14	su.suu (L.HL)	3060.66	2472.94
15	ka.kaa (L.HL)	2761.66	2334.31
16	sa.saa (H.LL)	1570.06	2496.91
17	ku.kuu (H.LL)	2647.31	2603.94
18	su.suu (H.LL)	2815.97	2604.13

It was found that ST16 was significantly different from ST13 ($p < .001$), ST14 ($p < .001$), ST15 ($p < .022$), ST17 ($p < .001$), and ST18 ($p < .001$).

In addition to the main effects, the Test x Stimulus Type interaction was significant: $F(5, 150) = 5.657$, $p < .001$, $\eta_p^2 = .159$ (Figure 66). The results of simple effects tests revealed that the differences between ST14 and ST16 were significantly greater in pretest than in TG2. The RT of ST14 decreased from the pretest to TG2; however, that of ST16 increased.

Figure 66: The comparison of perception RT of tokens in Group III (CV.CVV) between the pretest and TG2



Comparing RT in Posttest and TG1 (Novel Tokens): Perception RTs in posttest and TG1 were compared using a mixed ANOVA in order to examine whether the two tests were comparable (i.e., training effects were generalized in response speed in identifying vowel duration of novel tokens). Independent variables were test (2; posttest, TG1), group type (2; AV and A-only), and stimulus type (3 or 4 depending the group); the dependent variable was perception RT. Regarding the tokens in Group I (CVV.CVV), the results of a mixed ANOVA indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 10.963$, $p = .002$, $\eta_p^2 = .268$, and stimulus type, $F_{\text{Type}}(2, 60) = 7.591$, $p = .001$, $\eta_p^2 = .202$; however, group type, was not significant: $F_{\text{Group}}(1, 30) = 1.734$, $p = .198$. The mean RT of the posttest was 3149.44 milliseconds; that of TG1 was 2612.03 milliseconds. The RT significantly shortened in the TG1. Stimulus type had significant

effects; therefore, pairwise comparisons with Bonferroni correction were performed in order to locate where differences existed among the three stimulus types. Table 63 shows mean RT scores for each token in Group I.

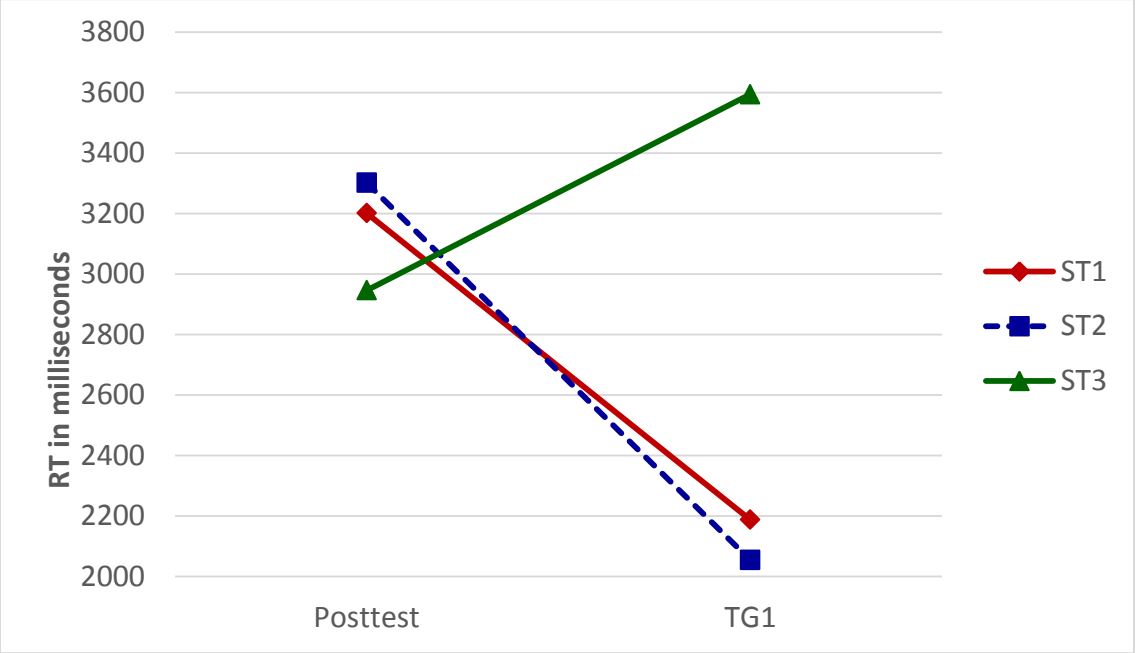
Table 63: Mean perception RT of the six tokens in Group I (CVV.CVV) in posttest and TG1 comparison

Stimulus Type (ST)	Posttest		TG1	
	Token	Mean RT (milliseconds)	Token	Mean RT (milliseconds)
ST1	kaa.kaa (LH.HL)	3201.00	taa.taa (LH.HL)	2187.50
ST2	saa.saa (LH.HH)	3301.38	see.see (LH.HH)	2054.34
ST3	saa.saa (HL.LL)	2945.94	see.see (HL.LL)	3594.25

The results showed ST3 was significantly different from ST1 (p = .008) and ST2 (p = .003). ST3 had a significantly longer RT than the other two tokens, which might be attributable to the pitch pattern.

In addition to the main effects above, the Test x Stimulus Type interaction was significant, $F(2, 60) = 13.362$, $p < .001$, $\eta_p^2 = .308$ (Figure 67). Results of simple effects tests revealed that the differences between ST3 and ST1, ST3 and ST2 were greater in TG1 than in the posttest. RT of ST1 and ST2 significantly shortened; however, that of ST3 increased.

Figure 67: The comparison of perception RT of the tokens in Group I (CVV.CVV) between the posttest and TG1



Regarding the tokens in Group II (CVV.CV), the results of a mixed ANOVA indicated significant main effects of test, $F_{Test}(1, 30) = 17.258$, $p < .001$, $\eta_p^2 = .365$; however, stimulus type, $F_{Type}(3, 90) = 1.393$, $p = .250$, and group type, $F_{Group}(1, 30) = 3.232$, $p = .082$, were not significant. None of the interactions was significant. The mean RT of the posttest was 3205.15 milliseconds; that of TG1 was 2773.82 milliseconds. The RT of all the tokens in Group II significantly shortened from the posttest to TG1.

Regarding the tokens in Group III (CV.CVV), the results of a mixed ANOVA indicated significant main effects of test, $F_{Test}(1, 30) = 19.354$, $p < .001$, $\eta_p^2 = .392$, and stimulus type, $F_{Type}(2, 60) = 5.459$, $p = .007$, $\eta_p^2 = .154$; however, group type was not significant, $F_{Group}(1, 30) = .262$, $p = .612$. None of the interactions was significant. The mean RT for the posttest was

3171.17 milliseconds; that for TG1 was 2605.95 milliseconds. The RT of all the tokens in Group II significantly shortened from the posttest to TG1. Stimulus type had significant effects; therefore, pairwise comparisons with Bonferroni correction were performed in order to locate where differences existed among the three stimulus types. Table 62 shows the mean RT of each stimulus type in Group III.

Table 64: Mean perception RT of the six tokens in Group III (CV.CVV) in posttest and TG1 comparison

Stimulus Type(ST)	Posttest		TG1	
	Token	Mean RT (milliseconds)	Token	Mean RT (milliseconds)
ST8	ka.kaa (H.LL)	3120.56	ta.taa (H.LL)	2422.38
ST9	sa.saa (L.HH)	2887.28	se.see (L.HH)	2494.38
ST10	sa.saa (H.LL)	3505.66	se.see (H.LL)	2901.09

Results indicated that ST9 and ST10 were significantly different (p = .010). The RT of ST10 was significantly longer than that of ST9 which may be attributable to the pitch pattern as the segmental information was the same.

Comparing RT in Posttest and TG2 (Novel Talker): Perception RT scores for the posttest and the TG2 were compared using a mixed ANOVA. Following previous analyses, the tokens used in the TG2 were divided into three categories (Group I, II and III) as shown in Figure 34. For each category, independent variables were test (2; posttest, TG2), group type (2; AV and A-only), and stimulus type (6); the dependent variable was perception RT. Regarding stimulus type in Group I (CVV.CVV), the results of a mixed ANOVA indicated significant main effects of test, F_{Test}(1,

30) = 46.802, $p < .001$, $\eta_p^2 = .609$; however, stimulus type, $F_{Type}(5, 150) = .320$, $p = .901$, and group type, $F_{Group}(1, 30) = 2.097$, $p = .158$, were not significant. None of the interactions was significant. The mean RT of the posttest was 2658.32 milliseconds; the mean RT of TG2 was 2943.43 milliseconds. Therefore, the RT was lengthened from the pretest to TG2.

Regarding stimulus type in Group II (CVV.CV), the results of a mixed ANOVA indicated significant main effects of test, $F_{Test}(1, 30) = 10.800$, $p = .003$, $\eta_p^2 = .265$, and stimulus type, $F_{Type}(5, 150) = 3.310$, $p = .007$, $\eta_p^2 = .099$; however, group type was not significant, $F_{Group}(1, 30) = .440$, $p = .512$. None of the interactions was significant. The mean RT of the posttest was 3125.75 milliseconds; the mean RT of TG2 was 2669.20 milliseconds. Therefore, the RT was shortened from the posttest to TG2. In order to locate where the differences existed among the six stimulus types, pairwise comparisons with Bonferroni correction were performed. Mean RT scores of each stimulus type are tabulated in Table 63 below. RTs for both ST7 and ST10 were faster than ST11; however, comparisons could not locate differences.

Table 65: Mean perception RT of the six tokens in Group II (CVV.CV) in posttest and TG2 comparison

Stimulus Type (ST)	Tokens	Mean RT (milliseconds)	
		Posttest	TG2
7	kaa.ka (LH.H)	3013.28	2382.78
8	kuu.ku (LH.H)	3315.13	3019.44
9	suu.su (LH.H)	3396.16	3053.59
10	kaa.ka (HL.L)	3253.22	2171.75
11	kuu.ku (HL.L)	2618.78	2599.19
12	suu.su (HL.L)	3157.94	2788.44

Regarding stimulus type in Group III (CV.CVV), the results of a mixed ANOVA indicated significant main effects of test, $F_{\text{Test}}(1, 30) = 18.760$, $p < .001$, $\eta_p^2 = .385$; however, stimulus type, $F_{\text{Type}}(5, 150) = 1.038$, $p = .397$, and group type, $F_{\text{Group}}(1, 30) = .1.320$, $p = .260$, were not significant. None of the interactions was significant. The mean RT of pretest was 2789.16 milliseconds; the mean RT of TG2 was 3014.15 milliseconds. Therefore, the RT was lengthened from the posttest to TG2.

In conclusion, as a result of comparing the RTs in pretest and posttest with the two TGs, it was found that the learners generally demonstrated faster RTs in TGs, compared to the pretest and posttest. Factors such as pitch patterns, vowel types, and preceding consonants affected perception accuracy and RT. However, there were not many meaningful differences between the two training groups (AV, A-only).

CHAPTER 4: DISCUSSION AND CONCLUSION

In this study, factors influencing L2 learners' perception, response latency, and production of vowel duration in Japanese were explored (Experiment 1). In addition, the efficacy of focused perceptual training on vowel duration and its influence on production were examined (Experiment 2). In this chapter, findings of Experiment 1 and 2 are discussed based on the research questions proposed for this study.

Factors Affecting Perception and Production of Vowel Duration in L2 Japanese (RQ1)

Experiment 1 examined whether preceding consonant, type of vowel, and pitch pattern for perception or token type for production had any influence on the production and perception of vowel duration in L2 Japanese. It was found that vowel type and token type significantly affected correct production of vowel duration. In general, the vowel /a/ had higher accuracy than the vowel /u/. Also, the CVV.CV token type had higher accuracy than the CV.CVV type as well as the CV.CV type. The error analysis of the token types showed that the learners had difficulties correctly producing vowel duration in the final syllable. There was an interaction between the preceding consonant and token type, which suggested that the CV.CV token with a stop consonant (/k/) had higher accuracy than that with a fricative consonant (/s/).

For the perception accuracy, the tokens used in this study were divided into four groups: (I) CVV.CVV, (II) CVV.CV, (III) CV.CVV, and (IV) CV.CV. For the tokens in Group I, it was found that pitch pattern affected perception accuracy; the LH.HH pattern had higher accuracy than LH.HL and HL.LL. For the tokens in Group II, it was found that all preceding consonants, pitch patterns, and vowel types affected perception accuracy although generally, a stop (/k/) and

a low vowel (/a/) revealed higher accuracy than a fricative (/s/) and a high vowel (/u/) respectively. In addition, the LH.H pitch pattern showed higher accuracy than the HL.L pattern. There was also an interaction between vowel type and pitch pattern; with the LH.H pitch, the vowel /a/ revealed higher accuracy than the vowel /u/. Regarding the tokens in Group III, vowel type and pitch pattern affected perception accuracy. A high vowel /a/ revealed higher accuracy than a low vowel /u/. Also, the L.HH pitch pattern showed higher accuracy than the H.LL pattern. There was an interaction among preceding consonant, vowel type, and pitch pattern; with the LH.H pitch, a combination of a consonant and vowel /ka/ showed higher accuracy than /ku/, /sa/, and /su/. Finally, for the tokens in Group IV, it was found that preceding consonant affected perception accuracy; tokens with a fricative /s/ showed higher accuracy than tokens with a stop /k/. Also, there was an interaction between vowel type and pitch pattern; with the vowel /a/, the H.L pitch showed higher accuracy than the L.H pitch.

Based on these findings regarding the pitch pattern, it was easier for the learners to correctly identify the vowel duration with the LH pitch in the first syllable and with the HH pitch in the second syllable. This finding is compatible with Minagawa (1997) who found that L2 learners including NSs of English more accurately identified long vowels with the HH pitch pattern than with the LL pitch pattern. The learners in the current study were all NSs of English; therefore, the higher pitch in word-final position may have been more perceptually salient. In addition, accented vowels, which can be perceptually salient, have higher pitch and are lengthened in a stress-timed language like English (Pennington, 1996). Therefore, it is easier for English NSs to correctly perceive the length of long vowels if high pitch is assigned. Also, the preference of high pitch on long vowels could demonstrate that the L2 learners were using English prosodic preferences when processing Japanese speech input, by associating high pitch

with an accented vowel that has longer duration. Furthermore, in English, the first syllable on many nouns and adjectives gets an accent (e.g., FA.ther) when the word does not have any prefix (Kubozono & Ohta, 1998). Therefore, the learners may have had higher accuracy with high pitch on the first syllable (i.e., CVV.CV or LH.H) versus the others (i.e., CVV.CV or HL.L).

Next, the overall findings of this study suggest that the L2 learners' perception tends to be continuous while NSs demonstrate categorical perception (Fujisaki, Nakamura, and Imoto, 1973, cited in Toda, 2003). As Figure 28 shows, the length of a consonant /k/ in *kaka* is 1.5 times as long as one in *kaaka*. As the error analysis of the CV.CV token in Figure 23 and Figure 24 showed, the slightly longer duration of /k/ may have confused the learners who perceived as a long consonant (i.e., geminate).

In addition, regarding the vowel type, it was easier to identify and produce vowel duration accurately when tokens contained a low vowel /a/, compared to a high vowel /u/. In Tokyo Japanese, the low vowel /a/ is considered the longest vowel in Tokyo Japanese (Shibatani, 1990) and the high back vowel /u/ is the shortest. Thus, the inherent length of the vowel might have been influential when the learners identified vowel duration. Next, the L2 learners had difficulty correctly producing and perceiveing accurate vowel length in the word-final position (i.e., the second syllable in this study), which supports what Koguma (2000) reported. The word-final position can be a very unstable position perceptually. Mutuskawa (2006) reported that Japanese long vowels in the word-final position (e.g., *konpyuutaa* 'computer') are often shortened (e.g., *konpyuuta*) especially in representing loanwords in Japanese. Finally, the interaction between a pitch pattern and a preceding consonant and/or vowel suggested that perception accuracy was influenced by a combination of the word-level and prosodic level factors.

Regarding the perception latency, the tokens used in this study were also divided into four groups: (I) CVV.CVV, (II) CVV.CV, (III) CV.CVV, and (IV) CV.CV. For the tokens in Group I, it was found that pitch pattern affected response time; the LH.HH pattern had shorter RT than LH.HL. For the tokens in Group II, it was found that pitch pattern influenced RT; LH.H had shorter RT than HL.L. In addition, there was an interaction between a preceding consonant and pitch pattern; with the HL.L pitch, a stop /k/ had shorter RT than a fricative /s/. Regarding the tokens in Group III, vowel type and pitch pattern affected perception RT. A high vowel /a/ revealed shorter RT than a low vowel /u/. Also, the H.LL pitch pattern revealed shorter RT than LH.H and LH.H pitch patterns. An interaction between vowel type and pitch pattern was found, and it suggested that the CV combination /ka/ revealed shorter RT than /ku/ with the LH.L pitch pattern. Finally, for the tokens in Group IV, it was found that preceding consonant affected perception latency; tokens with a fricative /s/ showed shorter RT than tokens with a stop /k/. Also, there was an interaction between vowel type and pitch pattern; a combination of consonant and vowel /su/ revealed shorter RT than /sa/ with the H.L pitch.

Based on these findings, regarding the pitch pattern, there was a tendency for the token ending with the HH pitch to show a shorter RT. In addition, the token with a stop /k/ and/or a low vowel /a/ revealed shorter RT. However, as the interactions between pitch pattern and consonant and/or vowel show, the three factors influenced perception RT together.

Effectiveness of Perceptual Training on Accuracy and RT (RQ 2)

Experiment 2 examined whether focused perception training was effective for the acquisition of vowel duration. In order to test the development of perception accuracy, the accuracy scores before and after training were compared. It was found that the two groups who

received the training, both auditory-visual with waveform input and auditory-only, improved in perception accuracy; the two groups demonstrated higher accuracy in identifying vowel duration after the training. On the other hand, the group that did not receive the training, which served as a control, did not improve their identification accuracy. Thus, it was concluded that the training was effective in enhancing correct perception of vowel length. This finding regarding the benefits of training on accurate perception of L2 contrasts confirmed what Bradlow and Pisoni (1999); Hardison (2003); Hirata and Kelly (2010); Lively, Logan, and Pisoni (1993); Logan, Lively, and Pisoni (1991); Motohashi (2007); Motohashi-Saigo and Hardison (2009) had found.

Regarding the influence of preceding consonant, vowel type, and pitch pattern, the results of the pretest and posttest comparison showed very mixed results. Regarding the tokens in Group I (the tokens with the CVV.CVV structure), among those with the LH.HH pitch, kaa.kaa showed lower accuracy then kuu.kuu and suu.suu; among the tokens with the HL.LL pitch, saa.saa showed lower accuracy than suu.suu. Thus, the learners demonstrated higher accuracy with the tokens with the vowel /u/ than ones with the vowel /a/. This finding did not support the results in Experiment 1 which showed that the vowel /a/, with a potentially longer duration, demonstrated higher accuracy. On the other hand, regarding the tokens in Group II (tokens with the CVV.CV structure), it was found that (1) perception accuracy was higher for those with LH.H pitch than ones with HL.L pitch; (2) the vowel /a/ showed higher accuracy than the vowel /u/ among the tokens with HL.L pitch. In Group III, the results showed that (1) the tokens with /sa/ had a tendency to have higher accuracy than ones with /ka/. Although the data from Group I showed slightly different patterns, generally, the learners demonstrated higher accuracy when they identified vowel duration for tokens with the vowel /a/ than ones with the vowel /u/.

Although perception accuracy showed improvement after perceptual training, both of the training groups showed that perception latency did not decrease. In other words, except for a few examples such as kuu.ku with the HL.L pitch, RTs to identify vowel duration generally became larger. Particularly, the RT of sa.saa with the H.LL pitch significantly lengthened. It was expected that the learners would demonstrate faster RT after the training. It is possible that as a result of receiving the training, the learners who had not been aware of or confident in their knowledge of the distinction noticed the difference and their processing time increased as they considered their response options.

Effectiveness of Training per Group (RQ 2)

The perception accuracy and response latency data obtained in the training sessions for each training group were analyzed in order to examine whether there was a development of perception accuracy and response latency and effects of other factors such as talker, pitch pattern, preceding consonant, and vowel type. For perception accuracy, the AV and A-only training groups demonstrated similar patterns. First, it was found that there were no significant differences in perception accuracy in the first and second week, except for tokens with the CV.CV structure for the AV group and ones with the CVV.CVV structure for the A-only group. Also, there were effects of talker on the perception accuracy. For example, tokens with the CV.CVV structure produced by Talker 5, a female talker, were easier than the other talkers for the AV group. In addition, tokens with the CVV.CV and CVV.CVV structures produced by Talker 4 were more difficult. There was an interaction between talker and stimulus type. For the AV group, tokens such as saa.sa with HL.L pitch as well as kuu.kuu with LH.HH pitch produced by Talker 4, kaa.kaa with LH.HH pitch produced by Talker 3, and suu.suu with LH.HL pitch

produced by Talker 2 were more challenging for the learners. For the A-only group, tokens such as kuu.kuu with LH.HH pitch, suu.suu with LH.HL pitch, saa.sa with HL.L pitch, and sa.saa with L.HL pitch produced by Talker 4 were more challenging for the learners. Finally, it was found that tokens with /sa/ were challenging in general for the learners. Perception accuracy for saa.sa with HL.L pitch as well as sa.saa with LH.H pitch was lower than the other tokens. The reasons why the tokens with /sa/ were more difficult than ones with /ku/ or /su/ could be related to the devoicing in Japanese. In general, the vowels between the two voiceless stops including /s/ and /k/ are devoiced or perceptually lost when they are not accented. By losing the vowel length by devoicing, the contrast between a devoiced short vowel which does not exist perceptually and a long vowel could become clearer, which resulted in higher perception accuracy for tokens with /ku/ and /su/. Also, as the waveform displays in Figure 28 shows, a fricative /s/ has a noise before the vowel, and sonority difference is clearer with a stop /k/ than a fricative /s/ (Hardison & Motohashi-Saigo, 2010).

The perception accuracy in each training session illustrated in Figure 40 suggests the arbitrary nature of time to give a posttest. The study by Logan et al. (1993) and Hardison (2003) administered perceptual training for three weeks. However, the training period in the current study was two weeks. It is not known how posttest results would look if the test had been done after Session 7 or after an additional session. The results show that the learners struggled at least in the first three sessions. Therefore, it is probably difficult to see the facilitative effects of training if the training period is very short.

Regarding the response latency, the AV and A-only groups showed slightly different patterns. First, for the AV group, it was found that response latency significantly shortened in the second week, compared to the first week. In addition, there were effects of talker. For

example, for the tokens with the CVV.CVV structure, suu.suu with LH.HL pitch produced by Talker 1 showed longer RT than kaa.kaa with LH.HH pitch; for the tokens with the CVV.CV structure, the RT for the tokens produced by Talker 4 was faster than those produced by Talker 2 and Talker 3; and for the tokens with the CV.CVV structures, RT for the tokens produced by Talker 1 was longer. Finally, the effects of pitch pattern, preceding consonant, and vowel type showed mixed results; therefore, it was difficult to draw a clear conclusion. However, there was a tendency for tokens with /k/ to have a faster RT than ones with /s/.

On the other hand, for the A-only group, it was found that the response latency was not always significantly shortened in the second week, compared to the first week. For example, the RTs of tokens with the CVV.CVV and CV.CV structures became significantly faster in the second week; however, the same pattern was not found for the CVV.CV and CV.CVV structures. Second, there were effects of talker. For example, RTs for the tokens with the CVV.CVV and CV.CV structure produced by Talker 1 significantly shortened in the second week; and RT for the tokens with CV.CV structures produced by Talker 2 were faster than ones produced by Talker1. In addition, because of the interaction between stimulus type and talker, suu.su produced by Talker 4 revealed shorter RTs while suu.suu produced by Talker 1 and saa.saa produced by Talker 5 revealed longer RTs. Finally, similar to the AV group, the effects of the pitch pattern, preceding consonant, and vowel type showed mixed results; therefore, it was difficult to draw a clear conclusion. However, there was a tendency for tokens with /k/ to have faster RTs than ones with /s/.

The data in the training strongly suggested that talker's voice had effects on the L2 learners' perception accuracy and RT while the study contained only four different talkers in the training sessions. Generally, the L2 learners revealed higher accuracy for the female talkers than

the male talkers. In addition, between the two male talkers (Talker 3 and 4), Talker 4 had lower accuracy. Bradlow, Torretta, and Pisoni (1996) reported six important factors that make a voice intelligible in American English: 1) female, 2) expanded vowel space, 3) precise articulation for the point vowels (i.e., /i/, /a/, /u/), 4) low degree of phonetic reduction, 5) regular rhythm in speech production, and 6) use of a relatively wide range in pitch at the sentence level. This may explain why the two female talkers had relatively higher perception accuracy. Also, as a result of examining Talker 4's voice, it was found that he had lower pitch range than the other male talker so that his voice does not show a wide pitch range.

Comparison between the Two Types of Training (RQ 3)

Although the training was beneficial to improve perception accuracy, the present study did not find significant overall differences in the modality of the training on perception accuracy or perception latency. Regardless of the types of perceptual training the learner took (i.e., AV or A-only), significant improvement occurred. There was only one set of data, tokens with the CVV.CV structure, which showed that the two groups were significantly different. For that set, the AV group had significantly higher accuracy than the A-only group.

Although the overall efficacy of the training type was not found, the interaction between the two points in time (i.e., before and after the training) and the training modality on perception accuracy suggested that the AV training group's rate of improvement was greater than the A-only group's. This finding partially supported Hardison (2003), and Motohashi (2007), and Motohashi-Saigo and Hardison (2009) where perceptual training with bimodal input was more effective than with unimodal input. Visual cues, including articulatory gestures involved in producing /l/ and /r/ as well as a visual display of durational contrasts can explicitly inform

learners about the difference between the two contrasts. On the other hand, the results of the current study showed that the learners were able to be trained to correctly identify vowel duration without the additional information; the focused training with only the auditory input facilitated the correct identification. It is because the waveform displays do not always show a clear distinction between a long and short vowel. As Figure 28 shows, the waveform with a preceding consonant /k/ shows a clear distinction, but not with /s/. Thus, the learners need to pay more attention to the auditory input with less clear visual cues. However, as the training data in Figure 41 show, the AV group revealed higher accuracy for Talker 4. Thus, the AV training could facilitate correct identification of vowel duration for a challenging context such as a difficult voice/talker.

Transfer to Production (RQ4)

Previous literature suggested the effects of perceptual training can transfer to production if the training is successful. This study found that overall production accuracy significantly improved after training for both of the training groups. Since the participants did not receive any specific training or practice on how to pronounce the words with short and long vowels, it was concluded that the effects of the perceptual training transferred to production. While the development of correctly producing vowel duration was observed, there was no effect of training modality or vowel type. Regarding the token type, there were significant differences on production accuracy. The tokens with CVV.CV, CV.CVV, and CV.CV structures significantly improved accuracy from the pretest to posttest; however, the CVV.CVV tokens did not show significant improvement. The CVV.CVV tokens were more difficult than the other types

because error analysis revealed that learners made more errors (i.e., the long vowel on the second syllable was shortened) for this token than the others.

Generalizability of the Training Effect on Perception Accuracy and RT (RQ5)

As Logan et al. (1991) argued, it is necessary to examine whether the effects of the training extend to identification of the L2 contrast in new tokens in order to determine the effectiveness of the training. Therefore, two tests of generalization were conducted: one with novel tokens produced by a familiar voice (TG1) and one with familiar tokens produced by a new voice (TG2). First, perception accuracy was examined. As a result of comparing the pretest data with the two TGs, the overall finding was that the learners demonstrated significantly higher accuracy on the two TGs. Therefore, it was confirmed that there was some development after the perceptual training. The only exception was for the tokens with the CVV.CV structures in the TG1; there were no significant differences in perception accuracy between pretest and TG1. The token se.see with H.LL pitch, which contained a novel vowel, was more difficult than ta.taa with L.HH and H.LL, pitch, which contained a novel consonant. It could suggest that generalization to a new vowel was more difficult than to a new consonant. It is also the case that /t/ and /k/ are both voiceless stops and have shown greater similarity in perception patterns (e.g., Hardison & Motohashi-Saigo, 2010).

Next, as a result of comparing the posttest data with the two TGs, the overall finding was that the learners demonstrated comparable performance. In other words, there was no significant difference between the posttest and the two TGs, except for the tokens with CVV.CVV in TG1 which showed higher accuracy in TG1 compared to posttest. Regarding the stimulus type, the

accuracy scores of see.se and se.see were significantly lower in TG1; therefore, the benefit of the training was not generalized to those two types of tokens containing a novel vowel /e/.

Regarding effects of the training modality on perception accuracy, the AV training was more effective for the development of accuracy for tokens with the CV.CVV structure, compared to the A-only training, in the comparison between the pretest and TG2. However, there were no other meaningful differences between the AV and A-only groups.

Next, the response latency was examined. As a result of comparing the pretest RT with the two TGs, it was found that the learners generally demonstrated significantly shorter response latency on the two TGs although there were some tokens that showed the opposite patterns. Next, as a result of comparing the posttest RT with the two TGs, it was found that the learners generally demonstrated significantly shorter response latency on the two TGs. Based on this finding, the learners were able to respond both accurately and quickly to novel stimuli and a new voice; however, we must also acknowledge that the RTs were significantly longer from the pretest to posttest. In addition, there were no meaningful differences in RTs between the AV and A-only groups. Based on these results, it was concluded that the learners' response time to correctly identify the vowel duration improved and the training effects were extended to the novel tokens as well as the novel voice, regardless of the training type.

Generalizability of the Training Effect to Production

In addition to the generalizability in the learner's perception, it was examined whether the training effects on production accuracy could be generalized to novel tokens. To test it, the test of generalization containing novel tokens was given and compared the data with the pretest and the posttest. As a result of comparing the pretest data with the TG, it was found that the learners

demonstrated significantly higher accuracy on TG. In the comparison between /ka/ and /ta/, where generalization to a novel consonant /t/ was examined, as well as between /sa/ and /se/, where generalization to a novel vowel /e/ was examined, the learners demonstrated higher accuracy in the TG. Therefore, it was concluded that there was a development from pretest to posttest. Also, the types of the tokens were significantly different; the CVV.CV token had the higher accuracy, compared to the CV.CVV tokens. The effects of the training modality were not found.

Next, as a result of comparing the posttest data with the TG, it was found that the learners demonstrated comparable performance. In other words, there was no significant difference between the posttest and the two TGs. The training modality was not significant, but the token type was significant. Based on the comparison of the four token types, it was found that the CVV.CVV type was more significantly difficult than the CVV.CV type. Thus, it was concluded that the learners' ability to correctly identify the vowel duration developed and the training effects were extended to the novel tokens.

Conclusion

In the present study, Experiment 1 explored a range of factors potentially affecting perception and production of vowel duration by L1 English learners of L2 Japanese. Based on the findings, Experiment 2 investigated the factors affecting the efficacy of training to increase learners' identification accuracy of vowel duration. These factors included modality of training (AV vs. A-only), preceding consonant, vowel type, talker's voice, and pitch pattern. Several of these factors had been the focus of some previous training studies.

In the few studies that have explored different modalities of learning, significant improvement was found for both AV and A-only training, with a significant advantage for AV training (Hardison, 2003; Motohashi Saigo & Hardison, 2009). In the current study, although the AV and A-only training groups began at comparable levels, and both showed significant improvement, the greater improvement in raw scores for the AV group compared to the A-only was not statistically significant. Previous research also demonstrated the influence on L2 perceptual identification accuracy of the position of a target sound (e.g., for AE /r/ and /l/) in a word, a talker's voice (e.g., Bradlow et al., 1997; Hardison, 2003; Lively et al., 1993), and an adjacent vowel (e.g., for AE /r/ and /l/, Hardison, 2003; for Japanese geminates, Motohashi Saigo & Hardison, 2009). To this knowledge of contextual influence, the current study adds the significant effects of the prosodic level of speech in the form of pitch pattern, which also encompasses the issue of syllabic position of the morae in a token (i.e., in the first and/or second syllable). Based on the significant complex interactions found in the earlier studies, it is not surprising that the interactions in the current study showed a similar level of complexity in the L2 learners' perceptual performance.

Such perceptual variability is best captured by exemplar-based models of learning in which the learners' stages of L2 perceptual development involve the evaluation of input based on context- and talker-dependent perceptual categories. The influence of context on perceptual identification, now, must be more broadly understood, at least for some target languages, as involving both the segmental and prosodic levels of speech.

In keeping with the hallmarks of successful training established by the past two decades of research (e.g., Hardison, 2012), the current study has also demonstrated the learners' ability to

generalize performance improvement from training to novel stimuli and a new voice, and to transfer an improved perception skill to production in the absence of explicit production training. Among the somewhat unexpected findings of Experiment 2 is the increase in response time for the posttest compared to the pretest stimuli. One might hypothesize that greater accuracy as a result of training would be accompanied by faster response time; however, the reverse finding may have been due to the learners' increased awareness of the range of stimulus cues following training, and their attempts to attend to several dimensions of the speech signal simultaneously. From a pedagogical standpoint, to focus learner attention on specific features of the speech event, teachers may find that visual displays of waveforms (for segmental duration) and pitch contours are helpful in the classroom or, for some learners, as self-study aids outside of class (e.g., Chun, Hardison, & Pennington, 2008; Motohashi Saigo & Hardison, 2009).

There are a few limitations in the current study. The original design called for a consideration of overall L2 proficiency as a factor. Other studies (Hardison & Motohashi Saigo, 2010; Toda, 1998) found an effect of L2 proficiency with regard to geminate perception. Although participants in the current study were recruited from a range of course levels, it was apparent that using exposure to instructed Japanese as a basis for proficiency was unfounded. A comparable number of participants from each year of the course were disqualified from the training study based on ceiling effects in terms of their accuracy in identifying vowel duration. A review of the literature does suggest that, in general, L2 learners of Japanese have less difficulty perceiving vowel duration compared to consonant duration, and the only available, albeit weak, measure of proficiency (i.e., semester of study) was not valid for the research objectives. Second, the current study focused on pseudo words in order to avoid the influence of vocabulary size and neighborhood density. Although this served well the objectives of the

current study and its range of learners, the findings may not be as generalizable to the perception of real words in the natural language environment. Third, the study focused on words produced in isolation. It may be the case that different results would obtain for words produced in context; however, the effect of connected speech on the perception of segmental duration is not clear. For example, while Motohashi Saigo and Hardison (2009) found no significant effect of condition (i.e., isolated word vs. carrier sentence context), a related study found significantly lower identification accuracy for words produced in a carrier sentence versus those produced in isolation (Hardison & Motohashi Saigo, 2010). Finally, to keep the stimulus set to a manageable size in the training study, not every consonant-vowel combination was used for every pitch pattern that can occur in the language. Future research could expand on this aspect.

APPENDICIES

Appendix A: List of target stimuli for production test in Experiment 1

Table 66: Target stimuli in production test

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

Appendix B: List of practice stimuli for production test in Experiment 1 and 2

Table 67: Practice stimuli in production test

Stimuli
noono
nono
rooro
roro

Appendix C: List of target stimuli for perception test in Experiment 1

Table 68: Target stimuli in perception test in Experiment 1

Stimuli	Pitch	Meaning
kaakaa	LH.HH	-----
kaakaa	LH.HL	-----
kaakaa	HL.LL	-----
kaaka	LH.H	-----
kaaka	HL.L	-----
kakaa	L.HH	-----
kakaa	L.HL	-----
kakaa	H.LL	-----
kaka	L.H	-----
kaka	H.L	flowers and fruits
saasaa	LH.HH	-----
saasaa	LH.HL	-----
saasaa	HL.LL	-----
saasa	LH.H	-----
saasa	HL.L	-----
sasaa	L.HH	-----
saasaa	L.HL	-----
saasaa	H.LL	-----
sasa	L.H	sake
sasa	H.L	bamboo leaves
kuukuu	LH.HH	-----
kuukuu	LH.HL	-----
kuukuu	HL.LL	-----
kuuku	LH.H	-----
kuuku	HL.L	-----
kukuu	L.HH	-----
kuuku	L.HL	-----
kuuku	H.LL	-----
kuku	L.H	cane
kuku	H.L	randomness
suusuu	LH.HH	-----
suusuu	LH.HL	-----
suusuu	HL.LL	-----
suusu	LH.H	-----
suusu	HL.L	-----
susuu	L.HH	-----
susuu	L.HL	-----

Table 68 (cont'd)

Stimuli	Pitch	Meaning
susuu	H.LL	-----
susu	L.H	-----
susu	H.L	dust

A dot shown with each pitch pattern represents a syllable boundary.
It is not separating morae.

Appendix D: List of practice stimuli for perception test in Experiment 1 and 2

Table 69: Practice stimuli in perception test

Stimuli	Pitch
noono	LH.H
nono	H.L
rooro	HL.L
roro	L.H

Appendix E: List of target stimuli for perception tests in Experiment 2

Table 70: Target stimuli in perception test in Experiment 2

Stimuli	Pitch
kaakaa	LH.HL
kaaka	HL.L
kaaka	LH.H
kakaa	L.HL
saasaa	LH.HH
saasaa	HL.LL
sasaa	L.HH
sasaa	H.LL
kuukuu	LH.HL
kuukuu	HL.LL
kuuku	LH.H
kuuku	HL.L
kukuu	H.LL
suusuu	LH.HH
suusu	LH.H
suusu	HL.L
susuu	L.HL
susuu	H.LL

Appendix F: List of stimuli for perception training in Experiment 2

Table 71: Stimuli in perception training

Stimuli	Pitch	Meaning
kaakaa	LH.HH	-----
kaakaa	HL.LL	-----
kakaa	H.LL	-----
kakaa	L.HH	-----
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
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	dust

Appendix G: List of practice stimuli for training sessions

Table 72: Practice stimuli in training

Stimuli	Pitch
noono	HL.L
nonoo	L.HH
rooro	LH.H
roro	H.L

Appendix H: List of target stimuli for production test in TG1 in Experiment 2

Table 73: Target stimuli in production test in TG1

Stimuli
seesee
seese
sesee
sese
taataa
taata
tataa
tata

Appendix I: List of target stimuli for perception test in TG1 in Experiment 2

Table 74: Target stimuli in perception test in TG1

Stimuli	Pitch
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

REFERENCES

REFERENCES

Asano, M. (2005). Boundary of sounds. In M. Minami (Ed.), *Linguistics and Japanese Language Education IV* (283 – 294). Tokyo, Japan: Kuroshio Publishers.

Aoyama, K., Flege, J., Guion, S., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, *32*, 233 – 250.

Archibald, J. (2005). Second language phonology as redeployment of L2 phonological knowledge. *Canadian Journal of Linguistics*, *50*, 284 – 315.

Bohn, O.S. (1995). Cros-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in corss-language reserach* (pp. 279 – 304). Timonium, MDL York Press.

Borden, G., Gerber, A., & Milsark, G. (1983). Production and perception of the /r/ - /l/ contrast in Korean adults learning English. *Language Learning*, *33*, 499 – 526.

Bradlow, A. R. & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America*, *106*, 2074 – 2085.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20,* 255-272.

Bradlow, A., Pisoni, D., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*, 2299 – 2310.

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.

Chun, D. M., Hardison, D. M., & Pennington, 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: Benjamins.

Enomoto, K. (1992). Interlanguage phonology: the perceptual development of durational contrasts by English-speaking learners of Japanese. *Edinburgh Working Papers in Applied Linguistics, 3,* 25 – 36.

Flege, J. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229 - 273). Timonium, MD: York Press.

Flege, J., & MacKay, I. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition, 26,* 1 - 34.

Fujisaki, H. & Sugitou, M. (1977). *Onsei no butsuriteki seishitsu* [The physical characteristics of speech]. In Iwanami Kouza Nihongo 5 On'in, pp. 65-105. Tokyo, Iwanami.

Hagiwara, R. E. (1995). Acoustic realization of American /r/ as produced by women and men (Doctoral dissertation, University of California, Los Angeles). *UCLA Working Papers in Phonetics, 90.*

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning, 49,* 213 – 283.

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

Hardison, D. M. (2005b). Variability in bimodal spoken language processing by native and nonnative speakers of English: A closer look at effects of speech style. *Speech Communication, 46,* 73 – 93.

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: 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.

Hayes, B. (1989). Compensatory lengthening in moraic phonology. *Linguistic Inquiry, 20,* 30 – 253.

Hayes, B., Kirchner, R., & Steriade, D. (2004). *Phonetically based phonology*. NY: Cambridge University Press.

Hirata, Y. (1990). Perception of geminated stops in Japanese word and sentence levels by English-speaking learners of Japanese language. *Journal of the Phonetic Society of Japan, 195,* 4 – 10.

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.

Imai, S., Walley, A., & Flege, J. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *Journal of the Acoustical Society of America, 117,* 896 – 907.

Ingram, J. C. L. & Park, S.-G. (1998). Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners. *Journal of the Acoustical Society of America, 103,* 1161 – 1174.

Ingvalson, E.M., McClelland, J.L., & Holt, L.L. (2011). Predicting native English-like performance by native Jaapanese speakers. *Journal of Phonetics, 39,* 571 – 584.

Jamieson, D. E., & Mooroson, D. E. (1986). Training non-native speech contrasts in adults: acquisition of the English /ð/ - /θ/contrast by francophones. *Perception & Psychology, 10,* 83 – 94.

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. (1999a). *The sound system of Japanese*. Tokyo, Japan: Iwanami.

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

Kubozono, H., & Ohta, S. (1998). *Onin koozoo to akusento* [Phonological structures and accent]. Tokyo, Japan: Kenkyuusha.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U. & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277,* 684 – 686.

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.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/. *Journal of the Acoustical Society of America, 89,* 874 – 886.

Metsala, J. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory and Cognition, 25*, 47 – 56.

McCandliss, B. D., Fiez, J. A, Protopapas, A., & Conway, M. (2002). Success and failure in teaching the [r] – [l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience, 2*, 89 – 108.

Minagawa, Y. (1997). 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. *Proceedings of the Spring Meeting of the Society Teaching Japanese as a Foreign Language*, 123 - 128.

Morosan, D. E. & Jamieson, D. G. (1989). Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word-position or task. *Journal of Speech and Hearing Research, 32*, 501 – 511.

Motohashi, M. (2007). Acquisition of geminates consonants in Japanese by American English speakers. Unpublished doctoral dissertation, Michigan State University, Michigan.

Motohashi-Saigo, M., & Hardison, D.M. (2009). Acquisition of L2 Japanese geminates: training with waveform displays. *Language Learning & Technology, 13*, 29 – 47.

Mutsukawa, M. (2006). Japanese loanword phonology in optimality theory: The nature of inputs and the loanword sublexicon. Unpublished doctoral dissertation, Michigan State University, Michigan.

Nagano-Madsen, Y. (1992). *Mora and prosodic coordination: A phonetic study of Japanese, Eskimo and Yoruba*. Lund: Lund University Press.

Ofuka, E. (2003). Perception of a Japanese geminate stop /tt/: the effect of pitch type and acoustic characteristics of preceding/following vowels. *Journal of the Phonetic Society of Japan, 7*, 70 – 76.

Okuno, T. (2009). Factors influencing L2 vowel perception in Japanese: Hyperarticulation, phonetic environment, and talker, American Association for Applied Linguistics Conference, Denver, Colorado, March 2009.

Pennington, M. C. (1996). *Phonology in English language teaching*. New York: Longman.

Port, R.F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America, 81*, 1574 – 1584.

Price, P.J. (1981). A cross-linguistic study of flaps in Japanese and in American English. Unpublished doctoral dissertation, University of Pennsylvania.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of phonetics*, *21*, 427 - 44.

Sheldon, A. (1985). The relationship between production and perception of the /r/ - /l/ contrast in Korean adults learning English: A reply to Borden, Gerber, and Milsark. *Language Learning*, *35*, 107 – 113.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*, 243 – 261.

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

Strange, W., & Dittman, S. (1984). Effects of discrimination training on the perception of /r – l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*, 131 – 145.

Takagi, N. (1993). Perception of American English /r/ and /l/ by adult Japanese learners of English. A unified view. Unpublished Ph.D dissertation, University of California-Irvine.

Toda, T. (1998). Nihongo gakushuusha ni yoru sokuon/chooon/hatsuon no chikakuhanchuuka [Categorical perception of geminates, long vowel, and moral nasals by Japanese learners]. *Bungee Gengo Kenkyuu*, *33*, 65 – 82.

Toda, T. (2003). *Second language speech perception and production: Acquisition of phonological contrasts in Japanese*. Lanham, MD: University Press of America.

Toda, T. (2009). Nihongo kyooiku niokeru gakushuusha onsee no kenkyuu to onsee kyooiku jissenn [Research on learners' speech sounds and practice of speech education in Japanese language education]. *Nihongo Kyooiku*, *142*, 47 – 57.

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

Uther, M., Knoll, M.A., & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, *49*, 2 – 7.

Ziegler, J., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*, 779 – 793.