THE EFFECTS OF TALKER VARIABILITY ON THE ACQUISITION OF NON-NATIVE SPEECH CONTRASTS

James S. Magnuson and Reiko A. Yamada
ATR Human Information Processing Laboratories, 2-2 Hikaridai, Seika, Soraku, Kyoto, 619-02, Japan; email: magnuson@hip.atr.co.jp, yamada@hip.atr.co.jp

ABSTRACT

In previous /r/-/l/ training studies that stressed talker variability, stimuli were blocked by talker. We compared mixed-talker training (in which the talker changed from trial to trial) with blocked-talker training, to see if blocked training gives subjects adequate experience to adapt to talker differences. We found that mixed training led to better talker-adaptation, although the effect was mitigated when the number of talkers used in training was increased.

INTRODUCTION

Previous /r/-/l/ training paradigms for Japanese listeners have shown that talker variability in training is a crucial element for post-test generalization to new talkers and stimuli. Subjects trained with stimuli produced by only one talker sometimes fail to acquire generalization ability [1], whereas subjects trained with stimuli produced by five talkers consistently show good post-test generalization (e.g., [2]). When high talker variability has been stressed, stimuli have been blocked by talker. Results from perceptual studies suggest that how such training is structured may be an important consideration.

Yamada and Tohkura [3] reported that when untrained Japanese adults attempt to identify English /r/ and /l/, they appear to set criteria based on the range of cues they hear in a series of trials. Given the entire series of stimuli from a synthesized /r/-/l/ continuum, Japanese response functions were quite similar to English speakers'. However, given only the most /r/-like or /l/-like half, Japanese boundaries shifted such that they continued to respond "R" approximately 50% of the time (as opposed to native speakers, whose nearly-categorical functions were not affected by such changes). It seems that Japanese subjects expected to hear equal numbers of /r/ and /l/ tokens, and set response criteria accordingly when the range of stimuli changed.

We have found a similar *range-bias* effect due to talker variability [4]. When Japanese adults identified /r/ and /l/ stimuli produced by only one talker, they responded "R" approximately 50% of the time. When stimuli from five talkers were mixed, the overall rate of "R"-response (*R-rate*) was still about 50%. However, the *R-rate* to particular talkers differed significantly between blocked-(single) and mixed-talker conditions. It seems that subjects set a single criterion based on the range of cues they heard within a block, rather than evaluating each stimulus independently.

This strategy was successful when the stimuli were produced by one talker. However, when stimuli from different talkers were mixed, some talkers' /r/s and /l/s sounded "R-like" or "L-like" relative to other talkers' productions, as was reflected in significant, talker-specific changes in bias (*R-rate*).

Given that previous studies which stressed the importance of talker variability in /r/-/l/ training have only presented stimuli blocked by talker [2], it follows that trained subjects might have difficulty adapting to between-talker differences, as did subjects in our perceptual tests [4]. The experiments we report here were designed to examine this possibility.

EXPERIMENT 1

We trained two groups of subjects to identify English /r/ and /l/. One group of subjects was trained in a mixed-talker condition (i.e., the talker could change on any trial), and the other was trained in a blocked-talker condition (i.e., the talker remained constant within a block). This comparison was made to determine whether either condition better promotes the ability to adapt to talker differences in non-native speech contrast perception.

Method

Subjects. 12 native speakers of Japanese with limited English training were paid to participate in Experiment 1.

Stimuli. The training stimuli were 79 minimal pairs of real English words contrasting /r/ and /l/ in 5 phonetic contexts: initial singleton, initial cluster, inter-vocalic, final singleton and final cluster. The stimuli were produced by 2 male native speakers of American English. One of the talkers was found to be "R-like" relative to several others (i.e., *R-rate* to his productions was ~.50 in a blocked-talker condition, but in a mixed-talker condition, increased significantly), and the other was found to be "L-like" [4]. The stimuli were selected from the set used in [2].

The test stimuli were 25 minimal pairs of real English words contrasting /r/ and /l/ in initial position (selected from the set used in [2]). The stimuli were produced by 4 native speakers of American English: the 2 training talkers and 2 female talkers (one R-like and one L-like [4]).

A set of generalization stimuli produced by a male talker not used in training or testing consisted of 32 minimal pairs of real English words contrasting /r/ and /l/ in the five training contexts, and 8 pairs of filler items contrasting other phonemes. These items were used in previous training studies [1,2] and were included to facilitate comparisons with those studies.

Procedure. A 2-alternative forced-choice paradigm was used for all sessions. On each trial, orthographic forms of a minimal pair of words were displayed on a CRT. Then, one of the pair was presented over headphones. The subject responded by pressing a key indicating the side of the screen on which he or she thought the word played over the headphones was displayed. The orthographic forms were randomly assigned to the right or left side of the CRT. Subjects received feedback about their responses only in training trials. If the subject answered correctly, a chime sounded and the next trial began. If the subject answered incorrectly, a buzzer sounded, the orthographic forms were randomly assigned to the left or right, and the word was played again. This continued until the subject answered correctly. For every three words identified correctly on the first attempt, subjects received an additional 1 yen as a monetary incentive.

There were 7 parts to the experiment: pretest, training sessions 1 and 2, mid-test, training sessions 3 and 4, and post-test. Subjects participated in the pretest and training session 1 on day 1, training session 2 on day 2, mid-test and training session 3 on day 3, training session 4 on day 4, and the post-test on day 5.

Test sessions consisted of nine blocks. Four were 50-trial blocks of stimuli produced by only one of each of the four testing talkers (blocked-talker). The same 200 stimuli were presented in four mixed-talker blocks. The order of blocked and mixed blocks was determined randomly. The ninth block consisted of the generalization items.

Subjects were randomly assigned to two training groups of six subjects. The blocked-training group heard training stimuli produced only by the "R-like" talker in the first 2 training sessions, and only stimuli produced by the "L-like" talker in the last 2 training sessions. Both groups were trained with four 948-trial sessions (total training trials: 3792). The mixed-training group heard equal numbers of stimuli produced by each talker presented in random order in each training session.

Results and Discussion

For this paper, we will focus on pretest--post-test comparisons; space constraints prevent us from including mid-test results. Both groups showed significant accuracy improvement on the /r/-/l/ items in the generalization set between pre- and post-tests ($F(1,10)>7$, $p<.05$; blocked group: .64 to .74; mixed: .62 to .72).

Both groups also improved on test items between pre- and post-tests, as can be seen in Table 1. An analysis of simple effects showed that the post-test difference between groups was significant for at the blocked level of talker condition ($F(1,10)=30$, $p<.001$). Note that subjects in the blocked group were significantly less accurate in the mixed-talker portion of the post-test than in the blocked-talker portion ($F(1,5)=19$, $p<.01$). That there was no such difference in the mixed group suggests that subjects in the mixed group applied

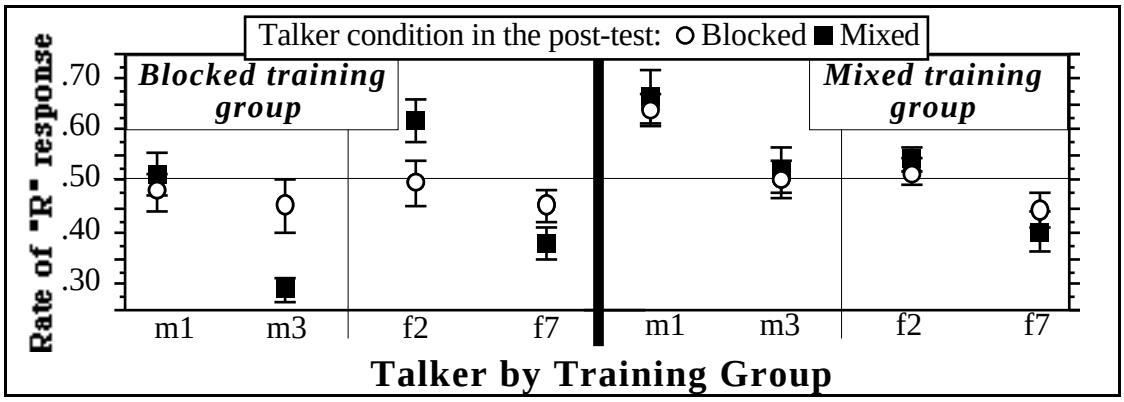

Figure 1: The talker by talker condition interaction by group in the Experiment 1 post-test. m1 and m3 were training talkers; f2 and f7 were heard only in tests. m1 and f2 appeared R-like to untrained subjects and m3 and f7 appeared L-like [4]. Bars show standard error.

similar response criteria in both talker conditions in the post-test.

ANOVAs on the *R-rate* data confirm that the mixed group used similar criteria in both talker conditions. Both groups showed interactions of talker and talker condition in the pretest ($p<.05$ for the mixed group, $p=.05$ for blocked). The blocked-talker training group still showed a strong post-test talker by talker condition interaction ($F(3,15)=16$, $p<.001$). In contrast, the interaction was not significant for the mixed training group ($F(3,15)<.5$, $p>.7$; see Figure 1). These subjects' responses to particular talkers were not affected by talker condition. Both groups of subjects responded "R" approximately 50% of the time in the blocked- and mixed-talker conditions. However, in the post-test, the blocked group showed large talker-specific differences in *R-rate*, similar to those observed previously with untrained subjects; *R-rate* increased for R-like talkers and decreased for L-like talkers.

Table 1: Accuracy in Experiment 1. All pre-post differences were significant (Tukey HSD post-hoc test, .05 level).

Group	*Talker condition*	*Pretest*	*Post-test*
Blocked training	blocked	.58	.73
	mixed	.59	.67
Mixed training	blocked	.58	.63
	mixed	.56	.63

It appears that the mixed training condition promoted greater ability to adapt to talker-specific differences in /r/ and /l/ productions, since subjects trained in that group responded "R" equally often to particular talkers -- even unfamiliar ones -- in both talker conditions in the post-test. However, the accuracy differences between groups suggest that there may be trade-offs between talker variability and stability in training: stability promoted higher accuracy for the blocked-training group.

EXPERIMENT 2

The results of Experiment 1 indicate that mixed-talker training may promote better adaptation to talker differences for even unfamiliar talkers. However, they do not provide an adequate basis for comparison with previous /r/-/l/ training paradigms that have stressed talker variability, as those studies have used 5 talkers in training [2]. In Experiment 2, we increased the number of training talkers to 5 and the number of testing talkers to 7, and added a third training condition.

Method

Subjects. 30 native speakers of Japanese with limited English training were paid to participate in Experiment 2.

Stimuli. We used the same sets of testing, training and generalization stimuli we used in Experiment 1. However, the test stimuli were produced by 7 talkers (the same 4 used in Experiment 1, with the addition of 2 males and 1 female). The training stimuli were produced by 5 talkers (the 2 males used in Experiment 1 with the addition of 1 male and 2 females).

Procedure. The same 2-alternative forced-choice paradigm used in Experiment 1 was used, although the

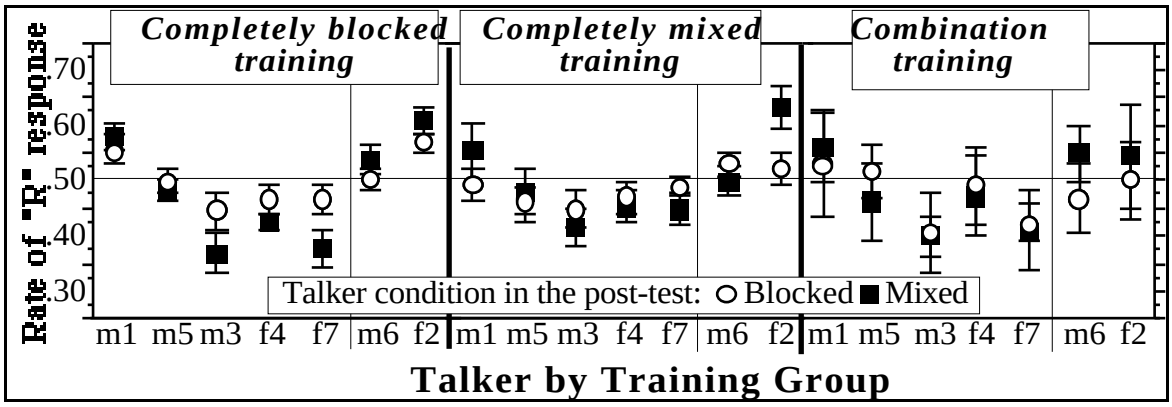

Figure 2. The talker by talker condition interaction by group in the post-test in Experiment 2. Talkers m1, m5, m3, f4 and f7 were training talkers; m6 and f2 were heard only in tests. m1, m5 and f2 have appeared R-like to untrained subjects, and m3, f4 and f7 were L-like [4] (m6 has not been tested). Bars represent standard error.

design was changed slightly. On the first day, subjects participated in a test with the generalization materials and a multiple talker pretest similar to the one used in Experiment 1, with the addition of 50 trials in blocked and mixed conditions for each of the 3 new talkers. On each of the next 5 days, subjects were trained in 158-trial sessions with feedback (total training trials: 3950). Subjects were assigned to one of three training groups (blocked, mixed, or combination, described below). At the time of this writing, 5 subjects had been assigned to the combination group, 12 to mixed, and 18 to blocked.

The blocked group heard 1 talker on each training day (a different talker each day). The mixed group heard equal numbers of stimuli produced by each of the training talkers mixed together in random order each day. The combination group also heard equal numbers of stimuli produced by each talker, but the stimuli were blocked by talker. On day 7, a post-test on the generalization and test materials used in the pretest was given.

Results and Discussion

Blocked, mixed and combination groups all improved significantly from pretest (.64, .65, and .63, respectively) to post-test (.76, .76, and .73) on the /r/-/l/ generalization items (Tukey's HSD post-hoc test, $p<.05$). All groups also showed significant improvement on the test materials between pretest and post-test in both talker conditions ($F>8$, $p<.05$). There were no significant differences between groups.

The interaction of talker and talker condition in the post-test *R-rate* results for each group are shown in Figure 2. The interaction was significant for subjects in the mixed ($F(6,66)=3.1$, $p<.01$) and blocked groups ($F(6,102)=3.7$, $p<.01$). The interaction was not significant for subjects in the combination group ($F(6,24)<1$, $p>.6$). It appears that the addition of more talkers increases the variability in the training set to such a degree that the mixed vs. blocked training advantage observed in the first experiment is diminished. This suggests that a combination of stability and variability may promote the ability to adapt to differences between talkers.

CONCLUSION

Our results indicate that talker variability, and also how it is organized, are important considerations in non-native speech contrast training. Further work is required to determine the nature of the observed trade-offs between stability and variability in learning non-native contrasts.

REFERENCES

[1] Magnuson, J.S., Yamada, R.A., Tohkura, Y., Pisoni, D.B., Lively, S.E., & Bradlow, A.R. (1995). *Proc. Acoust. Soc. Japan, Spring, 1995*, 393-394.

[2] Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., & Yamada, T. (1994). *J. Acoust. Soc. Am.*, 96, 2076-2087.

[3] Yamada, R.A., & Tohkura, Y. (1992). *Perception & Psychophysics*, 52, 376-392.

[4] Magnuson, J.S., & Yamada, R.A. (1994). *J. Acoust. Soc. Am.*, 95, 2872.