ABSTRACT

This paper reports some of the findings of a three-week perceptual training experiment involving Japanese and Korean learners of English as a second language. Other studies have shown that under appropriate conditions, adults can learn to identify nonnative speech sounds such as American English /r/ and /l/ with generalization to novel stimuli, transfer to new tasks and retention. Such conditions of “high variability” have included multiple talkers’ voices and a large stimulus training set involving these sounds across word positions. The current study investigated additional sources of variability by exploring the effects of the adjacent vowel and talkers’ facial cues on perceptual training. Thus, word position, adjacent vowel, and training type [auditory-visual (AV) vs. auditory-only (A-only); multiple talker vs. single talker] were independent variables. Results indicated significantly greater improvement in identification accuracy for AV vs. A-only training. Visual input contributed the most to the bimodal percept for the more difficult pre-training phonetic environments. Significant effects were found for vocalic environment as well as word position and training talker. Findings also revealed successful transfer to novel stimuli, a new talker, production improvement, and earlier word identification in connected speech.

1. INTRODUCTION

Various studies using a “high variability” stimulus set have demonstrated successful perceptual training of Japanese speakers to identify American English (AE) /r/ and /l/ with robust generalization to novel stimuli and a new talker’s voice [1] and transfer to production improvement [2]. A similar stimulus set has been used to train Korean speakers [3]. However, the focus of these studies has been limited to auditory input and to the consideration of word position as phonetic environment. The present study was conducted to expand our view of the sources of variability affecting the perceptual accuracy of AE /r/ and /l/ by both Japanese and Korean listeners by investigating the influence of the adjacent vowel and the contribution of visual input from talkers’ articulatory gestures. In an earlier study on the McGurk effect, a significant increase was noted in the identification accuracy of AE /r/ in the syllable /ra/ by nonnatives when corresponding visual cues were present [4] suggesting that training using videotaped stimuli could optimize the information value of visual cues in speech perception. Numerous studies have reported that initial position, and clusters versus singleton are more challenging for Japanese speakers, and final cluster position is more difficult for Koreans [1,3]. Based on the notion that second-language (L2) category formation and perceptual processing may be impeded by similarity between L2 and first-language (L1) sounds [5], one would predict that the more difficult contexts for Korean listeners would be the velarized syllable-final AE /Il/ where Korean has a nonvelarized sound, and intervocalic position where Korean has a voiced alveodental flap. Japanese has a voiced apical flap in the dental or alveolar region occurring utterance-initially and intervocically with acoustic, articulatory and perceptual similarities to the AE flap suggesting that perceptual accuracy of AE liquids in initial and intervocalic positions for them would be more difficult.

The present study was motivated by the following questions: a) Is there better performance with AV training than with A-only? The addition of the visual modality as a second channel of input for L2 learning provides another set of stimulus dimensions on which similarities and distinctions can be drawn to facilitate perceptual category development. b) How does identification accuracy vary in relation to both aspects of phonetic environment: word position and adjacent vowel? The speechreading literature has documented the effects of various vowels [6], word positions [7], and talkers [8], on the discernability of articulatory gestures by native observers. I hypothesized that visual information would be of less value when /rl/ and /Il/ were presented in clusters with labial consonants and in contexts with rounded vowels. c) What is the role of visual cues and vocalic context in the transfer of training to novel stimuli, a new talker and production improvement?

2. EXPERIMENT ONE – JAPANESE

Eight Japanese subjects ranging in age from 18-25 who were intermediate-level learners in an Intensive English Program (IEP) in the U.S. were randomly assigned to each of two experimental groups, AV and A-only in order to compare the two training methods (both involving multiple talkers). A control group was given the pretest and posttest (perception and production) but no training. Subjects had been in the U.S. for 1-7 weeks at the time of training. Prior English instruction had emphasized grammar and reading. Their IEP program did not focus on the perception or production of specific sounds. The perception phase of the experiment consisted of the sequence: pretest, training, posttest, and two tests of generalization (one with a familiar talker from training, and one with an unfamiliar talker). Stimuli for both the AV and A-only training groups involved multiple talkers and were presented via videotape. For the AV group, the testing segments involved three conditions (AV, A-only, and V(video)-
only) to provide a measure of the contribution made by each modality. A two-alternative forced choice task was used with /r/ and /l/ in a balanced design with word position and adjacent vowel as independent variables.

2.1. Materials and Procedure

Minimal pair stimuli contrasted /r/ and /l/ in initial singleton, initial cluster, final singleton and final cluster. For each of these positions in the pretest and posttest, four minimal pairs were selected with each of the following adjacent vowels: /u/, /o/, /l/, /e/ or /æ/, /i/ or /l/. Introvocalic position was also used but analyzed separately because of the two vocalic environments. For the testing segments, time (pretest, posttest), modality (for the AV group), word position, and vowel were within-subjects variables. There were insufficient minimal pairs to continue all vowels in all positions for the posttest. Thus, vowels were grouped following speechreading guidelines: /o, u, l, i/.

To prepare the stimuli, the audio and video signals were recorded as native speakers (NSs) of general American English pronounced the words in a randomized order, prompted by cue cards. Recordings were made in a television studio with ceiling-mounted lighting using a light gray background. A SONY Hi-8 videocamera was used with an Electrovoice lavaliere microphone. A full-sized image of each talker’s head was captured with a fully visible, lower jaw drop. These recordings were dubbed onto Betacam SP tape and loaded for editing into AVID Media Composer (MC8000) version 5.51 for Macintosh and were digitized at a sampling rate of 44.1 kHz. Edited stimuli were transferred to broadcast quality VHS format for presentation. To evaluate intelligibility, all stimuli were presented auditorily to a group of NSs for identification before the program began. There were no errors involving /r/ and /l/. Stimuli were then presented to subjects via a 27-inch TV monitor and videocassette recorder.

For both training groups, the sequence for each stimulus on videotape involved a tone, presentation of the utterance (AV or A-only), 4 seconds to circle the correct word on the response sheet, feedback (correct answer shown in 96 pt. font on the screen), and repetition of the utterance to promote the development of an association between the sound under test, its phonetic realization (and visual for the AV group) in context. Feedback versus no-feedback perceptual training is superior especially under conditions involving a high degree of within-category variability [9]. Each training session was observed to ensure that subjects made their responses prior to the feedback presentation. Training stimuli were spoken by three females and two males. The stimuli recorded by each of the talkers were transferred to broadcast quality VHS format for the AV group. Those mean pretest scores had differed by only 2.6%. Overall, AV training provided significantly greater improvement in perceptual accuracy \(F(1,14)=7.87, p<0.05\).

Further comparisons revealed that when visual information was present, the increase in scores after training was significantly greater for /r/ and /l/ in initial clusters versus singleton, and scores for initial positions showed greater improvement than those for final positions. Thus, the visual modality contributed strongly to the AV percept for the initial positions hypothesized to be more difficult based on L1 phonology, and these became more informative with training.

Across the three weeks of training, scores increased significantly from 72% at the end of Week 1 to 93% at the end of Week 3 for the AV group and 64% to 80% for the A-only group. Identification accuracy varied for the tokens produced by the different talkers. The significant difference was between Talkers 1 and 3 (both female).

Improvement in the identification accuracy of /r/ and /l/ with training varied significantly according to word position, vowel and talker \(F(24,1253)=3.37, p=0.0001\) for both training groups. This is evident in the contrast between Figures 2 (initial singleton) and 3 (final singleton) for the AV group. Initial singleton scores for Talker 2 were higher perhaps due to her very rounded articulation of /t/ and relatively salient apical contact for /l/. Across talkers, identification accuracy was lower for the liquids in initial positions with rounded vowels and higher in final positions especially with /æ/. Scores for Talkers 4 and 5 for initial clusters with /u, o/ were the lowest; however, the scores for these talkers for final clusters with /u, l/ were the highest.

In tests of generalization, mean accuracy was significantly affected by talker familiarity only in the V-only condition where scores were higher for TGI; otherwise, performance was comparable for the familiar and unfamiliar talkers.
In addition, perceptual training resulted in significant production improvement with an increase in mean ratings for the AV group from 5.2 to 6.3 and for the A-only group from 4.8 to 5.7. Post-hoc tests also revealed a significantly higher mean score for initial singleton and for contexts with /a/ compared to the high and rounded vowels. Following training, variability decreased in the ratings across word positions.

3. EXPERIMENT TWO - KOREAN

Eight Korean learners of English were assigned to each of four experimental groups. The AV and A-only training groups were further divided into two groups each: one received training with multiple talkers and the other with a single talker. The comparison of multiple- versus single-talker training had been conducted earlier with Japanese speakers using a population comparable to that of Experiment 1 [1]. This discussion will focus on the results for the multiple-talker AV and A-only training groups.

The subject population in Experiment 2 was comparable to the Japanese group in terms of age range, level of English proficiency, and previous English experience. Stimuli and procedural details were the same as in Experiment 1.

3.1. Results and Discussion

Identification scores for each experimental group were tabulated and analyzed as in Experiment 1.

As shown in Figure 4, for the AV group, mean identification accuracy in testing in each modality showed a significant increase as did the pre- and posttest scores for the A-only group as a result of training \((p=0.0001)\). The control group did not show improvement.

In contrast to the Japanese, scores for the Koreans across modalities were significantly higher for /r/ and /l/ in initial versus final positions. Similar to the Japanese though, scores overall were significantly higher for contexts with /i:/ and generally lower for the rounded vowels. The increase in mean scores for final singleton and cluster positions when visual input was present was significantly greater than the increase in the A-only condition.

To compare the effects of training type, again the A-only scores for the AV group were compared with those from the A-only group. Significantly greater improvement was noted for the AV versus the A-only group \([F(1,14)=7.46, p<.05]\).

Across the three weeks of training, scores increased significantly from 78% to 95% for the AV group. There was a significant effect of talker for this group also. Accuracy for tokens produced by Talker 1 was the highest for both the Japanese and Korean. As in Experiment 1, the interaction of Talker x Position x Vowel for the Koreans was also significant \([F(24,1253)=2.60, p=0.0001]\). The areas of greatest talker
difference were the final positions where accuracy rates were lower, especially with the rounded vowels and /i, I/ as shown in Figures 5 (initial singleton) and 6 (final singleton). In general, identification accuracy varied significantly within and across talkers of both genders for all training groups.

Figure 5: Korean AV Group: Mean percent correct in training for initial singleton position by talker (T) and vowel.

Figure 6: Korean AV Group: Mean percent correct in training for final singleton position by talker (T) and vowel.

A comparison of results between the AV multiple- and single-talker training groups in the tests of generalization revealed that in the absence of visual input, identification accuracy for the single-talker group was significantly lower for tokens produced by the unfamiliar talker in the untrained vowel context /e:/ (Group x Talker x Vowel [F(1,14)=5.97, p<.05]) and with the relatively difficult contexts for these learners – initial singleton with rounded vowels, and final singleton with unrounded ones.

The effects of perceptual training on production in Experiment 2 used the same stimuli and procedure. Within the multiple-talker groups, mean production ratings increased significantly from 5.01 to 6.39 (AV group) and from 4.84 to 5.92 (A-only). Variation in pretest scores across positions significantly decreased with time with the greatest improvement noted for /l/ in final singleton. Scores were higher for /l/ and /l/ with /a/ compared to the other vowels.

4. CONCLUSION

Recent studies in perceptual training have demonstrated that natural speech, stimulus variability, an identification paradigm and feedback during training contribute to the development of perceptual categories robust across phonetic environments and talkers, thus, enabling generalization to novel stimuli and talkers, and transfer to production improvement. In addition to replicating these findings, the present study revealed that simultaneously training two modalities, auditory and visual, was superior in improving perceptual accuracy than training only one [10]. The contribution of visible speech was most evident in the perception of /t/ and /l/ in the more difficult word positions (i.e., initial for Japanese and final for Koreans). It is also evident that in terms of phonetic environment, both adjacent vowel and word position have significant effects on perception and production of these sounds. Further studies have indicated that such focused perceptual training with minimal pairs including /l/ and /l/ also transfers to earlier word identification in connected speech [11]. The above findings support L2 speech acquisition as a bimodal process with obvious pedagogical implications (e.g. the use of an animated talking head such as “Baldi” that provides displays of articulatory movements inside the mouth) [12].

5. REFERENCES