Non-native speech perception training using vowel subsets: Effects of contrast choice and order of training

Journal:	*Journal of Speech, Language, and Hearing Research*
Manuscript ID:	JSLHR-S-07-0109
Manuscript Type:	Research Article
Date Submitted by the Author:	22-May-2007
Complete List of Authors:	Nishi, Kanae; Indiana University, Speech and Hearing Sciences Kewley-Port, Diane; Indiana University, Speech and Hearing Sciences
Keywords:	English language learners, Speech perception, Bilingualism, Korean

Non-native Speech Perception Training Using Vowel Subsets: Effects of Contrast Choice and Order of Training

Kanae Nishi and Diane Kewley-Port

Indiana University

Abstract

Purpose: Nishi and Kewley-Port (in press) trained Japanese listeners to perceive nine American English monophthongs and showed that a protocol using all nine vowels (Fullset) produced better results than the one using only the three more difficult vowels (Subset). The present study extended the target population to Koreans and examined whether protocols combining the two stimulus sets would provide more effective training.

Method: Three groups of five Korean speakers were trained on American English vowels for nine days using one of the three protocols: Fullset only, first three days on Subset then six days on Fullset, or first six days on Fullset then three days on Subset. Participants' performance was assessed by pre- and post-training tests, as well as by a mid-training test.

Results: 1) Fullset training was also effective for Koreans; 2) no advantage was found for the two combined protocols over the Fullset only protocol, and 3) sustained "non-improvement" was observed for training using one of the combined protocols.

Conclusions: In using subsets for training American English vowels, care should be taken not only in the selection of subset vowels, but also for the training orders of subsets.

Non-native Speech Perception Training Using Vowel Subsets:

Effects of Contrast Choice and Order of Training

In the past two decades, many speech perception training studies were conducted for second language learners (e.g., Strange & Dittmann, 1984; Jamieson & Morosan, 1986; Morosan & Jamieson, 1989; Logan, Lively, & Pisoni, 1991; Lively, Logan, & Pisoni, 1993; Wang, Spence, Jongman, & Sereno, 1999; Iverson, Hazan, & Bannister, 2005; Pruitt, Jenkins, & Strange, 2006; Nishi & Kewley-Port, in press; also see Bradlow, in press for detailed review). The methods and materials used by these studies vary, but overall, their results showed that structured intensive training using stimuli produced by native talkers that include an adequate amount of allophonic variation can improve perception by second language learners. Studies also showed that improvement due to training generalized to untrained voices, tokens, and positions in words. However, the reported results were typically on a small number of contrasts, and except for Nishi & Kewley-Port (in press), vowels were not extensively studied. Our previous study of vowel training with native Japanese listeners used nine American English (AE) monophthongs. Results found that unlike consonant training (i.e., training on /b/-/p/ voicing distinction generalized to /d/-/t/ contrast) (McClaskey, Pisoni, & Carrell, 1983; McReynolds & Bennet, 1972), perception of untrained vowels did not improve after training.

First, let's review the traditional minimal-pair based method often used in clinic and language classrooms in order to consider important issues about the selection of materials. In a typical minimal pair approach, a client (student) is asked to differentiate a pair of words that differ by one phonemic feature (e.g., "*k*ick" and "*p*ick" for initial consonant pair differing in place; "p*u*t" and "p*a*t" for vowel minimal pair; "bi*t*" and "bi*d*" for final consonant pair differing in voicing), a therapist, teacher, or a computer program (e.g., HearSay, see Dalby & Kewley-Port, 1999 for detailed description) provide feedback as to the correctness of the response. In addition, feedback may offer the client some opportunities to listen and compare the pair for differences. The assumption behind this technique is that training focusing on contrasts representing a difficult feature should generalize to the other contrasts that involve the same feature. However, Barlow and Gierut (2002) and Gierut (2004) reported that treatment using pairs, such as /m/-/tʃ/, that represent several feature differences, yielded more generalization than the traditional minimal pairs. This result is partially supported by our previous results (Nishi & Kewley-Port, in press) that a training vowel subset chosen based on difficulty can improve the identification of trained vowels, but does not generalize to untrained vowels. However, it is not obvious how to interpret the idea of multiple feature differences in the context of non-native vowel training. Vowels that present the maximal feature differences are the point vowels /iː-ɑː-uː/, but studies on cross-language vowel perception have shown that perceptual confusions rarely occur for these vowels. This is because these three point vowels or their allophones are commonly found among world's languages (Ladefoged & Maddieson, 1996). Rather, the majority of vowel errors made by non-native listeners were between spectrally adjacent non-native vowels that are assimilated into a single native vowel category (e.g., /iː/-/ɪ/ or /ʌ/-/ɑː/, /æː/-/ɛ/, etc.) (Strange et al., 1998; 2001; 2004; 2005). Apparently, non-native speakers need to learn differences among vowels realized by various gradient combinations of features or new features (e.g., nasality for French vowels).

The research presented in this paper includes two extensions of our previous study (Nishi & Kewley-Port, in press), in which two groups of Japanese listeners were trained on nine monophthongs in AE using two training sets: nine vowels (/iː, ɪ, ɛ, æː, ɑː, ʌ, ɔː, ʊ, uː/) covering the entire vowel space (Fullset protocol) or three out of nine vowels (/ɑː, ʌ, ʊ/) that were more difficult than the other six vowels (Subset protocol). Training results from the Japanese participants indicated that listeners who used the Subset protocol showed rapid improvement to high performance levels on the trained three vowels, however, when tested on the nine vowels, they showed no improvement for the untrained six vowels. In contrast, the listeners trained using the Fullset protocol improved gradually on all nine vowels. These results demonstrated the importance of including a large set in a vowel training protocol. However, the results also suggested the possibility of facilitating learning on the more difficult vowels by combining the Fullset and the Subset protocols in order to provide both problem-focused and large-set training in one protocol (hybrid protocols). For this reason, the present experiment was designed to address a practical issue, namely to evaluate the efficacy of two hybrid protocols (early or later training on the Subset) in comparison to the Fullset only protocol.

In addition, a more theoretical motivation for this experiment was to train a new group of second language (L2) learners (Koreans) so that the efficacy of the Fullset only training could be evaluated with listeners with different first language (L1). Models of second language speech learning (Best, 1995; Flege, 1995) have suggested that non-native phoneme perception is strongly influenced by differences in the phonetic inventories as well as in the realization of allophonic variation for “similar” phonemes in L1 and L2. Therefore, the same training protocol using the same material may not be similarly effective for listeners with different L1s. Korean learners of AE were chosen because Korean vowel system (described below) is substantially different from that of Japanese.

Briefly, AE has 10 monophthongs [iː, ɪ, ɛ, æː, ɑː, ʌ, ə, ɔː, ʊ, uː] and six diphthongs [e^{ɪ}, a^{ɪ}, a^{ʊ}, o^{ʊ}, ɔ^{ɪ}, ^{j}u] (Ladefoged, 1993). Many of these AE vowels are distinguished primarily by spectral properties, and the average duration ratio between inherently long [iː, e^{I}, æː, ɑː, ɔː, o^{ʊ}, uː] and short [ɪ, ɛ, ʌ, ʊ] vowels is approximately 1.3 (Strange et al., 2005).

Korean phonology has undergone many changes during the last 50 years, and the characteristics of modern Korean dialects are not fully described as yet. However, standard South Korean (a dialect spoken in the Seoul vicinity) has been reported to have 10 monophthongs [i, e, ɛ, a, ə(ʌ), u, o, y, ɨ, ø]. Among these monophthongs, the merger of [e-ɛ] distinction is prevalent in the Seoul dialect as well as in many others (Lee & Ramsey, 2000; Sohn, 1999). In addition, the great allophonic variation for [ə] observed for the older generation has been resolved to a vowel similar to AE [ɔː] in the younger generation (Lee & Ramsey, 2000). A recent acoustic study of the Korean vowels showed that the 10 monophthongs are still spectrally distinctive from each other (Yang, 1996). In addition to these 10 vowels, there are two semivowels [w, j] that form 10 onglide vowels [^{w}e, ^{w}ɛ, ^{w}ə, ^{w}a] and [^{j}ɛ, ^{j}e, ^{j}a, ^{j}ə, ^{j}o, ^{j}u] and an offglide vowel [u^{j}]. The two front rounded vowels [y, ø] are often realized as onglides [^{w}i, ^{w}e] in many dialects (Lee & Ramsey, 2000; Sohn, 1999). Vowel duration in Korean used to be phonemic, but majority of the generation born after the World War II uses duration for stress or rhythmic purposes, but not phonemically (Sohn, 1999).

In contrast to AE and Korean, Japanese has only five spectrally distinctive vowel categories. Each spectral category has clear long/short vowel pairs ([i-ii, e-ee, a-aa, o-oo, ɯ-ɯɯ]) where vocalic duration is strictly phonemic (Shibatani, 1990). The average duration ratio between long and short Japanese vowels is reported to range from 2.2 to 3.2, and the spectral differences between the five long/short pairs are very small (Hirata & Tsukada, 2004).

These differences in the vowel inventories among the three languages suggest that not only the number of spectrally distinctive categories but also the use of vocalic duration is clearly different among AE, Korean, and Japanese. Overall, given 10 spectral categories and vowel duration not being phonemic, the Korean vowel system appears to be more similar to that of AE than Japanese. However, as Flege (1995) predicts in his Speech Learning Model, similarity between L1 and L2 might hinder complete learning, although the specific influence cannot be determined without actually examining confusion patterns.

The specific research questions addressed in the present study concern: 1) whether the Fullset training protocol that was found effective for Japanese listeners is also effective for Korean listeners; 2) whether combining the Fullset (9V) and Subset (3V) training protocols (hybrid training) produces more improvement than the Fullset only protocol; and 3) whether the order of training sets influences the outcome of the hybrid training. In order to answer these questions, three training conditions were compared. The first condition was analogous to the Japanese Fullset training (9V-9V). The other two conditions were hybrids in which Fullset and Subset training protocols were combined with the orders of two protocols reversed between the conditions (9V-3V and 3V-9V). Training materials, number of trials and the number of sessions were maintained the same across the three conditions. The 9V-9V protocol was expected to improve Korean listeners' AE vowel perception to a similar level as Japanese listeners even though the vowel systems for Korean and Japanese were substantially different. The two hybrid training groups were also expected to show improved performance after training. In addition, the overall post-test performance for listeners trained using hybrid protocols was expected to be higher than the 9V-9V group's due to the focused training on the more difficult three vowels for three days. However, no prediction was made regarding the difference between the two hybrid conditions.

Method

Korean Participants

There were 15 Korean participants (10F, 5M; mean age = 23 years, range 19-30 years). All were native speakers of Korean who had never lived outside South Korea for more than one year. Participants were either students in the Intensive English Program, music school, or business school at Indiana University. One participant had already graduated from the Intensive English Program, and had been in the U.S. for 11 months when she completed training, but the data of this participant was not excluded because her response patterns at pre-test did not show remarkable difference from those of the other participants. All other participants were in the U.S. less than four months. All except one considered that their main dialect region was Seoul. The one participant was from the Kyengsang dialect region, where vowel duration is used phonemically (Lee & Ramsey, 2000; Sohn, 1999), but the data of this participant was not excluded because her response patterns at pre-test were similar to those of the other participants.

Participants were divided randomly into three groups of five. All participants were trained on AE vowels nine days. First group was trained using only the Fullset protocol for nine days (9V-9V); the second group was trained using the Fullset protocol for the first six days, then the Subset protocol for the last three days (9V-3V); the last group was trained using the Subset protocol for the first three days, then the Fullset protocol for the last six days (3V-9V). Figure 1 presents the schedule comparison between the groups. All participants were given pre-test on the first day, and post-test on the last day. A short mid-test probe (split into two halves, A and B) were given to assess improvement due to the specific protocol. Thus, the hybrid groups had the probe when training sets were changed: the seventh and the eighth days for 9V-3V group, and the fourth and the fifth days for the 3V-9V group. The probe schedule for the 9V-9V group was the same as the 9V-3V group to ensure that difference observed at post-test, if any, can be attributed to the Subset protocol. One male participated in the 9V-9V group, two in the 9V-3V and the 3V-9V groups.

Stimulus materials

The stimulus materials were the same as the previous study (Nishi & Kewley-Port, in press). For details of the recording apparatus, recording procedures, and stimulus preparation refer to Nishi and Kewley-Port (in press).

Briefly, there were two categories of stimulus materials: 36 monosyllabic consonant-vowel-consonant (C_1VC_2) real words (RW) and 54 disyllabic nonsense words (NSW). The NSW were /C_1VC_2ə/ where C_1-C_2 combinations were /b-b, b-p, d-d, d-t, g-g, g-k/. The consonants in the RW stimuli were /b, p, d, t, k, h, s, z, ʃ, tʃ, dʒ, m, n, l, w/ for C_1, and /b, p, d, t, g, k, s, z, ʃ, m, n/ for C_2. The NSW were used both in training and tests, but the RW were used only in the test to examine generalization in more varied consonantal contexts. Each of the stimulus words included one of the nine AE monophthongs /iː, ɪ, ɛ, æː, ɑː, ʌ, ɔː, ʊ, uː/. All stimulus words were produced in a carrier sentence, “The first word is ___, isn’t it?” with a falling intonation before the tag question. All carrier sentences were digitized at a sampling rate of 24.414 kHz, and the stimulus words were excised from the sentences.

Five native speakers of AE (2 female (F1, F2), 3 male (M1, M2, M3), age = 20-27 years old) from the North Midland dialect region (Labov, Ash, & Boberg, 2006) recorded two tokens for each NSW and RW. Tokens by speaker M1 were used only for task familiarization. Speakers F2 and M3 served as speakers for both test and training stimuli (Trained speakers); speakers F1 and M2 served as speakers for test stimuli (New speakers). The selection procedure of speakers was reported previously (Nishi & Kewley-Port, in press).

Stimulus materials for the tests, probe, and training are described below.

Tests. All stimuli (NSW and RW) produced by both Trained and New speakers were presented in both pre- and post-tests.

Probe. In the probe, NSW tokens that contain nine vowels produced in three (/b-b, d-d, g-g/) out of the six consonantal contexts were presented. The probe had two parts (A and B, respectively. See Figure 1), and were presented when stimulus sets were switched. Probe-A contained the stimuli produced by the New speakers, and probe-B contained the stimuli produced by the Trained speakers. All participants took probe-A first, but the orders of two speakers in each probe were counterbalanced among participants.

Training. Training presented only the NSW tokens. There were two sets of training stimuli that were identical to our previous study with Japanese participants (Nishi & Kewley-Port, in press). The first set included all nine vowels and was used for the Fullset protocol. The other set included only the three more difficult vowels, /ɑː, ʌ, ʊ/ and was used for the Subset protocol. Performance observed during a pilot study with Korean participants that used the Fullset protocol confirmed that these three vowels were always more difficult than the other six vowels.

Apparatus

Tests and probe. Stimulus files were low-pass filtered at 5 kHz using a built-in10th-order Butterworth filter on the TDT PF1. All stimuli were presented only to participant’s right ear through earphones (TDH-39A) at a fixed listening level determined as comfortable during pilot testing.

Training. Seven of the participants (two 3V-9V, two 9V-3V, three 9V-9V) were trained using different apparatus from the tests and probe. For these participants, stimuli were presented only to the right ear through Sennheiser HD250 linear II headphones. Stimulus files were low-pass filtered at 5 kHz using an IIR filter approximated the 10th-order Butterworth filter used for the tests and probe. The listening level was made equivalent to that for tests and probes. The other eight participants were trained using the same apparatus and settings as tests and probes.

Procedures

Familiarization for response system. Prior to pre-test, all participants were familiarized with the responses used in all training sessions and tests. The detailed procedure was the same as the previous study (Nishi & Kewley-Port, in press), except that the instructions were written in Korean. All participants passed a written confirmation test in which they were required to identify an International Phonetic Alphabet (IPA) symbol representing the vowel sound in a keyword correctly more than 90% of time (9 vowels × 2 key words × 5 randomizations = 90 total questions). Two participants (not among the 15 trained here) did not pass this IPA-keyword familiarization and were not included in the study.

Tests. All participants were tested on NSW tokens first. Stimulus presentation was blocked by speaker. Presentation order of speakers was counterbalanced among the participants. Stimuli within a block were randomized. The participants' task was to identify the first AE vowel in the /C_1VC_2ə/ NSW by choosing one of the nine IPA vowel symbols displayed on the computer screen. The 18 keywords used during the response system familiarization always appeared under each IPA symbol. The stimulus word was presented only once. In total, there were 108 trials (6 consonantal contexts × 9 vowels × 2 tokens) per block for NSW, 72 trials (4 words × 9 vowels × 2 tokens) per block for RW. Participants were allowed to take breaks between blocks. All participants completed tests in approximately 90 minutes.

Training. Training presented only the NSW tokens. The participants had nine training sessions between pre- and post-tests. One session lasted approximately 90 minutes and consisted of six blocks of 108 trials (Fullset: 9 vowels × 6 contexts × 2 tokens; Subset: 3 vowels × 6 contexts × 2 tokens × 3 repetitions). Among the six blocks, participants alternated the blocks by the two speakers, F2 and M3, three times. Half of the participants began training with speaker F2 and the other half began training with M3.

The procedures for the training were similar to the vowel identification task given in the tests, except that feedback (adopted and modified from Miller et al., 2004) was provided for each trial. In a particular trial, when a participant identified a target vowel correctly, the text feedback “Correct” appeared on the computer screen and the next trial began; when the answer was wrong, the text feedback “Incorrect” and a subwindow with three buttons (Correct, Wrong, and Stop) appeared on the screen. The participant was encouraged to listen to the sound of the correct answer and the wrong answer by clicking on the Correct or Wrong buttons up to ten times in any combination. The listener was allowed to use the Stop button to proceed to the next trial at any time during the feedback, or not to use the listening opportunity at all. The sound of the correct answer was always the stimulus presented for the trial, but the sound of the wrong answer was randomly chosen by the computer program from the two tokens produced by the same speaker for each button press. Participants completed all sessions including pre- and post-tests within one month.

Probe. Two halves of the probe (A and B, see Figure 1) were given when the hybrid groups switched training stimulus sets. Specifically, the 3V-9V group had probe-A after the completion of Subset training on the third training session and probe-B before starting the Fullset training on the fourth training session. The 9V-3V group had probe-A after the sixth training session and probe-B before the seventh training session. The probe schedule for 9V-9V group was the same as the 9V-3V group. The participants' task was the same as the tests.

Results

Pre-training Confusions

Initial vowel confusions were summarized by aggregating all responses collected for the 15 Korean participants at pre-test. Table 1 presents the summary. The first column lists the stimulus vowels. Column 2 lists most frequent response for a stimulus vowel along with the percentage of frequency out of 720 total opportunities (column 3). Columns 4 and 5 present the same results as columns 2 and 3 for the second modal responses, respectively. Columns 6 and 7 are for the other responses that occurred more than 5%, if any. The last column shows the number of times the target vowel was used as incorrect response for other vowels.

The average percent correct identification across all participants for all talkers was 48%. Two (/ʌ, ʊ/) of the three vowels (/ɑː, ʌ, ʊ/) chosen for the Subset training were among the least accurately identified vowels (39% and 31%, respectively). Although the low vowel /ɑː/ (72%) was by far the most accurately identified among the nine vowels, when the incorrect responses for all target vowels were considered (column 8), apparently Korean participants had a strong bias towards choosing this vowel as response (c.f. 68 times for /i/). This, in turn, inflated correct identification of /ɑː/ target. A further examination revealed that the most frequent incorrect use of /ɑː/ was for the target vowel /ʌ/ (248 responses out of 639). On the other hand, the response vowel /ʌ/ was frequently used for the target /ʊ/ (150 responses). Overall, pre-test confusions indicated that the Subset /ɑː, ʌ, ʊ/ was indeed confusable for Korean listeners.

Time Course of Training

The time course of nine-day training (Tr-1 to Tr-9) was summarized as daily average scores for the trained tokens. The four panels in Figure 2 present the results for the three Korean groups (n = 5) and for the Japanese Fullset group (upper right, n = 6) from the previous study (Nishi & Kewley-Port, in press). Each line in Figure 2 shows a participant's performance over the nine training sessions calculated as an average across the trained vowels and six daily training blocks. In the figure, along with the scores during training sessions, averages calculated over the nine vowels for pre-, post-test, and probe (only for Korean) are plotted.

As can be seen, all participants showed improvement over the nine-day training protocol. The performance by the 9V-9V group (upper left), where the greatest positive change in performance was observed from the pre-test to session Tr-1, followed by a gradual improvement over the next eight sessions. This pattern was similar for all four groups including the Japanese Fullset group. Generally, improvement during the Fullset training (Tr-1 to Tr-9 for 9V-9V; Tr-1 to Tr-6 for 9V-3V; and Tr-4 to Tr-9 for 3V-9V) was gradual and resembled that of the Japanese Fullset group. For the hybrid group, performance during the Subset training (Tr-7 to Tr-9 for 9V-3V; Tr-1 to Tr-3 for 3V-9V) was with much higher accuracy. Interestingly, for the 3V-9V group, even though they showed rapid and large improvement during the Subset training, all participants' performance suddenly declined when the other six vowels were introduced in the probe and Fullset training in session Tr-4. This suggests that the high accuracy observed during the Subset training may be strongly related to lowered uncertainty or simple procedural learning.

Performance Change for Trained Stimuli

This section reports the changes in participants' performance between pre- and post-tests. All the results presented in the figures are in percent correct identification scores, but these scores were converted into rationalized arcsine units (rau, Studebaker, 1985) for statistical analyses in order to allow score distributions at the extremes (under 15% and over 85%) to be more linear. In the analyses described below, participant group was treated as between-subject variable; test, speaker, and vowel were treated as within-subjects variables.

Figure 3 presents the summary of pre- and post-test average scores on the NSW stimuli produced by the Trained speakers (top panel) and New speakers (lower panel). The average pre- and post-test percent correct scores calculated across nine vowels and speakers are presented for the three Korean groups (3V-9V, 9V-3V, and 9V-9V) as well as Japanese Fullset group (J_Fullset). Error bars indicate the within-group standard deviations.

First, the three Korean groups' pre-test average vowel scores for Trained speakers were subjected to a between-subject ANOVA to find out whether the three groups were similar before training. Despite some differences among groups noted by visual inspection, no statistical difference was found among the groups ($F(2, 12) = 1.26$, $p = .32$). This result, therefore, warrants that other differences found among the three Korean groups at post-test can be attributed to training.

Next, pre- and post-test scores were subjected to a mixed design ANOVA (group × test). The results indicated that the main effect of test was significant ($F(1, 12) = 154.96$, $p < .001$), but main effect of group ($F(2, 12) = .19$, $p = .83$) and group × test interaction ($F(2, 12) = 1.65$, $p = .23$) were not. These results indicate that the three training protocols were overall equally effective to improve AE vowel perception by Korean participants.

The last statistical analysis on the Trained speaker scores was performed to confirm that the Fullset training was similarly effective for participants with different L1 backgrounds. The results of a mixed design ANOVA (group × test) comparing Korean 9V-9V group and Japanese Fullset group revealed that the both groups showed significant improvement after training ($F(1, 9) = 68.40$, $p < .001$), but no significant difference between the two language groups ($F(1, 9) = .27$, $p = .61$) or group × test interaction was observed ($F(1, 9) = 3.58$, $p = .09$). Therefore, overall, a training protocol incorporating a large set of vowels was effective for participants with different L1 backgrounds.

Generalization of Training Effects

Next, generalization of training was examined for tokens produced by New speakers and RW tokens that were not presented during training.

New speakers. As was described earlier, NSW tokens produced by the New speakers were presented only in the tests so that generalization to new voices can be observed. The lower panel of Figure 3 presents the result for all four groups. The results for the four groups were not only very similar to one another but also closely resembled the results for the Trained speakers (upper panel). In order to confirm this observation, a mixed-design ANOVA (group × test) was performed for the Korean groups. Significant main effect was observed only for test ($F(1, 12) = 152.65$, $p < .001$) but not for group ($F(2, 12) = .12$, $p = .89$) or group × test interaction ($F(2, 12) = 1.39$, $p < .29$). These results indicate that all three groups could generalize improved vowel identification to the NSW tokens spoken by New speakers.

Real words. The next analysis was performed for the RW tokens. The mean percent correct scores for Trained and New talkers for the three training groups at pre- and post-tests are presented in Table 2. In the table, the first column lists participant groups, columns 2 and 3 present the average pre-test scores for the Trained and New talkers, respectively. Columns 4 and 5 present the post-test scores.

In the table, changes in performance after training were smaller for RW than for NSW tokens, but similar overall improvement for the three groups was observed. Indeed, a test × group × talker ANOVA on the RW data indicated only a significant main effect for test ($F(1, 12) = 73.49$, $p < .001$), but none of the other main effects or interactions were significant. These results are similar to those of the previous study with Japanese participants (Nishi & Kewley-Port, in press), and suggest that the training on NSW also improved identification of vowels in the untrained RW stimuli. Note that vowels in these RW were presented in more varied consonantal contexts than the NSW used for training. Therefore, these results also indicate that a training method using highly variable naturally produced stimulus materials can improve L2 listeners' identification of non-native vowel categories even in untrained consonantal contexts.

Effects of Order of Training Sets

As can be seen in Figure 2, even though no difference was found for overall improvement among the three training groups, when training sets were switched, identification accuracy of the two hybrid training groups suddenly declined. This implies the possible influence of the order of training sets. For this reason, the same data presented in Figure 2 were summarized for each vowel and presented as Figure 4. The three panels are for the three Korean training groups, 9V-9V, 9V-3V, and 3V-9V, respectively. Individual lines in each panel show per-session correct identification for a vowel averaged across five participants in a group, six training blocks, and six consonantal contexts. The three Subset vowels are indicated by solid lines. Similar to Figure 2, scores for pre-, post-tests and probe are also shown.

Obviously, the 9V-9V and 9V-3V groups' performance during the Fullset training are very similar. For these groups, scores for the Subset vowels tended to be lower than the other vowels during the Fullset training. The 3V-9Vgroup showed remarkably high accuracy of identifying vowel /ʊ/ even at Tr-1 (notice that other two groups scored about 60% on that vowel at Tr-1). However, when other six vowels were introduced in Tr-4 for Fullset training, performance for /ʊ/ suddenly declined and remained around 60% all through the Fullset training. Interestingly, the amount of decline for the other two vowels in the Subset, /ɑː/ and /ʌ/, was not as large as /ʊ/.

The above observations was confirmed by an ANOVA (group × training session × vowel) performed for average scores for Subset vowels during the six Fullset training sessions. Results indicated significant main effects of training session ($F(5, 60) = 20.31$, $p < .001$) and vowel ($F(2, 24) = 14.99$, $p < .001$), as well as significant interactions of vowel × group ($F(4, 24) = 4.08$, $p < .02$) and vowel × session ($F(10, 120) = 2.13$, $p < .03$). None of the other effects were significant. A further examination was performed on the vowel × group interaction using Tukey HSD. Results indicated that only 3V-9V group showed significant difference between vowel /ʊ/ and other two Subset vowels, /ɑː/ and /ʌ/ ($p < .001$ for /ʊ/-/ʌ/ comparison, $p < .002$ for /ʊ/-/ɑː/ comparison).

These results indicate that the rapid improvement on /ʊ/ observed for the 3V-9V group during the Subset training may not be due to perceptual learning of appropriate acoustic cues for AE vowels. Rather, participants might have learned the strategy to maximize their identification accuracy by relying on acoustic cues available when uncertainty was low during the Subset training. However, as the results suggests, they perpetuated this strategy even though it was not effective during the Fullset training.

Use of Feedback and Performance on Subset

As shown in Figure 4, the correct identification for the target vowels /ɑː/ and /ʌ/ increased steadily for the three groups, while only 3V-9V group did not improve on /ʊ/ from session Tr-1 to Tr-9. We interpret these results as due to the particular training set order. To investigate whether the training set order influenced how participants engaged in training, the three groups' use of feedback during training was analyzed. Recall that the feedback provided in the present study was interactive – when a response was incorrect, a participant was afforded 10 opportunities to listen to the sound of correct and incorrect answers in any combination, with an option to terminate feedback and continue on to the next trial. The training program recorded the number of button presses during feedback for correct sound, wrong sound or to stop feedback. Figure 5 presents the frequency of errors (left panels) and the number of erred trials for which participants used feedback listening at least once (right panels) for the three groups. The horizontal axis indicates training sessions (Tr-1 to Tr-9) and the vertical axis shows the frequency of occurrence summed across six blocks and five participants. The solid vertical lines in the panels for 9V-3V and 3V-9V groups indicate when these hybrid training groups switched protocols. Results are shown only for individual vowels used in Subset training.

Comparison between the frequencies of errors and feedback use suggests that 9V-9V group used feedback listening for almost all errors throughout all training sessions. A similar trend can be found for 9V-3V group during the Fullset training (Tr-1 to Tr-6). As for the 3V-9V group, their feedback use during the Fullset training (Tr-4 to Tr-9) for vowel /ɑː/ and /ʌ/ revealed rather similar patterns as the other two groups. For vowel /ʊ/, however, as suspected from the percent correct vowel scores presented in Figure 4, their feedback use was only one half or less of the error frequencies.

Although descriptive, these results illustrate that the participants in the 3V-9Vgroup did not use feedback listening as often as the other two groups during Fullset training, and this tendency was the most prominent for the vowel /ʊ/ that showed high error frequency and little improvement during the Fullset training. This is in stark contrast to the 3V_9V group with near-perfect performance on /ʊ/ for the final three training sessions. Due to the post-hoc nature of the analysis, clear causality cannot be drawn from these results, but considering the difference found among the three groups was only for vowel /ʊ/, and 3V-9V is the only group that was trained on the Subset vowels first, it is reasonable to conclude that the order of training set and possibly the vowels included in the subset influenced the participants' learning strategies.

Discussion

Summary of Results

As an extension of our previous vowel training study with Japanese learners of American English (AE) (Nishi & Kewley-Port, in press), the present study trained native speakers of Korean for nine sessions to perceive nine AE monophthongs using three different training protocols. The first protocol 9V-9V was a control condition in the present study and used nine vowels for nine training sessions; the second protocol 9V-3V used nine vowels for the first six sessions, then the training set was reduced to three more difficult vowels for the last three sessions; the third protocol 3V-9V presented the three more difficult vowels during the first three sessions, then the set size was expanded to nine vowels for the last six sessions. The 9V-3V and 3V-9V protocols were designed to test whether focused training on three more difficult vowels facilitates higher overall performance than the 9V-9V protocol.

Results showed that all groups perceived AE vowels more correctly after training. Participants trained using 9V-9V protocol made comparable improvement to the Japanese equivalent group in our previous study (Nishi & Kewley-Port, in press), indicating that a vowel perception training protocol using a large set of highly variable, naturally produced stimuli is effective for learners from first languages with different vowel inventories. However, although it was hoped that the hybrid protocols, 9V-3V and 3V-9V, would be found more effective than the 9V-9V protocol due to their "problem-focused" training component, results showed that neither of them were more effective than 9V-9V. Instead, a closer look at the changes in identification accuracy over the nine-day training sessions, as well as between pre- and post-tests, revealed that the 3V-9V hybrid group failed to improve on one of the vowels among the three difficult vowels, /ʊ/, probably because of the particular order of the training stimulus sets.

Why the 3V-9V Protocol Failed to Train /ʊ/?

The most intriguing result here was that training first on the smaller set had a strong negative influence on the learning of the vowel /ʊ/ for the 3V-9V group but not for 9V-3V group. Because the obvious difference between these groups was the training set orders, it is reasonable to conclude that the order of training was the major cause. However, considering that the same was not true for the other vowels (/ɑː/ and /ʌ/) included in the smaller set, the specific training set might have contributed to this difference.

What mechanism might we posit to establish that the order of training and the specific training set caused the above difference? One plausible explanation is that the Subset and Fullset protocols might have been different in terms of task quality. Even though the same identification tasks and feedback systems were used in both protocols, the cognitive mechanisms involved during training might be different. Juslin, Olsson and Olsson (2003) explored the difference between categorization and multiple-cue judgment tasks, and found that people shift strategies from simple memorization of exemplars to abstraction of acoustic cues when the criterion changed from a binary to a continuous, probabilistic variable. Similarly, Smith and Minda (2000) pointed out that a training condition that repeatedly presents a small set of stimuli can invite trainees to simply remember the association between a category label and each exemplar without deeper processing to learn the characteristics of a category. After such learning, the task simply becomes a memory retrieval process. On the other hand, if trainees process stimuli deeper and learn particular characteristics for categories, then the decision making is based on probability calculation in reference to the learned category characteristics. Applying this explanation, the Subset protocol represents a categorization task in which the number of categories presented is limited to three. On the other hand, the Fullset protocol approximates a continuum of categories (i.e., multiple-cue judgment task) by presenting three times more categories than the Subset protocol. Intentionally or unintentionally, trainees may have discovered the more effective strategy by weighing which requires less effort – i.e., number of label-exemplar associations or number of cues to be evaluated to determine vowel categories. Obviously, memorizing exemplars for the Subset training is efficient, but not for the Fullset training.

Then why 9V-3V did not show the same results as 3V-9V? The following may be a plausible explanation. The 3V-9V protocol initially guided learners to make label-exemplar associations for the three vowels (/ɑː, ʌ, ʊ/). Then when task demands expanded to nine vowels, trainees had not only to use the newly learned labels, but also to modify a still unstable, complicated system again to accommodate more new vowels. It is as if learning two sets of foreign vowels successively. On the other hand, the 9V-3V protocol began with guiding trainees to establish a multiple cue weighting strategy for many vowels, and the subsequent reduction of categories in Subset training does not require any further modification. If this explanation is reasonably correct, then training first on a smaller set – even though it includes more difficult vowels produced in the same number of consonantal contexts as in the larger set – may not produce optimal learning.

What We Need to Know about Perceptual Learning

As reviewed in the introduction, the focus of research on cross-language speech perception training to date has been typically on the three areas: 1) the plausibility of modifying adult speech perception; 2) the influence of previous linguistic experience; and 3) the influence of experimental variables (stimuli, feedback, and task) on the training outcome. Views are converging that adult speech perception can be modified through rather short-term but structured laboratory training regardless of previous linguistic experience. Although it has never been compared directly, a training protocol using an identification task with highly variable, naturally produced training stimuli has been assumed to produce more successful outcomes than discrimination task using synthetic stimuli (see Bradlow, in press for summary). The present results can suggest at least three topics for future research.

First, the results of the present study showed that non-native vowel training requires additional care in protocol design, namely the order of training sets and the number of vowels included in a set. Because no previous training study has used a large set of consonants, it cannot be determined whether the present results are true for vowels only. Thus further research on consonant training is suggested.

Second, while the present results indicated that the use of large training sets may prevent certain learning problems, it is not known how large a set should be to guide trainees to use multiple-cue judgment rather than exemplar-based judgment.

Lastly, the results of feedback use showed that learners may lose interest or "give up" on making efforts on one or more of the items. Future studies on speech perception training should investigate how trainees' engage themselves in learning activities during training, and how feedback should be structured within a training protocol.

Conclusions

The results of the present study add evidence that adult non-native vowel perception can be improved through structured laboratory training. In addition, the present results suggested possible negative influence of using smaller training sets. In conclusion, we suggest that human language learning requires an adequate amount of complexity that enables one to attend to the cues that differentiate categories rather than memorizing the association between a category and exemplars. Furthermore, results caution against the casual use of a "subset then larger set" training protocol, or presumably the use of several subsets, because such protocols may cause unexpected, perhaps long lasting, unfavorable influence on learning strategies.

References

Barlow, J. A., & Gierut, J. A. (2002). Minimal pair approaches to phonological remediation. *Seminars in Speech and Language, 23,* 57-67.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171-204). Baltimore: York Press.

Bradlow, A. R. (in press). Training non-native language sound patterns: Lessons from training Japanese adults on the English /r/-/l/ contrast. In J. Hansen & M. Zampini (Eds.), *State-of-the-art issues in second language phonology.*

Dalby, J., & Kewley-Port, D. (1999). Explicit pronunciation training using automatic speech recognition. In M. Holland (Ed.), *Tutors that Listen: Speech Recognition for Language Training, Special Issue of the Journal of the Computer Assisted Language Learning Consortium (CALICO), 16*, 425-445.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-277). Baltimore: York Press.

Gierut, J. A. (2004). Enhancement of learning for children with phonological disorders. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 50+ years of discoveries in speech communication* (pp. 164-172). Boston: MIT.

Hirata, Y., & Tsukada, K. (2004). The effects of speaking rates and vowel length on formant movements in Japanese. In A. Agwuele, W. Warren, & S.H. Park (Eds.), *Proceedings of the 2003 Texas Linguistics Society Conference* (pp. 73-85). Somerville, MA: Cascadilla Proceedings Project.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America, 118,* 3267-3278.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception and Psychophysics, 40,* 205-215.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132,* 133-156.

Labov, W., Ash, S., & Boberg, C. (2006). *Atlas of North American English: Phonology and Phonetics.* Berlin: Mouton de Gruyter.

Ladefoged, P. (1993). *A course in Phonetics.* Fort Worth: Harcourt Brace.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages.* Malden, MA: Blackwell.

Lee, I., & Ramsey, S. R. (2000). *The Korean language.* New York: State University of New York Press.

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. *The 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/: A first report. *The Journal of the Acoustical Society of America, 89,* 874-886.

McClaskey, C. L., Pisoni, D. B., & Carrell, T. D. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception and Psychophysics, 34,* 323-330.

McReynolds, L. V., & Bennet, S. (1972). Distinctive feature generalization in articulation training. *Journal of Speech and Hearing Disorders, 37,* 462-470.

Miller, J. D., Dalby, J. M., Watson, C. S., & Burleson, D. F. (2004). Training experienced hearing-aid users to identify syllable-initial consonants in quiet and noise. *The Journal of the Acoustical Society of America, 115,* 2387.

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.

Nishi, K., & Kewley-Port, D. (in press). Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets.

Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America, 119,* 1684-1696.

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

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 3-27.

Sohn, H.-M. (1999). *The Korean language.* New York: Cambridge University Press.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., & Nishi, K. (2001). Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *The Journal of the Acoustical Society of America, 109,* 1691-1704.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., Nishi, K. & Jenkins, J. . (1998). Perceptual assimilation of American English vowels by Japanese listeners. *Journal of Phonetics, 26,* 311-344.

Strange, W., Bohn, O.-S., Nishi, K., & Trent, S. A. (2005). Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *The Journal of the Acoustical Society of America, 118,* 1751-1762.

Strange, W., Bohn, O.-S., Trent, S. A., & Nishi, K. (2004). Acoustic and perceptual similarity of North German and American English vowels. *The Journal of the Acoustical Society of America, 115,* 1791-1807.

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

Studebaker, G. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research, 28,* 455-462.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America, 106,* 3649-3658.

Yang, B. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics, 24,* 245-261.

Author Note

Kanae Nishi and Diane Kewley-Port, Department of Speech and Hearing Sciences, Indiana University.

A portion of this paper was presented at the 150th Meeting of the Acoustical Society of America in Minneapolis, Minnesota, 2005. This research was supported by the NIH postdoctoral fellowship to the first author (F32 DC-006313) and a NIH research grant to the second author (R01 DC-02229). Manuscript preparation was in part supported by a NIH research grant to Patricia G. Stelmachowicz (R01 DC-04300). William B. Mills wrote the training program used here. James J. Jenkins and Winifred Strange made significant contribution to the development of the initial research design. The comments and consultation with Charles S. Watson and James D. Miller regarding the interactive feedback system were invaluable. The authors also wish to thank Jaehee Lee for translating consent forms and instructions into Korean. Finally, the authors wish to thank all the participants.

Correspondence should be addressed to Kanae Nishi, Boys Town National Research Hospital, 555 N. 30th street, Omaha, NE 68131. E-mail: nishik@boystown.org.

Figure Captions

Figure 1. Summary of participation schedule for the three training groups: 9V-9V, 9V-3V, and 3V-9V. “Fullset” and “Subset” denote the vowel sets used for training. “A” and “B” indicate the two halves of the probe.

Figure 2. Percent correct identification scores on the trained tokens for the individual participants in the three Korean training groups (9V-9V, 9V-3V, 3V-9V) at pre- and post-tests, probe, and during training (Tr-1 to Tr-9). Equivalent data for the Japanese Fullset group (upper right) from Nishi and Kewley-Port (in press) is also presented for comparison with the Korean 9V-9V group (upper left).

Figure 3. Percent correct identification scores obtained for the nonsense word tokens spoken by the Trained speakers (top panel) and the New speakers (bottom panel) observed at pre- and post-tests for the three Korean groups (3V-9V, 9V-3V, 9V-9V) and Japanese Fullset group (J-Fullset).

Figure 4. Percent correct identification scores for the individual vowels by the three Korean groups (3V-9V, 9V-3V, 9V-9V) at pre- and post-tests, probe, and during training (Tr-1 to Tr9). Thicker lines indicate the three vowels used in the Subset training.

Figure 5. The frequency of errors in each session for the three vowels used in the Subset training (left panels), and the frequency of feedback use at least once per erred trial for the three vowels in the Subset training (right panels).

Table 1

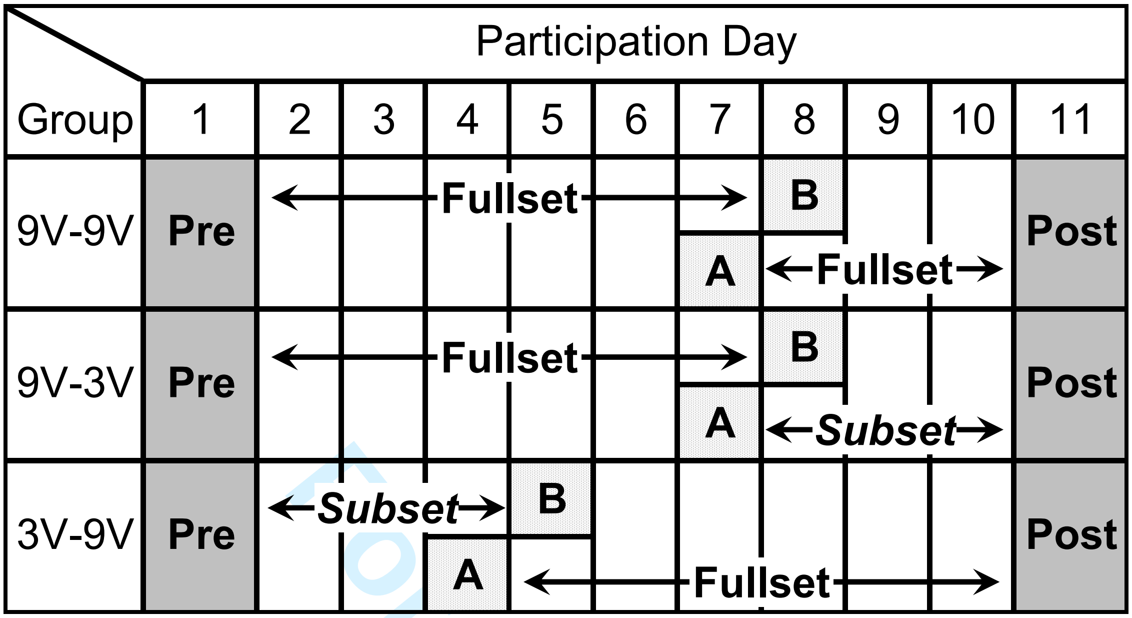
Vowel Confusion Patterns in Percentage at Pretest for All Korean Participants (N=15). Overall percent correct identification = 48%. Total number of observations per stimulus vowel is 720.

Stimulus Vowel	**Modal**		**Second Modal**		**Other**		**Incorrect Use Count**
	Response	**%**	**Response**	**%**	**Response**	**%**	
iː	iː	58	ɪ	42			68
ɪ	ɪ	53	ɛ	24	æː	13	309
					iː	8	
ɛ	ɛ	41	æː	40	ɑː	11	426
æː	æː	55	ɛ	30	ɑː	11	435
ɑː	ɑː	72	ɔː	17	ʌ	5	639
ʌ	ʌ	39	ɑː	34	ɔː	18	440
ɔː	ɔː	46	ɑː	29	ʌ	23	379
ʊ	uː	32	ʊ	31	ʌ	21	398
					ɔː	10	
uː	ʊ	51	uː	41	ʌ	7	249

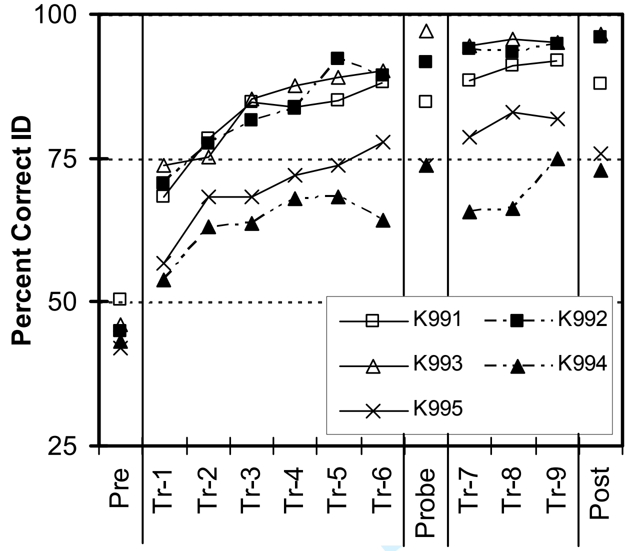
Table 2

Average percent correct identification of vowels in real words spoken by Trained and New talkers for three subject groups. Based on 1440 total observations.

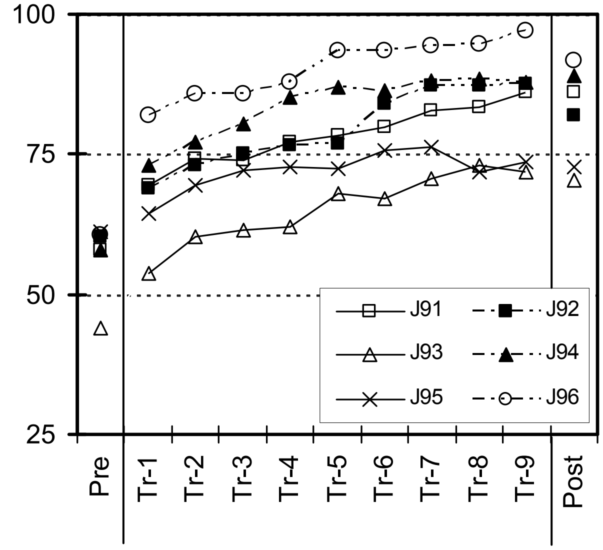
	Pre-test		Post-test	
Group	Trained (%)	New (%)	Trained (%)	New (%)
3V-9V	62.08	60.14	69.72	73.06
9V-3V	66.11	76.53	62.08	72.36
9V-9V	65.56	69.44	60.97	71.39


Participation Day
Group
1
2
3
4
5
6
7
8
9
10
11
9V-9V
Pre
Fullset
B
A
Fullset
Post
9V-3V
Pre
Fullset
B
A
Subset
Post
3V-9V
Pre
Subset
B
A
Fullset
Post

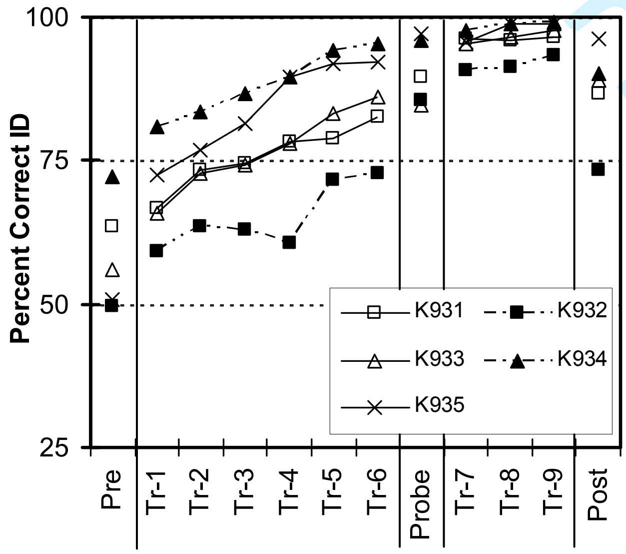
Korean 9V-9V



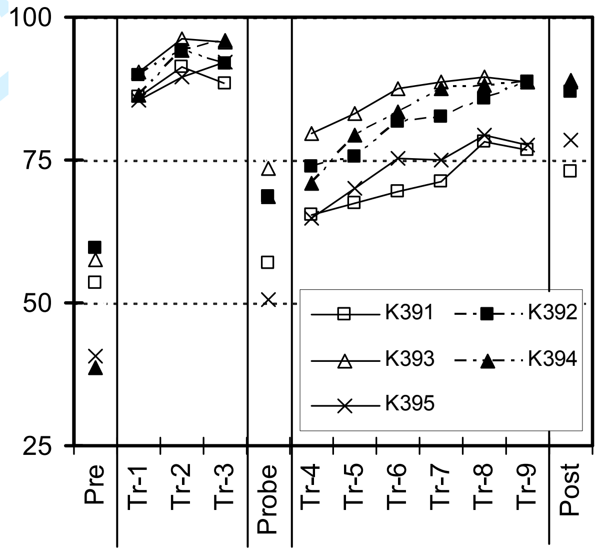
Japanese Fullset



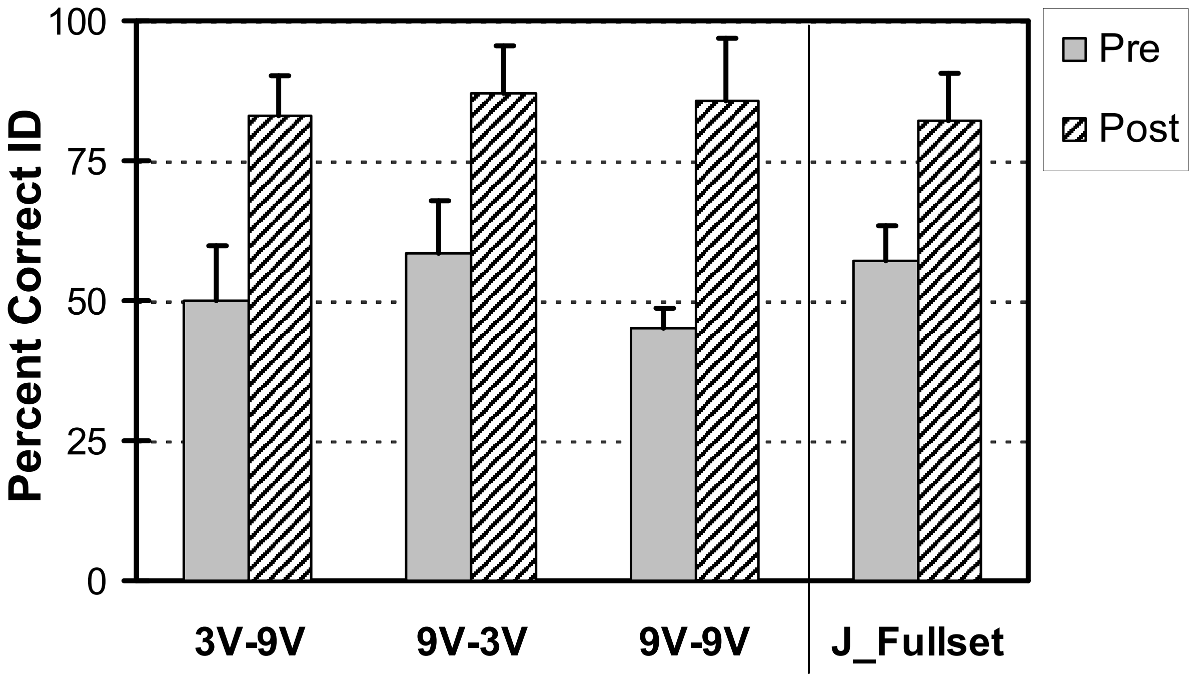
Korean 9V-3V



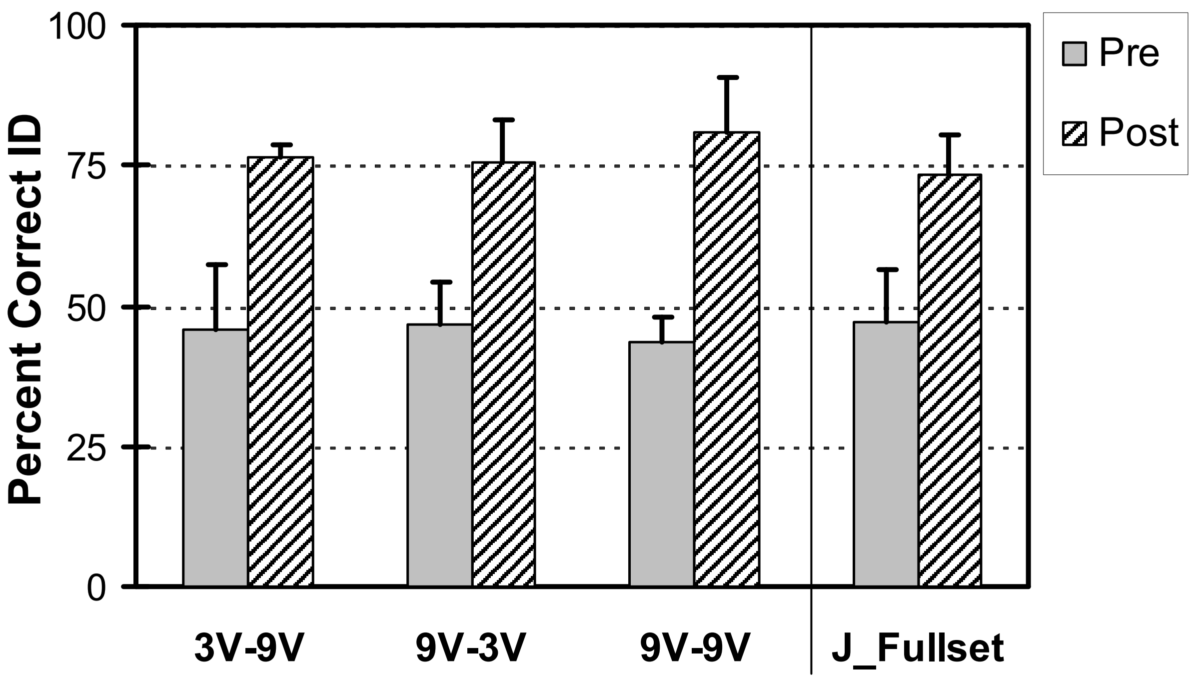
Korean 3V-9V

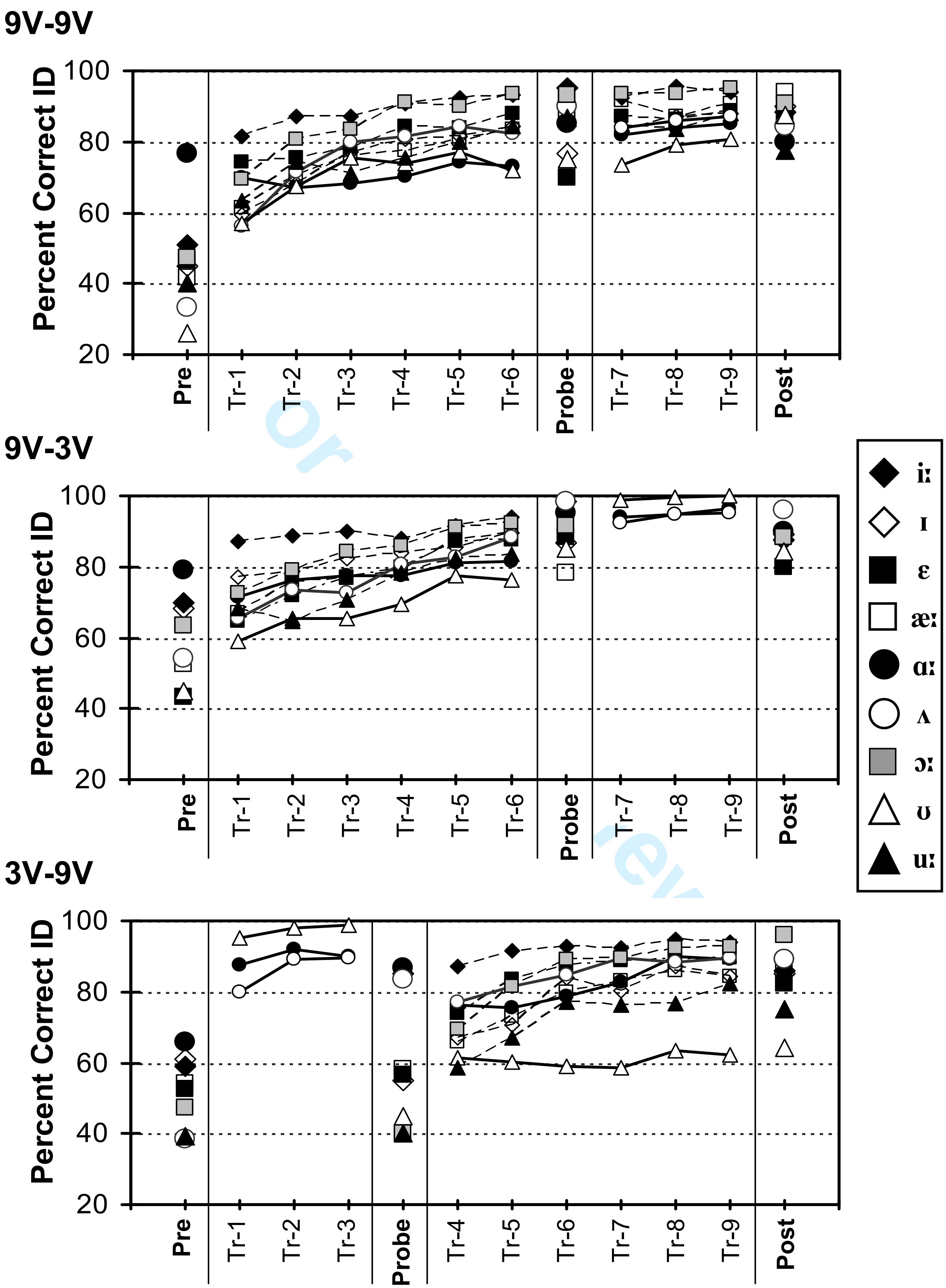


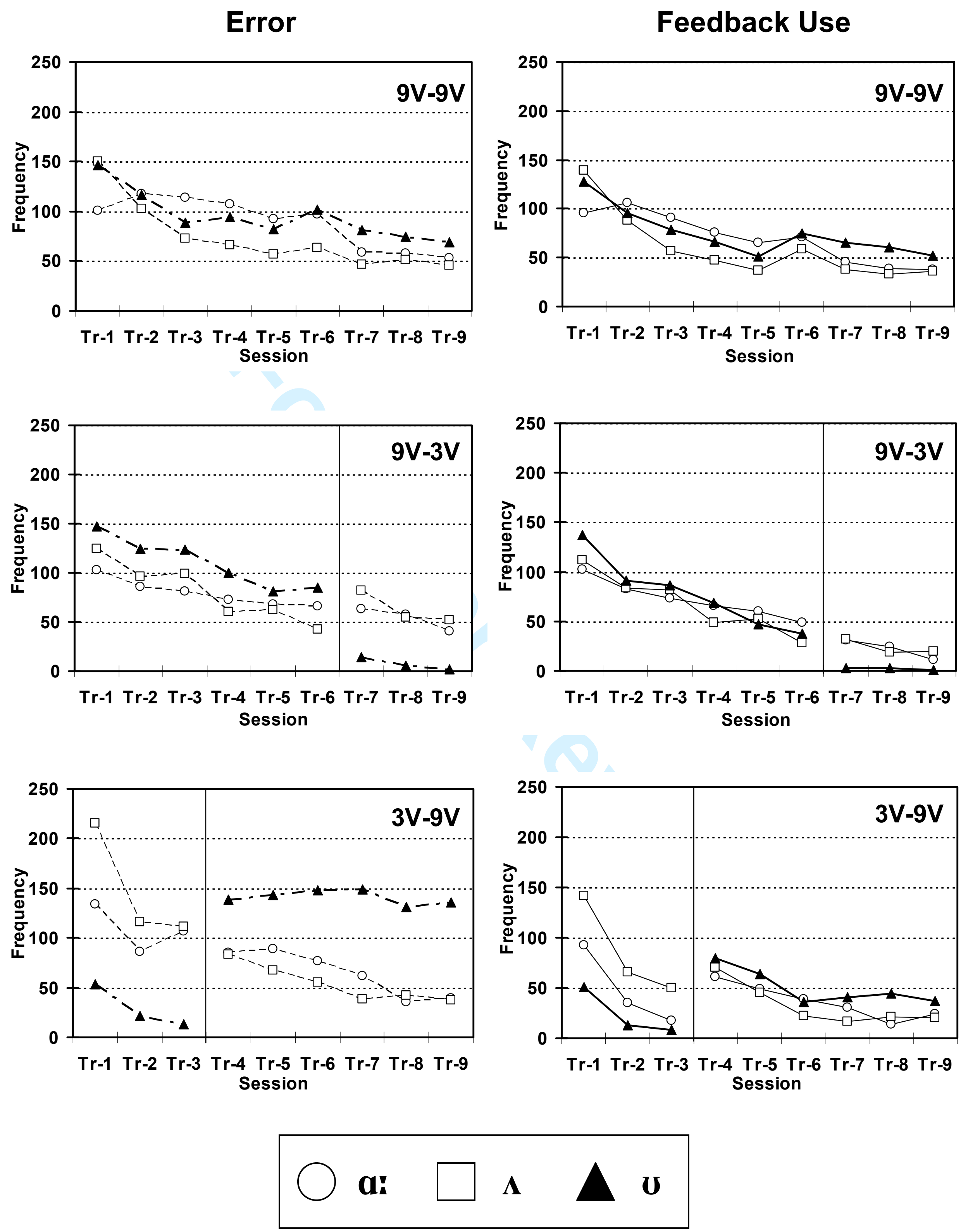
Trained Speakers



New Speakers




9V-9V
Percent Correct ID
100
80
60
40
20
Pre
Tr-1
Tr-2
Tr-3
Tr-4
Tr-5
Tr-6
Probe
Tr-7
Tr-8
Tr-9
Post
9V-3V
3V-9V
iː
ɪ
ɛ
æː
ɑː
ʌ
ɔː
ʊ
uː


Error
Feedback Use
9V-9V
9V-3V
3V-9V
Frequency
Session
Tr-1 Tr-2 Tr-3 Tr-4 Tr-5 Tr-6 Tr-7 Tr-8 Tr-9
0 50 100 150 200 250
ɑː
ʌ
ʊ