Published in final edited form as:

J Speech Lang Hear Res. 2008 December ; 51(6): 1480–1493. doi:10.1044/1092-4388(2008/07-0109).

Non-native Speech Perception Training Using Vowel Subsets: Effects of Vowels in Sets and Order of Training

Kanae Nishi and Diane Kewley-Port

Department of Speech and Hearing Sciences, Indiana University

Abstract

Purpose—Nishi and Kewley-Port (2007) 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 listeners 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.

In the past two decades, many speech perception training studies were conducted for second language (L2) 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, 2007; also see Bradlow, in press for review). The methods and materials used by these studies vary, but overall, their results showed that structured intensive training using stimuli produced by native speakers that include an adequate amount of allophonic variation can improve perception by L2 learners. Studies also showed that improvement due to training generalized to untrained voices, tokens, and positions in words. However, previous studies typically trained on a small number of contrasts, and except for Nishi & Kewley-Port (2007), 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.

Several important issues about the selection of materials are considered in the following review of the traditional minimal-pair based method that is often used in clinic and language classrooms. In a typical minimal pair approach, a client (student) is asked to differentiate a pair

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

A portion of this paper was presented at the 150th Meeting of the Acoustical Society of America in Minneapolis, Minnesota, 2005.

of words that differ by one phonemic feature (e.g., “kick” and “pick” for initial consonant pair differing in place; “put” and “pat” for vowel minimal pair; “bit” and “bid” 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 study (Nishi & Kewley-Port, 2007) that a training vowel subset chosen based on difficulty can improve the identification of trained vowels, but improvement 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 /i:/-/ɑ:/-/u:/, 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., perception of AE /i:/-/ɪ/ or /ʌ/-/ɑ:/, /æ:/-/ɛ/, etc.) (Strange et al., 1998; 2001). Apparently, L2 learners 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, 2007) in which two groups of Japanese listeners were trained on AE vowels using two training sets: nine vowels (/i:/, ɪ, ɛ, æ, ɑ:/, ʌ, ɔ:/, ʊ, u:/) covering the entire vowel space (fullset protocol) or three out of nine vowels (/ɑ:/, 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 three trained 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). Therefore, the present experiment was designed to evaluate the efficacy of hybrid protocols in comparison to the fullset only protocol. Furthermore, if hybrid protocols are found more effective than fullset only protocol, it is also helpful in practice to know whether early or later training on the subset is more effective. For this reason, two hybrid protocols were devised by combining the fullset and subset protocols but with their orders in training reversed between the two protocols.

A more theoretical motivation for this experiment was to examine the efficacy of the fullset only training with listeners with different first language (L1). Models of L2 speech learning (Best, 1995; Flege, 1995) have suggest that initial non-native phoneme perception is an assimilation process in which L1 categories are used to represent L2 categories. Naturally, in such processes, a strong influence from the L1 phonetic inventories as well as in the realization of allophonic variation for “similar” phonemes in L1 and L2 are inevitable. Flege specifically predicts that “similar” L2 phonemes are more difficult than “new” ones to master perfectly because L1 categories will continue to be used for such L2 phonemes. If this prediction is correct, then the more similar the L1 vowel system is to L2's, the more difficult learning would be. In our previous study (Nishi & Kewley-Port, 2007), Japanese was chosen because its vowel system was substantially different from that of AE, both in the number of spectral categories

and the phonemic use of vowel duration. In the present study, a new L1 group, Korean, was chosen because it is more similar to AE both in the number of spectral categories and the use of vowel duration (see description below). To examine the similarity issue, the results from Japanese fullset group in Nishi & Kewley-Port study were compared with those of a Korean group that completed a comparable training protocol. It was predicted that Korean listeners will initially identify more AE vowels correctly than Japanese group due to the similarity of vowel system. After training, however, there are three possible outcomes: 1) similarity between L1 and L2 vowel systems hinders learning, and Korean listeners will not perform as well as Japanese listeners at post-training test; 2) similarity promotes learning because less modification of vowel system is required and Korean listeners will perform more poorly than Japanese listeners; and 3) similarity does not influence learning and both Japanese and Korean listeners will perform equally well after training.

Briefly, AE has 10 monophthongs [i, ɪ, e, æ, ɛ, ʌ, ə, ɔ, ʊ, u], four true diphthongs [aɪ, aʊ, ɔɪ, ju], and two “smaller” diphthongs [eɪ, oʊ] that involve less spectral movements than the true diphthongs and the extent of diphthongization vary depending on speaker (Ladefoged, 1993). Many of these AE vowels are distinguished primarily by spectral properties, but these vowels can be grouped as 11 inherently long [i, eɪ, æɪ, ɛɪ, ɔɪ, oʊ, u, aɪ, aʊ, ɔɪ, ju] and four short [ɪ, ɛ, ʌ, ʊ] vowels (Peterson & Lehiste, 1960). Strange et al. (2005) reported that the average duration ratio between seven of the inherently long [i, eɪ, æɪ, ɛɪ, ɔɪ, oʊ, u] and four short [ɪ, ɛ, ʌ, ʊ] vowels is approximately 1.3.

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, 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 [ɔ] 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 [w̥]. The two front rounded vowels [y, ø] are often realized as onglides [wⁱ, 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 do not use duration phonemically, but for stress or rhythmic purposes (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, u-uu]) where vocalic duration is strictly phonemic (Shibatani, 1990). The average long/short Japanese vowel duration ratio 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 is.

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. As stated above, there are three possible outcomes for the efficacy of the 9V-9V protocol. If L1 vowel system is found to hinder/promote learning, then such a result indicates that a vowel training requires an L1-specific protocol. If Korean 9V-9V achieves similar performance as Japanese fullset, then it indicates that the structure of vowel categories in L1 and their similarity to L2 vowels do not influence the outcome as long as many vowels are included. Naturally, the two hybrid training groups were also expected to show improved performance after training. However, we expected that the overall post-test performance for listeners trained using hybrid protocols would 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 (K394) had already graduated from the Intensive English Program, and had been in the U.S. for 11 months when she completed training. 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 (K933) was from the Kyengsang dialect region, where vowel duration is reported to be used phonemically (Lee & Ramsey, 2000; Sohn, 1999). K394 and K933 were not excluded from the study because their response patterns at pretest 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. The 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 pretest 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. As a result of the random assignment of listeners to the conditions, one male participated in the 9V-9V group, and two each in the 9V-3V and the 3V-9V groups.

Stimulus materials

The stimulus materials were the same as the previous study (Nishi & Kewley-Port, 2007). Additional details of the recording apparatus, recording procedures, and stimulus preparation, can be found in Nishi and Kewley-Port (2007).

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 /C₁VC₂ə/ where C₁-C₂ 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 C₁, and /b, p, d, t, g, k, s, z, ʃ, m, n/ for C₂ (see Appendix A). The purpose of including RW stimuli was to examine generalization in more varied consonantal contexts than NSW, and no effort was made to make all consonantal contexts for RW stimuli equivalent across the vowels. The NSW stimuli were used both in training and tests, but the RW stimuli were used only in the test. Each stimulus included one of the nine AE monophthongs /i:, ɪ, ε, æ:, ɑ:, ʌ, ɔ:, ʊ, u:/. All stimuli 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 stimuli were excised from the sentences. Stimulus materials for the tests, probe, and training are described below.

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 produced both test and training stimuli (trained speakers); speakers F1 and M2 produced test stimuli (new speakers). The selection procedure of speakers was reported previously (Nishi & Kewley-Port, 2007).

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 relative to when stimulus sets were switched. Probe-A contained the tokens produced by the new speakers, and probe-B contained the tokens produced by the trained speakers. All participants took probe-A first, but the orders of two speakers in each probe part 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, 2007). The first set included all nine vowels and was used for the fullset protocol. The other set included only the three more difficult vowels, /ɑ:, 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-in 10th-order Butterworth filter on the Tucker-Davis Technologies PF1. All stimuli were presented only to a 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 in training only to the right ear through Sennheiser HD250 linear II headphones. Stimulus files were low-pass filtered at 5 kHz using an IIR filter that approximated the 10th-order Butterworth filter used for the tests and probe. The listening level was adjusted to be 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 procedures were the same as the previous study (Nishi & Kewley-Port, 2007), 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 $\times$ 2 key words $\times$ 5 randomizations = 90 total questions; for key words, see Appendix B). 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 RW 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 AE vowel in the stimuli (the first vowel in /C₁VC₂ə/ 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 72 trials (4 words $\times$ 9 vowels $\times$ 2 tokens) per block for RW and 108 trials (6 consonantal contexts $\times$ 9 vowels $\times$ 2 tokens) per block for NSW. Participants were allowed to take breaks between blocks. All participants completed tests in approximately 90 minutes.

Training—Training presented only the NSW tokens in 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 $\times$ 6 contexts $\times$ 2 tokens; subset: 3 vowels $\times$ 6 contexts $\times$ 2 tokens $\times$ 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 also 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

Table 1 presents the initial vowel confusions observed for the 15 Korean participants at pre-test. Stimulus vowels are listed in column 1. The second row presents the response categories. Numbers in individual cells are frequencies of responses observed for a particular stimulus-response combination. The average percent correct identification across vowels, participants, and speakers was 48%, which is comparable to that of Japanese listeners (53%) Nishi & Kewley-Port (2007) reported. Two (/ʌ, ʊ/) of the three vowels (/ɑ:, ʌ, ʊ/) chosen for the subset training were among the least accurately identified vowels (39% and 31%, respectively). Although the low vowel /ɑ:/ (72%) was by far the most accurately identified among the nine vowels, when the use of response categories were examined, /ɑ:/ was also most frequently used (639 times) as incorrect responses (at least 1.45 times more often than other categories). A further examination revealed that the most frequent incorrect use of /ɑ:/ was for the target vowel /ʌ/ (248 responses out of 639), whereas /ɑ:/ was identified as /ʌ/ only 38 times. This suggests that these Korean participants did not discriminate between /ʌ/ and /ɑ:/, but responded with /ɑ:/ for both /ɑ:/ and /ʌ/. This, in turn, limited incorrect use of /ʌ/ for /ɑ:/ and increased correct identification of /ɑ:/. The response vowel /ʌ/ was frequently used for the target /ʊ/ (150 responses). Overall, pre-test confusions indicated that the vowels /ɑ:, ʌ, ʊ/ in subset 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, 2007). 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 sessions. 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 group¹. For the hybrid groups, 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.

¹The learning curve of K994 revealed a somewhat different pattern from the other participants. Although speculative, we believe this was due to the specific strategy this participant employed during training. According to the casual queries of each participant (to obtain information about training difficulties for future use), this particular participant attempted to discover and memorize small token-specific acoustic differences between the "Correct" and "Wrong" sounds played during feedback. None of the other participants (either Korean or Japanese in previous study) mentioned this strategy. However, feedback stimuli for the "Wrong" sound were randomly selected from the two tokens spoken by the same talker for each button press. Apparently, this particular participant's strategy led to confusion and poorer performance during training. Our speculation is based on the sudden improvement in her performance at probe suggesting that she was less confused because there was no feedback during tests.

Overall Performance Change

This section reports the changes in participants' overall 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 (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 a 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 and control groups (J_fullset and J_control, respectively). Error bars indicate the within-group standard deviations.

All five groups' pre- and posttest average vowel scores for trained speakers were subjected to a mixed ANOVA to find out whether the groups were similar before and after training. Results indicated there was no significant main effect of group, but the main effect of test ($F(1, 21) = 151.53, p < .001$) and test $\times$ group interaction ($F(4, 21) = 10.53, p < .001$) were significant. A follow-up analysis on the interaction using Tukey's HSD revealed that there was no difference among the five groups at pre-test, but all groups except the J_control performed significantly better at posttest than pretest ($p < .001$ for all). These results warrant that the Korean listeners' pre-test performance was equivalent to that of Japanese listeners, and that other differences found among the three Korean groups at post-test can be attributed to training. These results also indicate that 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 five groups. The results closely resembled the results for the trained speakers (upper panel), and the four trained groups revealed trends very similar to one another. A mixed-design ANOVA (group $\times$ test) revealed a significant main effect for test ($F(1, 21) = 148.17, p < .001$) and a group $\times$ test interaction ($F(4, 21) = 7.55, p < .001$). A follow-up analysis on interaction using Tukey's HSD revealed that the five groups performed similarly at pre-test, the four trained groups considerably improved after training ($p < .001$ for all), but that no difference was found between pre- and posttest for J_control ($p = .81$). These results indicate that all three Korean groups generalized their 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 speakers for the three Korean 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 speakers, respectively. Columns 4 and 5 present the post-test scores. Numbers in the parentheses are standard deviations.

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 $\times$ group $\times$ speaker 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, 2007), and suggest that the training on NSW also improved identification of vowels in the untrained RW stimuli. Note that vowels in these RW tokens 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 3V-9V training group suddenly declined. This implies the possible influence of the order of training sets. For this reason, the Korean data presented in Figure 2 were summarized for each vowel and presented as Figure 4. 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-9V group 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, /*α*/ and /*Λ*/, was not as large as /*ʊ*/.

The above observations were confirmed by an ANOVA (group $\times$ training session $\times$ vowel) performed for average scores for subset vowels during the six fullset training sessions. Results revealed 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 $\times$ group [$F(4, 24) = 4.08$, $p < .02$] and vowel $\times$ session [$F(10, 120) = 2.13$, $p < .03$]. None of the other effects were significant. A further examination was performed on the vowel $\times$ group interaction using Tukey HSD. Results indicated that only the 3V-9V group showed significant difference between vowel /*ʊ*/ and other two subset vowels ($p < .001$ for /*ʊ*-/*Λ*/ comparison, $p < .002$ for /*ʊ*-/*α*/ 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 developed a task-specific strategy that maximized their identification accuracy by relying on some non-major acoustic cues for AE vowels. 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 /*α*/ and /*Λ*/ increased steadily for the three groups, while only the 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 /ɑ:/ and /Λ/ 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-9V group 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 9V-3V 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, 2007), 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 used nine vowels for nine training sessions; the second protocol 9V-3V used nine vowels for the first six sessions, and 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, and 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.

Despite the similarity of Korean vowel inventory to that of AE, Korean listeners' pre-training performance did not differ from that of previously reported Japanese listeners. 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, 2007), indicating that a vowel perception training protocol using a large set of highly variable, naturally produced stimuli is equally effective for learners from first languages (L1) with different similarities of vowel inventories to that of AE. Apparently, the similarity of vowel inventories between the L1 and L2 does not influence the success of vowel training as long as many vowels are used. 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. In fact, the 9V-9V and 9V-3V were essentially equivalent. However, a

closer look at the changes in performance 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 (/ɑ:/ and /Λ/) 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 calculations in reference to the learned category characteristics. Applying this explanation, the subset protocol represents an association task between three categories and their 12 exemplars. 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. Following a principle of least effort, trainees may have discovered the more effective strategy by choosing one requiring less effort. Obviously, memorizing 12 exemplars is efficient for the subset training, but not for the fullset training.

Then why were the results of 9V-3V different from 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 (/ɑ:, Λ, 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 they were 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 did 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 that are produced in the same number of consonantal contexts as in the larger set – may lead to inefficient learning strategies.

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. Research has

produced converging results 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 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 the effects of consonant set size on 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 for effective learning 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.

Acknowledgements

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 Jae Hee Lee for translating consent forms and instructions into Korean.

Appendix A

List of Real Word Materials

i:	ɪ	ɛ	æ:	ɑ:	ʌ	ɔ:	ʊ	u:
cheese	chip	check	chat	chop	bus	cause	bush	boom
deep	hit	deck	mat	job	cut	chalk	could	lose
heap	miss	net	pass	knock	hug	dog	hood	soon
seat	pick	tech	tab	shop	tuck	talk	wood	zoom

Appendix B

List of Key Words Used in Response System

ir	ɪ	ɛ	æɪ	ɑɪ	ʌ	ɔɪ	ʊ	uɪ
team meet	hit kids	set ten	hat map	hot mop	gum luck	sauce walk	foot look	two who

References

- Barlow JA, Gierut JA. Minimal pair approaches to phonological remediation. *Seminars in Speech and Language* 2002;23:57–67. [PubMed: 11938491]
- Best, CT. A direct realist view of cross-language speech perception. In: Strange, W., editor. *Speech perception and linguistic experience: Issues in cross-language research*. Baltimore: York Press; 1995. p. 171-204.
- Bradlow, AR. Training non-native language sound patterns: Lessons from training Japanese adults on the English /r/-/l/ contrast. In: Hansen Edwards, JG.; Zampini, ML., editors. *Phonology and Second Language Acquisition*. John Benjamins; in press
- Dalby, J.; Kewley-Port, D. Explicit pronunciation training using automatic speech recognition. In: Holland, M., editor. *Tutors that Listen: Speech Recognition for Language Training, Special Issue of the Journal of the Computer Assisted Language Learning Consortium (CALICO)*. 16. 1999. p. 425-445.
- Flege, JE. Second language speech learning: Theory, findings, and problems. In: Strange, W., editor. *Speech perception and linguistic experience: Issues in cross-language research*. Baltimore: York Press; 1995. p. 233-277.
- Gierut, JA. Enhancement of learning for children with phonological disorders. In: Slifka, J.; Manuel, S.; Matthies, M., editors. *From sound to sense: 50+ years of discoveries in speech communication*. Boston: MIT; 2004. p. 164-172.
- Hirata, Y.; Tsukada, K. The effects of speaking rates and vowel length on formant movements in Japanese. In: Agwuele, A.; Warren, W.; Park, SH., editors. *Proceedings of the 2003 Texas Linguistics Society Conference*. Somerville, MA: Cascadilla Proceedings Project; 2004. p. 73-85.
- Iverson P, Hazan V, Bannister K. 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* 2005;118:3267–3278. [PubMed: 16334698]
- Jamieson DG, Morosan DE. Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception and Psychophysics* 1986;40:205–215. [PubMed: 3580034]
- Juslin P, Olsson H, Olsson AC. Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General* 2003;132:133–156. [PubMed: 12656301]
- Labov, W.; Ash, S.; Boberg, C. *Atlas of North American English: Phonology and Phonetics*. Berlin: Mouton de Gruyter; 2006.
- Ladefoged, P. *A course in Phonetics*. Fort Worth: Harcourt Brace; 1993.
- Ladefoged, P.; Maddieson, I. *The Sounds of the World's Languages*. Malden, MA: Blackwell; 1996.
- Lee, I.; Ramsey, SR. *The Korean language*. New York: State University of New York Press; 2000.
- Lively SE, Logan JS, Pisoni DB. 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* 1993;94:1242–1255. [PubMed: 8408964]
- Logan JS, Lively SE, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America* 1991;89:874–886. [PubMed: 2016438]
- McClaskey CL, Pisoni DB, Carrell TD. Transfer of training of a new linguistic contrast in voicing. *Perception and Psychophysics* 1983;34:323–330. [PubMed: 6657433]
- McReynolds LV, Bennet S. Distinctive feature generalization in articulation training. *Journal of Speech and Hearing Disorders* 1972;37:462–470. [PubMed: 4648936]

- Miller JD, Dalby JM, Watson CS, Burleson DF. Training experienced hearing-aid users to identify syllable-initial consonants in quiet and noise. *The Journal of the Acoustical Society of America* 2004;115:2387.
- Morosan DE, Jamieson DG. Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word-position or task. *Journal of Speech and Hearing Research* 1989;32:501–511. [PubMed: 2779195]
- Nishi K, Kewley-Port D. Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets. *Journal of Speech, Language and Hearing Research* 2007;50:1496–1509.
- Peterson GE, Lehiste I. Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America* 1960;32:693–703.
- Pruitt JS, Jenkins JJ, Strange W. 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* 2006;119:1684–1696. [PubMed: 16583912]
- Shibatani, M. *The languages of Japan*. New York: Cambridge University Press; 1990.
- Smith JD, Minda JP. Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 2000;26:3–27.
- Sohn, HM. *The Korean language*. New York: Cambridge University Press; 1999.
- Strange W, Akahane-Yamada R, Kubo R, Trent SA, Nishi K. Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *The Journal of the Acoustical Society of America* 2001;109:1691–1704. [PubMed: 11325137]
- Strange W, Akahane-Yamada R, Kubo R, Trent SA, Nishi K, Jenkins J. Perceptual assimilation of American English vowels by Japanese listeners. *Journal of Phonetics* 1998;26:311–344.
- Strange W, Bohn OS, Nishi K, Trent SA. Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *The Journal of the Acoustical Society of America* 2005;118:1751–1762. [PubMed: 16240833]
- Strange W, Bohn OS, Trent SA, Nishi K. Acoustic and perceptual similarity of North German and American English vowels. *The Journal of the Acoustical Society of America* 2004;115:1791–1807. [PubMed: 15101657]
- Strange W, Dittmann S. Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics* 1984;36:131–145. [PubMed: 6514522]
- Studebaker G. A “rationalized” arcsine transform. *Journal of Speech and Hearing Research* 1985;28:455–462. [PubMed: 4046587]
- Wang Y, Spence MM, Jongman A, Sereno JA. Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America* 1999;106:3649–3658. [PubMed: 10615703]
- Yang B. A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics* 1996;24:245–261.

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

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.

Korean 9V-9V **Japanese Fullset** **Korean 9V-3V** **Korean 3V-9V** **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 (2007) is also presented for comparison with the Korean 9V-9V group (upper left).

Trained Speakers

New Speakers

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 two Japanese groups (J_Fullset, J_Control) from Nishi and Kewley-Port (2007).

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 observed at pretest for all Korean participants (N=15). Overall percent correct identification = 48%.
Total number of observations per stimulus vowel is 720

Stimulus	Response (Count)									
	i:	ɪ	ɛ	æ:	ɑ:	ʌ	ɔ:	ʊ	u:	
i:	415	300	1		1	1	1		1	1
ɪ	57	381	175	94	3		1	3	6	1
ɛ	3	4	296	287	77	28	24			
æ:	4		214	396	80	5	21	1		
ɑ:	1		8	29	519	38	124			1
ʌ			19	19	248	280	132	15		1
ɔ:			3	5	207	166	328	10		7
ʊ	3	5	6	2	22	150	75	224		1
u:					1	52	2	369		232
										296

Table 2
Average percent correct identification of vowels in real words spoken by trained and new speakers for three Korean groups. Based on 1440 total observations

Group	Pre-test		Post-test	
	Trained (<i>SD</i>)	New (<i>SD</i>)	Trained (<i>SD</i>)	New (<i>SD</i>)
3V-9V	62.08 (7.38)	60.14 (10.65)	69.72 (6.07)	73.06 (7.95)
9V-3V	66.11 (9.06)	62.08 (3.76)	76.53 (8.22)	72.36 (6.35)
9V-9V	65.56 (7.06)	60.97 (4.97)	69.44 (2.69)	71.39 (5.16)