Published in final edited form as:
J Acoust Soc Am. 1993 September ; 94(3 Pt 1): 1242–1255.

Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories

Scott E. Lively, **John S. Logan**[a], and **David B. Pisoni**[b]
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405

Abstract

Two experiments were carried out to extend Logan *et al.*'s recent study [J. S. Logan, S. E. Lively, and D. B. Pisoni, J. Acoust. Soc. Am. **89**, 874–886 (1991)] on training Japanese listeners to identify English /r/ and /l/. Subjects in experiment 1 were trained in an identification task with multiple talkers who produced English words containing the /r/–/l/ contrast in initial singleton, initial consonant clusters, and intervocalic positions. Moderate, but significant, increases in accuracy and decreases in response latency were observed between pretest and posttest and during training sessions. Subjects also generalized to new words produced by a familiar talker and novel words produced by an unfamiliar talker. In experiment 2, a new group of subjects was trained with tokens from a single talker who produced words containing the /r/–/l/ contrast in five phonetic environments. Although subjects improved during training and showed increases in pretest–posttest performance, they failed to generalize to tokens produced by a new talker. The results of the present experiments suggest that variability plays an important role in perceptual learning and robust category formation. During training, listeners develop talker-specific, context-dependent representations for new phonetic categories by selectively shifting attention toward the contrastive dimensions of the non-native phonetic categories. Phonotactic constraints in the native language, similarity of the new contrast to distinctions in the native language, and the distinctiveness of contrastive cues all appear to mediate category acquisition.

INTRODUCTION

In a recent study, Logan *et al.* (1991) demonstrated that native speakers of Japanese could be trained to improve their perception of the English /r/–/l/ contrast by using a two-alternative forced-choice identification paradigm and a highly variable stimulus set. Moderate improvements in identification accuracy and response latency were observed after only three weeks of laboratory training. Performance also generalized to a limited extent to a new voice and to new words. The present paper addresses several issues concerning stimulus variability raised by Logan *et al.*'s study. Their training materials included both talker variability and variability due to phonetic environment. In the present investigation, the contribution of each of these sources of variability is considered separately. In experiment 1, a group of Japanese listeners was trained with a set of tokens produced by five talkers. The English phonemes /r/ and /l/ appeared in initial singleton, initial consonant clusters, and intervocalic positions. In experiment 2, a new group of listeners was trained with a single talker who produced tokens contrasting /r/ and /l/ in all five of the original phonetic

[b] To whom correspondence should be sent.

[a] Currently at the Department of Psychology, Carleton University, Ottawa, Ontario K1S 5B6, Canada.

environments used by Logan *et al*. The relative importance of each source of variability to generalization was assessed by examining each source separately and comparing the results to Logan *et al*.'s earlier findings.

According to the prevailing view of cross-language speech perception, when listeners acquire new phonetic contrasts, they develop abstract, context-invariant units, such as phonemes, templates, or prototypes (Best, 1992; Henly and Sheldon, 1986; Jamieson and Morosan, 1986, 1989; Strange and Dittmann, 1984). Prototypes are defined in Rosch's sense of the term and are assumed to have a number of special psychological properties (Rosch, 1975a,b; Rosch and Mervis, 1975). First, the mental representations of phonetic categories are assumed to have foci, which serve as the "best" or most representative member of the category. Second, phonetic segments are judged according to their "distance" from the focus of the relevant category: good category members are close to the prototype of the category and poor category members are further away from the prototype. As such, category goodness is a function of distance from the prototype (Samuel, 1982). Finally, the foci of phonetic categories are assumed to be better perceived and have more stable representations in long-term memory (Kuhl, 1991a).

Psychological evidence for phonetic prototypes comes from several sources (see Kuhl, 1991a). First, some members of a phonetic category are responded to faster, more accurately, and with higher confidence ratings or goodness judgments than others (see Kuhl, 1991a,b; Miller and Volaitis, 1989; Pisoni and Tash, 1975; Volaitis and Miller, 1992). Second, some within-category discriminations are better than others. For example, Kuhl found that tokens of /i/ surrounding a good exemplar were more poorly discriminated than tokens of /i/ that surrounded a poor category exemplar (Kuhl, 1991a,b; Kuhl *et al*., 1992). Third, selective adaptation studies have demonstrated that tokens rated as good exemplars of a category make better adapters than tokens either closer to or further away from category boundaries (Miller *et al*., 1983; Samuel, 1982). Finally, Miller (1977) and Repp (1976) demonstrated that good category members are more effective competitors in dichotic listening conditions than poor category members.

In addition to psychological evidence for prototypical representations of phonetic categories, formal linguistic analyses have encouraged the view that representations for speech sounds are abstract and context-invariant. The sound system of a language is typically defined in terms of phonemic categories and category membership is based on traditional linguistic criteria, such as complementary distribution, free variation, and phonetic similarity. According to this perspective, only linguistically significant information is represented in phonemic categories. As a consequence, a great deal of other detailed information about a talker's voice, speaking rate, or the surrounding phonetic context is assumed to be filtered out or normalized so that only linguistically distinctive information is preserved (see Pisoni, 1992).

I. EXPERIMENT 1: TRAINING IN THREE PHONETIC ENVIRONMENTS

The assumption that phonetic categories are represented by prototypes in long-term memory has had several important consequences for cross-language studies of speech perception. For example, Strange and Dittmann (1984) trained Japanese listeners in a discrimination paradigm with a synthetic "rock"–"lock" stimulus continuum. They assumed that training listeners with tokens that contrasted /r/ and /l/ in initial singleton position would allow subjects to form a prototype that could be applied to other phonetic environments. However, this strategy ignores the spectral and durational differences between /r/ and /l/ in different phonetic environments (Dissosway-Huff *et al*., 1982; Lehiste, 1964). As a result, only half of Strange and Dittmann's subjects showed any improvement in identifying naturally produced /r/–/l/ contrasts in initial position. Moreover, only one of eight subjects showed any improvement for contrasts in initial consonant clusters. These findings suggest that discrimination training with a small set of tokens from one phonetic environment may be ineffective in modifying listeners' phonetic perception, particularly when generalization is tested with tokens from different phonetic environments in an identification paradigm.

Jamieson and Morosan (1989) found similar results using an identification training paradigm. They attempted to train native speakers of French to identify the English /θ/ and /ð/ distinction. Although training with a prototypical member of each phonetic category produced some increase in identification accuracy, improvements were not as large as those obtained by subjects who were trained with a much more variable stimulus set using a perceptual fading technique (Jamieson and Morosan, 1986). The authors concluded that stimulus variability in an identification training paradigm was important in establishing new phonetic categories because it lessens the importance of within-category differences, while increasing the importance of between-category differences.

Logan *et al.* (1991) addressed the issue of context-invariant phonetic categories by training listeners with stimuli from a wide range of phonetic environments. Subjects were presented with the /r/–/l/ contrast in initial singleton, initial consonant clusters, intervocalic, word-final consonant clusters, and final singleton positions. Increases in performance that were observed from the pretest to the posttest and during training demonstrated the effectiveness of the training paradigm. Identification performance was most accurate for /r/ and /l/ in final consonant clusters and final position and least accurate for /r/ and /l/ in initial consonant clusters. Goto (1971), Dissosway-Huff *et al.* (1982), Mochizuki (1981), and Sheldon and Strange (1982) all reported similar patterns of performance across phonetic environments for Japanese listeners. Logan *et al.* argued that during perceptual learning listeners formed context-sensitive representations for the new phonetic contrast. To account for the effect of phonetic environment, we assumed that selective attention was more easily drawn to some dimensions than to others (Jusczyk, 1993a,b; Nosofsky, 1986, 1987; Strange and Jenkins, 1978). Phonotactic constraints in the native language, coarticulatory effects and durational cues all appear to influence changes in selective attention across different phonetic environments (Dissosway-Huff *et al.*, 1982; Henly and Sheldon, 1986; Sheldon and Strange, 1982).

The differential effects of phonetic environment provided the motivation for experiment 1. We hypothesized that training would be more effective if additional trials were dedicated to phonetic environments in which identification performance was initially poor (cf. Atkinson, 1972). In experiment 1, words contrasting /r/–/l/ in final consonant clusters and final singleton position were removed from the training set because numerous studies have demonstrated that Japanese listeners' performance is at ceiling for contrasts in these environments (Dissosway-Huff *et al.*, 1982; Goto, 1971; Mochizuki, 1981; Sheldon and Strange, 1982). Only tokens containing /r/ and /l/ from the difficult environments (initial singleton, initial consonant clusters, and intervocalic positions) were presented during training. Robust generalization to novel tokens and a novel talker was encouraged by training listeners with words from several talkers. Although the number of phonetic environments was reduced compared to the earlier Logan *et al.* study, we predicted that the training set was sufficiently variable to provide listeners with a rich variety of acoustic cues to the /r/–/l/ contrast that could be brought to bear on new tokens produced by a novel talker during the generalization test. We also predicted that accuracy across all phonetic environments would be similar when subjects were trained selectively on only the most difficult phonetic environments.

A. Training and testing procedure

The training and testing procedures used in the present experiments were identical to those used by Logan *et al.* (1991). Subjects responded in a two-alternative forced-choice identification task throughout the experiment. Immediate feedback was given only during the training phase. The minimal uncertainty of the two-alternative forced-choice procedure, combined with immediate feedback, was used to promote the formation of robust new phonetic categories (Jamieson and Morosan, 1986). Furthermore, the identification training procedure avoids the attentional demands of discrimination training by encouraging subjects to group similar objects into the same category (Lane, 1965, 1969). Discrimination training, in contrast, requires listeners' to attend to within-category differences and may not promote robust category formation (Carney *et al.*, 1977; Liberman *et al.*, 1961; Pisoni, 1973; Werker and Logan, 1985; Werker and Tees, 1984).

B. Method

1. Subjects—Subjects in experiment 1 were six native speakers of Japanese. Each had been in the United States for two months prior to the beginning of the experiment. Listeners reported that they had studied English for a mean of 7.5 yr before coming to the United States, although their English training in Japan stressed reading and writing rather than speaking skills. Subjects rated their ability to read and understand spoken English as fair to poor and reported that they communicated in English about 50%–75% of the time. All subjects were enrolled in English language acquisition classes at Indiana University in Bloomington at the time of testing and were paid $5.00 for each laboratory session. None of the subjects reported any history of speech or hearing problems at the time of testing.

2. Stimulus materials—The stimuli used in the pretest and posttest phase were identical to those used by Logan *et al.* (1991).[1] A male native speaker of English produced 48 words contrasting /r/ and /l/ in four phonetic environments. Four minimal pairs each were drawn from initial singleton, initial consonant clusters, intervocalic, and final singleton positions. Sixteen additional minimal pairs that did not contrast /r/ and /l/ were also collected. In addition, a matched set of nonwords was developed to examine the effects of lexical status on Japanese listeners' ability to identify /r/ and /l/. The pseudowords were generated by substituting two phonemes in each of the original words used in the pretest and posttest (bleed → gleep; breed → greep).

The training stimuli were a subset of the words used by Logan *et al.* (1991). Thirteen minimal pairs contained initial singleton contrasts, 24 pairs contained initial consonant cluster contrasts, and five contained intervocalic contrasts. Each of the 84 stimulus words was produced by five native speakers of English—three males and two females. Thus, subjects were trained with a total of 420 unique stimulus tokens. None of the words used in training was presented in the pretest or posttest or in the subsequent tests of generalization.

The stimuli used in the tests of generalization were identical to those used by Logan *et al.* (1991). Two sets of tokens were recorded. The first set consisted of 99 additional items from one of the female training talkers (talker 4). Five phonetic environments were represented in this stimulus set. Thirty-eight words had /r/ or /l/ in initial singleton position, 32 had /r/ or /l/ in initial consonant clusters, 3 had an intervocalic /r/ or /l/, 11 had /r/ or /l/ in final consonant clusters, and 15 had /r/ or /l/ in final singleton position. The second set of 95 novel tokens was produced by a new male native speaker of English. In this set of stimuli, 38 items had /r/ or /l/ in initial singleton position, 29 had /r/ or /l/ in initial consonant clusters, 4 had an intervocalic /r/ or /l/, 7 had /r/ or /l/ in final clusters, and 17 had /r/ or /l/ in final singleton

[1] The words we chose for the pretest and posttest were identical to those used by Strange and Dittmann (1984).

position. None of the words used in generalization was presented in the pretest or posttest or during training. The number of words containing /r/ or /l/ was approximately balanced across the two stimulus sets.

Stimulus preparation for all phases of the experiment was identical. Talkers sat in a sound-attenuated IAC booth and recorded isolated words using an Electro-Voice DO54 microphone and an Ampex AG-500 tape deck. No special pronunciation instructions were given. Words were presented individually to talkers on a CRT monitor inside the recording booth during each recording session. After the recordings were made, the items were low-pass filtered at 4.8 kHz and were digitized at 10 kHz using a 12-bit analog-to-digital converter. The digitized files were edited into word-length stimuli and were equated for RMS amplitude using a signal processing package running on a PDP 11/34 laboratory computer. Each item was tested for intelligibility in an open-set identification task by a separate group of native speakers of English. Any item that was not intelligible to at least 85% of the listeners or produced an /r/–/l/ confusion was replaced by a new token from the same talker.

3. Procedure—The effects of training were assessed by using a pretest–posttest procedure and several tests of generalization. A two-alternative forced-choice identification procedure was used throughout the training and testing phases of the experiment. All training and testing took place in a sound-attenuated room equipped with individual cubicles for each subject, which contained a desk, a two-button response box, and a CRT monitor. Stimuli were presented to subjects at 80 dB SPL over a set of TDH-39 matched and calibrated headphones using a PDP 11/34 laboratory computer. The computer collected individual responses and latencies during all phases of the experiment. Response latencies were measured from the onset of the stimulus item to the subjects' button press response. All subjects were tested and trained individually.

Pretest and posttest performance were assessed with two sets of stimulus materials. In the lexical test, subjects were randomly presented with a token from a set of 16 minimal pairs that contrasted /r/ and /l/ in four phonetic environments. Eight filler pairs that did not contain /r/ and /l/ were also presented. In the pseudoword test, subjects were presented with 16 minimal pairs of pronounceable nonwords that contrasted /r/ and /l/. Eight minimal pairs of pseudowords that did not contain /r/ and /l/ were also presented in this test. Each set of stimuli was presented twice and the order of presentation was counterbalanced across subjects.

On each trial of the pretest and posttest, the two members of a minimal pair were displayed in the lower left- and lower right-hand corners of the CRT monitor and a “GET READY” prompt appeared 500 ms later in the middle of the screen. One member of the minimal pair was presented over the headphones 500 ms after the warning prompt was displayed. Subjects were asked to identify the stimulus by pressing the button on the response box that corresponded to the word displayed on the screen. Responses to /l/'s were always made with the left hand and responses to /r/'s were always made with the right hand. Subjects were given a maximum of 4 s to respond. The next trial began 1000 ms after the previous trial ended. No feedback was given during the pretest or the posttest phases of the experiment. Altogether, both tests lasted approximately 30 min.

Subjects began the 15-day training period after they completed the pretest. The training procedure used the same two-alternative forced-choice identification task employed by Logan *et al*. The only difference between the pretest and training was that feedback was given on every trial during the training phase. A light on the response box was illuminated corresponding to the correct response alternative. The next trial was presented 1000 ms later.

If subjects made an error, the feedback light remained illuminated and the stimulus token was repeated again over the headphones. Subjects were not required to respond during the repetition.

Stimuli used for the training portion of the experiment were produced by five talkers. Subjects heard only one talker per training session and talkers were presented in the same order throughout training. Thus, subjects heard each talker three times during training. Each talker produced 84 tokens that contrasted /r/ and /l/ in initial position, initial consonant clusters, and intervocalic position. Tokens produced by each talker were presented three times each during a training session, yielding a total of 252 training trials per session. Subjects were given a 5-min break after 176 trials. Training sessions lasted approximately 30 min each.

Subjects received two tests of generalization following the completion of training and the posttest. The procedure was identical to the pretest and the posttest. Listeners identified the items produced by the familiar talker first, followed by the items from the unfamiliar talker. Each item was presented twice, in random order, during each generalization test. No feedback was given during generalization. The two tests of generalization lasted about 30 min.

C. Results

1. Pretest–Posttest—Mean accuracy and mean response latencies were submitted to separate analyses of variance, (ANOVAs). Lexical status, phonetic environment, and pretest-versus-posttest were within-subjects variables in both analyses. *Post hoc* tests were conducted using Tukey’s HSD procedure.

Figure 1 displays accuracy and latency measures for the pretest and posttest as a function of phonetic environment. Separate ANOVAs revealed a main effect for test in the accuracy and latency data [$F_{pc}(1,5) = 7.76$, $p < 0.01$; $F_{rt}(1,5) = 14.51$, $p < 0.01$]. Mean accuracy increased from 79.96% on the pretest to 85.57% on the posttest. This finding replicates the earlier results of Logan *et al.* (1991). Mean response times decreased from 1165 to 931 ms. Equivalent changes in performance were obtained for the three training environments. Interestingly, Tukey’s HSD test revealed a significant decrease in response latencies for contrasts in word final singleton position, an environment in which no training was given. No main effects or interactions involving the lexical status variable were observed.

In addition to the main effects for test (pretest versus posttest), main effects for phonetic environment were also observed in both the accuracy and latency analyses [$F_{pc}(3,15) = 10.97$, $p < 0.01$; $F_{rt}(3,15) = 8.95$, $p < 0.01$]. Tukey’s HSD tests showed that accuracy was poorest for contrasts in initial consonant clusters. Mean response times tended to be slowest for targets in final position, although *post hoc* tests failed to localize this difference more precisely.

2. Training—Separate ANOVAs were conducted on mean accuracy and response latency scores. Week of training, talker, and phonetic environment were all treated as within-subject variables in the analysis.

The main effects for week of training indicate the success of the training procedure. As shown in Fig. 2, responses became significantly more accurate [$F(2,10) = 14.08$, $p < 0.01$] and faster [$F(2,10) = 17.09$, $p < 0.01$] as training progressed over the three weeks. Increases in accuracy were localized between weeks 1 and 2 of training, although response latencies decreased significantly during all three weeks of training.

As shown in Fig. 3, the effects of talker variability obtained by Logan *et al.* (1991) were replicated in experiment 1 [$F_{pc}(4,20) = 25.40$, $p < 0.01$; $F_{rt}(4,20) = 11.45$, $p < 0.01$]. Tukey's HSD tests revealed that subjects responded more accurately to talker 4 than to any other training talker, except to talker 5. Responses to talker 4 were significantly faster than to all other talkers, except to talker 3. Several other pairwise comparisons were also significant: Responses to talker 1 were slower than to talkers 3 and 5. Subjects responded more rapidly to talker 3 than to talker 2.

Main effects for phonetic environment were also observed in the accuracy and latency analyses [$F_{pc}(2,10) =40.77$, $p < 0.01$; $F_{rt}(2,10) = 16.80$, $p < 0.01$]. Responses to contrasts in initial singleton position were more accurate and faster than responses in initial consonant clusters. Accuracy was also greater for /r/'s and /l/'s in intervocalic position than in initial consonant clusters. Finally, response latencies were faster for contrasts in initial singleton position than for /r/'s and /l/'s in intervocalic position.

In addition to the main effects for phonetic environment and talker, an interaction between both variables was obtained in the analysis of the accuracy data [$F(8,40) = 2.21$, $p < 0.05$]. Figure 4 shows mean accuracy as a function of talker and phonetic environment. Tukey's HSD tests revealed that responses to /r/'s and /l/'s in initial singleton position were significantly more accurate than responses to targets in initial consonant clusters for talkers 2, 3, and 4. Identification of /r/'s and /l/'s in intervocalic position was also significantly higher than in initial consonant clusters for talkers 2 and 4.

3. Tests of generalization—Separate ANOVAs on the mean accuracy and latency scores collected during the tests of generalization revealed no significant effects of talker [$F_{pc}(1,5) < 1$; $F_{rt}(1,5) < 1$]. Mean accuracy was 88% for each talker. Mean response latency was 883 ms for the familiar talker and 854 ms for the unfamiliar talker. Accuracy and latencies were comparable to those observed in week 3 of training. Only a main effect for phonetic environment was observed in the accuracy analysis [$F(4,20) = 3.43$, $p < 0.05$]. Although Tukey's HSD tests failed to localize any significant differences, targets in final position and final consonant clusters were responded to most accurately, while targets in initial consonant clusters and intervocalic position were responded to least accurately.

D. Discussion

The results of experiment 1 extend Logan *et al.*'s (1991) previous findings: A high variability identification procedure is effective in training Japanese listeners to acquire English /r/ and /l/ in a laboratory setting. Three findings indicate the success of this paradigm. First, performance improved from the pretest to the posttest, even for a phonetic environment in which no training was given. Second, increases in accuracy and decreases in response latency were observed during all three weeks of training. Finally, subjects demonstrated evidence of robust generalization: that is, performance with both a familiar and unfamiliar talker was equivalent to performance observed at the end of training.[2]

Overall, the results of experiment 1 suggest three conclusions. First, the present findings show that moderate increases in identification performance can be obtained, even if the variability due to different phonetic environments is reduced. The size of the training benefits obtained in experiment 1 was comparable to those observed by Logan *et al.* (1991)

[2] Claims concerning robust generalization should be taken with one qualification. Given the large differences in performance as a function of talker during training, it could be the case that subjects would not generalize well to a different talker. This possibility could not be assessed in this experiment because of a lack of minimal pairs that contrast /r/ and /l/ in the different phonetic environments. The pretest–posttest stimuli, training tokens and test of generalization stimuli exhaust almost all of the /r/–/l/ minimal pairs in English. Thus, no additional new words were available to be produced by other talkers.

using a larger number of phonetic environments. By eliminating the phonetic environments in which Japanese listeners' *a priori* accuracy is high and devoting more training trials to difficult phonetic environments, performance increases were still obtained.

Second, effects of talker variability were observed throughout training. Accuracy and response latency varied widely as a function of talker. Talker 4 was responded to most accurately while talker 2 was responded to least accurately, again replicating the earlier results of Logan *et al*. Acoustic analyses of utterances produced by these talkers may be useful in determining the source of the differences in intelligibility. Despite these differences, however, we observed generalization to new words produced by an old talker and novel tokens produced by a new talker. Indeed, performance during generalization was equivalent to performance during the third week of training. These findings suggest that our training procedure encouraged robust acquisition of the /r/–/l/ contrast by Japanese listeners.

Third, differences in accuracy and response latency as a function of phonetic environment continued to be observed in each phase of the experiment, despite the fact that the number of training trials on difficult phonetic environments was increased compared to Logan *et al*. Initially, we hypothesized that additional training with the difficult contrasts would lead to equivalent performance across all phonetic environments. Indeed, we observed changes in identification accuracy from the pretest to the posttest for each of the training environments. However, even after training, performance for targets in initial consonant clusters was still substantially worse than for /r/'s and /l/'s in the final singleton position. Optimistically, this finding indicates that additional training may be necessary to obtain equivalent performance among all phonetic environments, particularly for difficult phonetic contexts such as initial consonant clusters.

To our knowledge, every experiment that has examined Japanese listeners' perception of /r/ and /l/ in different phonetic environments has found that contrasts in initial consonant clusters are the most poorly identified (see Gillette, 1980; Goto, 1971; Logan *et al*., 1991; Mochizuki, 1981; Sheldon and Strange, 1982; Strange and Dittmann, 1984). Several studies have attempted to localize the critical cue responsible for the differential performance across phonetic environments (Dissosway-Huff *et al*., 1982; Henly and Sheldon, 1986; Mann, 1986). In general, these studies have compared English /r/ and /l/ to their Japanese equivalent and have also examined spectral and temporal differences between these contrasts across different environments. In terms of phonetic similarity, Japanese has an /r/-like sound that resembles English /d/ or /t/. Depending on the vowel environment, it is produced either as a stop consonant or as a flap (Price, 1981; Yamada and Tohkura, 1992). In contrast to /r/, Japanese has no equivalent of the English /l/.

Acoustic analyses suggest differences in temporal and spectral characteristics across phonetic environments may also play a role in accounting for the differences in performance. Dissosway-Huff *et al*. (1982) examined gross temporal aspects of English /r/ and /l/ and found large differences in duration in different phonetic environments: Contrasts in final singleton position were consistently longer in duration than those in initial singleton position or initial consonant clusters. These acoustic measurements correlate well with perceptual data collected from Japanese listeners: Performance tends to be the best in environments in which the cues to /r/ and /l/ are the longest and worst in environments in which the cues are the shortest (Logan *et al*., 1991; Mochizuki, 1981; Sheldon and Strange, 1982).

However, Henly and Sheldon (1986) showed that duration is not the only cue that predicts the perceptual difficulty of different phonetic environments. They tested Japanese and Cantonese listeners on the /r/–/l/ contrast and found that Japanese listeners demonstrated their typical pattern of confusions across phonetic environments. In contrast, Cantonese listeners had the most difficulty with /r/ and /l/ in final singleton position and initial consonant clusters. Thus, the phonological system of the listener's native language also appears to play an important role in determining perception of non-native contrasts in different phonetic environments (Best, 1993; Flege, 1989b; Flege and Wang, 1989; Polka, 1991, 1992). According to Henly and Sheldon's argument, Japanese listeners have difficulty with initial consonant clusters because of the phonotactic constraints in the native language: Japanese does not generally allow consonant clusters (Mann, 1986).

In addition to durational cues and phonotactic constraints, coarticulatory effects also play a role in determining the difficulty non-native listeners have with different phonetic environments. Sheldon and Strange (1982) argued that /r/ and /l/ are coarticulated with the preceding consonant in initial consonant clusters. As a result, target steady-state values of the critical third formant may not be present. If subjects are attempting to match phonetic segments to some type of context-invariant template, coarticulated segments may provide a poor match to this template in memory. A similar argument can be made for the advantage observed for targets in final singleton position and final consonant clusters. In this case, the postvocalic /r/'s and /l/'s color the preceding vowel and provide a more robust and redundant set of cues to aid identification. In short, a combination of these factors may be responsible for the observed pattern of results. Explanations of the pattern of performance based on only one factor cannot provide an account of the effects of phonetic environment on perception. In the future, the relative contributions of phonotactic constraints, segment durations and coarticulatory cues should all be examined systematically to determine the locus of the effects of phonetic environment.

II. EXPERIMENT 2: EFFECTS OF TRAINING WITH A SINGLE TALKER

Another important source of information that needs to be addressed is talker variability, which refers to the number and nature of speakers used during training and generalization. Many of the previous cross-language training studies have used synthetic speech stimuli because the critical acoustic cues to non-native contrasts could be specified precisely and manipulated systematically. For example, Pisoni *et al.* (1982) and McClaskey *et al.* (1983) successfully trained English listeners to identify synthetic stop consonants that varied in voice onset time (VOT) into three perceptual categories. Strange and Dittmann (1984) trained Japanese listeners with a synthetic "rock"–"lock" series that varied in the third formant transition of the /r/ and /l/ segment and found that subjects improved in their discrimination of the synthetic training stimuli. However, while they found some generalization of the training to a new synthetic continuum ("rake"–"lake"), they failed to observe any generalization whatsoever to tokens of /r/ and /l/ in naturally produced English words.

The problem with using synthetic speech during the training phase is exemplified by Strange and Dittmann's subjects' failure to generalize to the natural speech tokens. Synthetic speech is an impoverished acoustic signal that lacks the rich redundancy of acoustic cues found in natural speech (see Pisoni *et al.*, 1985). Although experimenters may be able to isolate and manipulate the contrastive cues used by native speakers (see, however, Yamada and Tohkura, 1991, 1992), listeners in these experiments were not exposed to stimuli that reflect the diversity of variability found in natural speech. Consequently, training with synthetic speech may not be very effective when listeners are required during a generalization test to identify more natural stimuli that contain a great deal of superfluous variability.

Logan *et al.* (1991) addressed the problem of using synthetic speech stimuli in training by presenting listeners with a large number of natural speech stimuli. We assumed that differences in talkers' glottal waveforms, vocal tracts, dialects, and natural speaking rates would be important to establishing perceptual constancy over a wide range of stimulus variability (Kuhl, 1983). In our earlier experiment, listeners were trained with five different talkers and heard new talkers during the pretest and posttest and generalization. This stimulus set provided listeners with a very large number of acoustic cues to the /r/–/l/ contrast. The success of the high stimulus variability training paradigm was shown by the significant increases in performance from the pretest to the posttest and during training. Moreover, robust category acquisition was displayed by the finding that listeners performed only marginally better with a familiar talker than with an unfamiliar talker during the final tests of generalization.

We have argued recently that in the course of perceptual learning listeners developed highly context-dependent representations that contain information about the voice of the talker producing the contrast instead of abstract, invariant prototypes (Lively *et al.*, 1991). Recall and recognition memory studies with native speakers of English have also provided converging evidence for the retention of talker-specific details (Goldinger, 1992; Goldinger *et al.*, 1991; Palmeri *et al.*, 1993; Schacter and Church, 1992). Indeed, Mullennix and Pisoni (1990) showed that voice information is not filtered out or lost by listeners in a Garner selective attention task (Garner, 1974). These findings suggest that voice information and the phonetic forms of spoken words are perceived integrally and stored in memory in some composite representation.

The findings concerning talker variability have several important implications for our understanding of listeners' abilities to generalize to new talkers after training. When listeners are trained with many talkers, they may generalize from a large collection of tokens stored in memory. However, if listeners are trained with only a single talker, then only a small set of training exemplars can be brought to bear on the generalization task. Previous work in categorization and memory suggests that such high variability conditions will lead to robust generalization, whereas low variability conditions will not (Posner and Keele, 1968). Although listeners may improve in their ability to identify /r/ and /l/ from the pretest to the posttest and during training due to memory for specific test items, they should not generalize very well to new tokens because of the reduction in stimulus variability. This hypothesis was examined directly in experiment 2 by training a group of Japanese listeners with only a single talker who produced tokens contrasting /r/ and /l/ in five phonetic environments. Generalization was then tested with new tokens produced by the talker used in training and novel tokens produced by an unfamiliar talker. According to the present hypothesis, if talker variability is a critical factor in determining subjects' robust generalization to new tokens produced by novel talkers, then performance on the generalization tests should be poor. Listeners should be significantly faster and more accurate in the tests of generalization with the familiar talker than with the unfamiliar talker.

A. Method

1. Subjects—Six additional native speakers of Japanese were recruited for experiment 2. Subjects were obtained from the same source and met the same requirements as the subjects used in experiment 1.

2. Stimulus materials—Stimulus materials for the pretest and posttest and the tests of generalization were identical to those used in experiment 1. However, the stimuli presented during training were all produced by talker 4. In addition to the 42 minimal pairs produced by this talker in experiment 1, 26 additional minimal pairs were collected for use in training. Fifteen of these pairs contrasted /r/ and /l/ in final singleton position and the remaining pairs contrasted in final consonant clusters. Subjects were trained with a total of 136 stimuli in experiment 2. The new stimuli were prepared in the same manner as in experiment 1.

3. Procedure—The procedure for experiment 2 was identical to that used in experiment 1. Subjects participated in a two-alternative forced choice identification task in the pretest–posttest phase, training, and tests of generalization. Feedback was given only during the training phase. The only modification to the procedure occurred in the training regimen. Subjects were exposed only to the stimulus tokens produced by talker 4 during each of the 15 days of training. Each token was presented twice during every training session, for a total of 272 trials per session. Sessions lasted approximately 30 min. As in the previous experiment, accuracy and response latencies were collected during all phases of the experiment.

B. Results

1. Pretest–Posttest—Mean accuracy and latency scores were submitted to separate ANOVAs. Lexical status, phonetic environment, and pretest-versus-posttest were treated as within-subjects variables. Tukey's HSD procedure was used to analyze significant effects. Data from one subject were lost because of equipment failure. Thus, the final analyses reported in this section reflect data obtained from five subjects.

The top panel of Fig. 5 displays accuracy in the pretest and posttest as a function of phonetic environment. Although the main effect for test (pretest versus posttest) failed to reach significance in the analysis of the accuracy data [$F(1,4) =4.72$, $p < 0.1$], a significant interaction of test with phonetic environment revealed that subjects improved in their ability to identify /r/ and /l/ in some environments [$F(3,12) =4.67$, $p < 0.05$]. Accuracy of responses to contrasts in initial consonant clusters increased significantly from the pretest to the posttest. A nonsignificant improvement was also observed for contrasts in initial singleton position. The lower panel of Fig. 5 shows response latencies in the pretest and posttest as a function of phonetic environment. Both the main effect for pretest-versus-posttest [$F(1,4) = 381.5$, $p < 0.01$] and the interaction of test with phonetic environment [$F(3,12) = 5.92$, $p < 0.05$] were significant. Mean response times decreased significantly from the pretest to the posttest in all phonetic environments. The largest improvement was observed for contrasts in initial consonant clusters.

The main effect for phonetic environment was also significant in the analysis of the accuracy data [$F(3,12) = 9.94$, $p < 0.01$]. This result reflects the finding that responses to contrasts in final singleton position were more accurate than responses in any other phonetic environment. Finally, we failed to find any significant effects of lexical status, either alone or in conjunction with any other variable.

2. Training—Mean accuracy and latency scores from each day of training were submitted to separate ANOVAs. Week of training and phonetic environment were within-subject variables in each analysis.

Changes in performance during training were demonstrated by the main effect for week in the accuracy and response latency analyses [$F_{pc}(2,58) = 97.11$, $p < 0.01$; $F_{rt}(2,58) = 37.32$, $p < 0.01$]. As expected, accuracy increased and latency decreased significantly from week 1 to week 2. No significant increase in accuracy or decrease in latency was observed from week 2 to week 3.

The success of training was localized to improvements observed in initial singleton and intervocalic positions and initial consonant clusters. This finding was confirmed by the significant interaction of week with phonetic environment in the analysis of the accuracy data [$F(8,232) = 10.82$, $p < 0.01$] and is displayed in the top panel of Fig. 6. Accuracy was initially at ceiling levels for contrasts in final singleton position so no changes in performance could be observed for this environment during training. However, significant increases were obtained from week 1 to week 2 for contrasts in initial singleton, initial consonant clusters, intervocalic positions, and final consonant clusters. Subjects' accuracy increased significantly from week 2 to week 3 for initial consonant clusters. The lower panel of Fig. 6 shows a similar interaction in the response time analyses [$F(8,232) = 6.83$, $p < 0.01$]. Mean response times to targets in all phonetic environments decreased significantly during each week of training. Larger decreases were observed for contrasts in initial consonant clusters than for contrasts in final singleton position.

The main effect for phonetic environment was also significant in the analyses of the accuracy and latency data [$F_{pc}(4,116) = 53.59$, $p < 0.01$; $F_{rt}(4,116) = 34.39$, $p < 0.01$]. Responses were most accurate when the /r/–/l/ contrast was in word final clusters or final singleton position. Subjects responded faster to /r/ and /l/ in initial singleton position than to targets in word final clusters. No significant differences in performance were observed among any of the remaining phonetic environments. This pattern of results suggests a speed–accuracy trade-off: Fast responses were made for contrasts in difficult phonetic environments while slower, more deliberate responses were made to contrasts in easier environments.

3. Tests of generalization—Mean accuracy and latency scores from each test of generalization were submitted to separate ANOVAs. Latency means were log-transformed prior to analysis to compensate for the large differences in the variability between the two talker. Talker and phonetic environment were treated as within-subject variables in each analysis.

The results of the tests of generalization clearly revealed the limitations of the single-talker training procedures. As Fig. 7 shows, differences in accuracy between talkers were not equivalent across phonetic environments. This observation was confirmed by a significant talker by environment interaction [$F(4,20) = 6.12$, $p < 0.01$]. Subjects identified /r/'s and /l/'s in initial singleton and intervocalic positions more accurately when they were produced by a familiar talker. Responses were marginally faster for the old talker [old talker: 1045 ms; new talker: 1275 ms; $F(1,5) = 4.09$, $p < 0.1$].

Although these results show that subjects failed to generalize to new tokens produced by a new talker, the outcome of the test of generalization also indicated that listeners failed to generalize to new tokens produced by the old talker used in training. Mean accuracy on the test of generalization with the familiar talker was 83%. This was the same level of performance obtained during the first week of training. The pattern of response times was also similar. Mean response time in the test of generalization with the familiar talker was 1045 ms. This falls midway between the mean response times obtained in weeks 1 and 2 of training, 1167 and 942 ms, respectively.[3] This observation suggests that subjects developed highly detailed representations for the training stimuli, but that this knowledge did not transfer very well to new tokens.

In addition to the differences observed in performance as a function of talker, a main effect for phonetic environment was also obtained in both the accuracy and the latency data [$F_{pc}(4,20) = 6.12$, $p < 0.01$; $F_{rt}(4,20) = 7.59$, $p < 0.01$]. These findings reflect the fact that

[3] The results contrast with those obtained in experiment 1. Recall that in experiment 1 we observed performance during generalization that was comparable to performance during week 3 of training.

responses were most accurate to targets in final position and final consonant clusters, whereas the response latencies were fastest to contrasts in initial position.

C. Discussion

In this experiment, Japanese listeners were trained to identify /r/ and /l/ with stimulus tokens produced by a single talker. The results showed that listeners improved from the pretest to the posttest and during training in some phonetic environments. However, performance during the tests of generalization was not as good with an unfamiliar talker as it was with a familiar talker in the most difficult phonetic environments. Moreover, generalization to novel words produced by a familiar talker was mediocre, as shown by the finding that performance during generalization was no better than the performance observed during the first week of training. These findings suggest two conclusions. First, the improvements obtained during training reflect stimulus-specific learning, rather than robust abstract category acquisition. True category acquisition would be demonstrated by generalization to new talkers and new tokens over many different environments. Second, the presence of talker variability in the stimulus set during training appears to be an important condition for demonstrating robust generalization in this type of training paradigm.

Several important methodological and theoretical issues can be addressed by the present results. One methodological issue concerns the validity of our test of generalization. In the experiments reported here and in the earlier Logan *et al.* (1991) study, talker 4 was responded to fastest and most accurately. In both of these experiments, talker 4 was also used as the familiar talker during the test of generalization. However, responses to the unfamiliar talker were only marginally different from responses to the familiar talker in Logan *et al.*'s (1991) study and no difference was observed in experiment 1 of the present report. One might argue that the new talker is somehow inherently "good" or more intelligible to Japanese listeners and that our test is not a good measure of generalization. The results of experiment 2 provide strong evidence against this claim. If the unfamiliar talker was inherently good, then generalization performance in experiment 2 should not have been affected by the talker that was used in training. Good talkers should be responded to quickly and accurately, regardless of the composition of the stimulus set or prior training experience. We did not find this result. Instead, we suggest that our test of generalization does provide useful information about the effectiveness and generalizability of our training procedures. Moreover, the results demonstrate that the perceptual learning during training is talker-specific and that subjects retain detailed information about the stimulus ensemble in memory.[4]

Another interesting theoretical point concerns Goto's (1971) observation that Japanese listeners who are learning English become accustomed to a single talker and that they require additional training to adapt to new voices. Clearly, listeners in experiment 2 became very familiar with the talker used during training. In fact, the generalization data suggest that listeners became attuned to certain characteristics of that talker's voice and that these properties did not transfer very well to tokens produced by a new talker. More interesting, however, was the finding that training did not transfer well to new tokens produced by the familiar talker. These results demonstrate clearly that talker variability is an important factor in perceptual learning, generalization, and robust category acquisition in speech perception.

[4] In a pilot study carried out with untrained monolingual Japanese listeners, we found no difference in intelligibility between the familiar and unfamiliar talkers (Lively *et al.*, 1993). This finding supports our claim that differences in the composition of the training set, rather than absolute differences in intelligibility, are responsible for the pattern of generalization we observed.

III. GENERAL DISCUSSION

The experiments reported here were motivated by the pattern of results observed in our earlier training study. In that experiment, Japanese listeners were trained to identify English words containing /r/ and /l/ with highly variable stimulus sets. Both the voice of the talker and the phonetic environment of the contrast were varied during training. In the present investigation, the relative contributions of these two types of variability to generalization were studied separately. Our results are directly relevant to the theoretical construct of a phonetic prototype and to recent claims about the mental representations of categories in speech perception and spoken word recognition.

In experiment 1, Japanese listeners were trained with tokens that contrasted /r/ and /l/ only in the three most difficult phonetic environments (initial singleton and intervocalic positions and initial consonant clusters). Multiple talkers produced each of the training stimuli. Identification accuracy increased and response latency decreased from the pretest to the posttest and during the three weeks of training. However, significant differences were still observed among the phonetic environments. Factors such as durational differences (Dissosway-Huff *et al.*, 1982), coarticulatory effects (Sheldon and Strange, 1982), phonotactic constraints in the native language (Henly and Sheldon, 1986; Mann, 1986), and similarity to existing contrasts (Best, 1993; Polka, 1991, 1992) conspire to make these environments difficult for Japanese listeners.

Despite the fact that the amount of stimulus variability was reduced in the training set by eliminating contrasts in final singleton position and final consonant clusters, robust generalization was still observed in the first experiment. This finding suggests that the variability added by “easy” environments is not necessary for subjects’ subsequent generalization. Training with multiple talkers who produce only difficult contrasts appears to be a sufficient condition for generalization to new tokens produced by a familiar talker and to novel tokens produced by an unfamiliar talker. An important question for future research concerns how much variability in the difficult phonetic environments is critical for successful category acquisition. Training with a single token from each non-native category (Jamieson and Morosan, 1989) or a continuum of items from the same phonetic environment (Strange and Dittmann, 1984) appears to be insufficient to promote robust category formation. Thus, the minimal amount of environmental variability necessary for generalization remains to be determined.

The results of the tests of generalization obtained in experiment 2 differed from those found in the first experiment and our earlier study. In experiment 2, subjects were trained with tokens from a wider variety of phonetic environments. However, each of these tokens was produced by only a single talker. Subjects improved in their identification accuracy during training and from the pretest to the posttest. However, this improvement clearly reflected stimulus-specific learning because very little generalization was observed to new tokens produced by familiar and unfamiliar talkers. This pattern of results provides further evidence that talker variability is a critical factor in obtaining generalization in cross-language training studies (Goto, 1971).

The present findings parallel results reported a number of years ago by Posner and Keele (1968), who trained subjects to identify and categorize visually presented random dot patterns. In their classic categorization studies, subjects were presented with dot patterns that were either high or low statistical distortions of a prototype. Posner and Keele found that subjects who learned to classify the high variability patterns generalized better to new patterns than subjects trained with low variability patterns. Moreover, they found that the statistical prototype of the distortions was recognized and remembered more accurately than other distortions that had the same statistical relationships to the training patterns even when the prototype was never presented during training (Posner and Keele, 1970). Posner and Keele argued from these results that subjects stored an abstract prototype as well as specific training exemplars in memory (see, however, Hintzman, 1986).

A strong version of prototype theory suggests that a prototype for a non-native phonetic category could be acquired through abstracting a central tendency from the training stimuli (see Posner and Keele, 1968; Smith and Medin, 1981). If all stimuli contribute equally to the developing representation, the strong version of the prototype model makes three predictions that preclude it from accounting for the present data. First, an averaging model predicts that the developing representation would be biased toward contrasts in initial consonant clusters, given that the majority of the trials in experiment 1 were dedicated to this phonetic environment. Clearly, the results do not support this prediction. Large effects of phonetic environment were still observed despite the fact that all of the training occurred on the difficult phonetic environments.

Second, the averaging model predicts no differences among talkers because all stimuli are assumed to contribute equally to the prototype and all talkers were presented the same number of times during training. The results of experiment 1 do not support this prediction either. Reliable differences among talkers were observed in both the accuracy and latency analyses. Performance on tokens from talker 4 was consistently better than performance with any of the remaining training talkers.[5]

Finally, the prototype model predicts that generalization to novel words produced by the training talker in experiment 2 should be good, given that listeners were exposed to tokens produced by this talker. The results again do not support the averaging model: Generalization to novel tokens produced by the familiar talker was no better than performance during the first week of training.[6]

Recent findings in the categorization literature have suggested a viable alternative to the prototype view. Nosofsky (1986, 1987) and Kruschke (1992) have argued for exemplar-based models of categorization. In these models, subjects are assumed to store in memory multidimensional representations of objects presented during training. A selective attention mechanism weights the importance of various stimulus dimensions. The critical dimensions for category membership are given strong weights, while dimensions that are less important receive smaller weights. Changes in selective attention “stretch” and “shrink” the perceptual space for these dimensions and in turn alter the internal category structure: Objects become less similar to each other as dimensions are stretched and more similar to each other as dimensions are shrunk. According to this view, categorization of new items occurs in the context of items stored in memory. When a new item is presented, it is compared on a dimension-by-dimension basis to stored items in memory. Items that are very similar to stored items along critical dimensions are likely to be accepted as category members, whereas items that are low in similarity are likely to be rejected.

[5] The pattern of performance observed across different talkers in this study and the results of the earlier Logan *et al*. experiment have recently been replicated in a collaborative study at the ATR Laboratory in Japan using 19 monolingual Japanese subjects (see Lively *et al.*, 1993).

[6] The averaging model could be modified by assuming that stimuli contribute to a weighted average via a selective attention mechanism. In this case, factors such as native-language phonotactic constraints might contribute to the differential weighting of stimuli. For example, Japanese listeners might weight contrasts in initial consonant clusters less heavily than contrasts in final singleton position because of the phonological rules of their native language. Although recourse to the properties of the native language provides a principled extension to the model and may account for the effects of phonetic environment, it is difficult to understand how the model could be extended to account for the highly selective effects of talker.

Two issues need to be addressed by the exemplar-based selective attention framework with regard to non-native phonetic category acquisition. First, what information is preserved in memory? Second, how is the selective attention mechanism drawn to the relevant contrastive stimulus dimensions and how it is influenced by sound contrasts of the native language?

From the results presented in this report and our previous investigation, it appears that listeners encode talker-specific information in memory. Support for this claim may be seen in the generalization results of experiments 1 and 2. One consequence of the high variability training in experiment 1 was that representations for the new phonetic categories were stretched according to the parameters of many voices. As a result, subjects had a relatively unconstrained set of exemplars from which to generalize. Listeners in experiment 2, in contrast, were trained with a highly constrained stimulus set and as a consequence showed poor generalization to a new talker. It appears that training with a single talker was a relatively ineffective means for stretching listeners' perceptual spaces for non-native contrasts. Rather, subjects engaged in stimulus-specific category learning.

Several other findings also support the assumption that listeners encode voice-specific information in memory. Goldinger *et al.* (1991) found that listeners are able to use voice-specific information as a retrieval cue in a serial recall task when stimulus items are presented at long inter-stimulus intervals. More recently, using both implicit and explicit memory tests Goldinger (1992) showed that highly detailed, talker-specific representations are developed for spoken words and that this information can be exploited during word recognition. As mentioned earlier, Goto (1971) observed that Japanese listeners become attuned to particular voices and that more training is required for generalization to new voices. Finally, Flege (1989a) has also suggested that information about talkers' voices is encoded into memory as part of the developing representation for non-native phonetic categories.

Another important theoretical issue concerns aspects of the selective attention mechanism. First, the level at which the selective attention mechanism operates needs to be specified in much greater detail. Second, the interaction between the selective attention mechanism and the listeners' native language must be considered. Recently, these issues have been addressed in some detail by Jusczyk (1989, 1993a,b), who has proposed a model for the acquisition of phonetic categories and lexical items (WRAPSA) by infants that uses a selective attention mechanism similar to the one employed in Nosofsky's (1986, 1987) model. In Jusczyk's model, the selective attention mechanism operates at a perceptual level. He argues that acoustic dimensions are extracted from the incoming signal by a bank of auditory processing filters (Sawusch, 1986) and that these dimensions are weighted automatically in terms of their importance by the selective attention mechanism. Weights for each dimension are determined by the degree to which they signal contrastive meanings. These weights are then used to generate a precompiled interpretive scheme that can be applied automatically to incoming fluent speech (Jusczyk, 1989, 1993a; Klatt, 1979).

A similar process may occur when Japanese listeners are trained to identify words containing English /r/ and /l/. Over the course of training, these weights are changed to facilitate development of new phonetic categories and in turn they cause a reorganization of listeners' phonological spaces or filters (Flege, 1989b; Terbeek, 1977). Alternatively, depending on the acoustic characteristics of the novel contrast and phonological system of the listener, the attention weights for contrasts in the native language may not change during training. Rather, an unused portion of a general acoustic-phonetic space may be used for the new phonetic categories (Best, 1993).

The difficulty of retuning the selective attention mechanism and acquiring new contrasts appears to be mediated by several factors. First, the salience of the contrastive cues must be considered (Underbakke *et al.*, 1988). Strange and Jenkins (1978) have argued that contrasts based on temporal cues are easier to learn than contrasts based on spectral cues. Thus, acquiring the /r/–/l/ contrast may be more difficult than acquiring VOT contrasts. This assumption is supported by comparing a number of cross-language training studies. Pisoni *et al.* (1982) and McClaskey *et al.* (1983) showed that English listeners could acquire a third VOT category in a single training session. Japanese listeners, in contrast, appear to require several weeks of training to acquire the /r/–/l/ contrast.

Second, Best (1993) has argued that the acquisition of new phonetic contrasts is dependent upon the degree to which the novel contrast is similar to contrasts in the native language. Non-native contrasts that are highly similar to those in the native language may be easily acquired (Flege, 1990). According to a selective attention model, only small adjustments in the perceptual weighting scheme are required for the acquisition of these types of contrasts. Contrasts that are allophonic in the native language, but contrastive in the second language, may be more difficult to acquire, particularly if each member of the contrast is an equally good member of the native language phonetic category (Best, 1993; Flege, 1990).

Third, the phonological constraints of the native language also play a role in determining how difficult a new contrast will be to acquire (Flege, 1989a; Flege and Wang, 1989; Henly and Sheldon, 1986). The difficulty that Japanese listeners have with /r/ and /l/ in initial consonant clusters supports this conclusion. Retuning the selective attention mechanism to detect and recognize a contrast in an environment governed by a different set of phonotactic constraints may present special difficulties for nonnative listeners. Thus, predicting the difficulty of acquiring new phonetic contrasts will be dependent on the constraints of the native language and on the similarity of the new contrast to preexisting categories in the listeners' first language (Best, 1993; Best and Strange, 1992; Flege, 1990; Polka, 1991, 1992).

In summary, Japanese listeners were trained to identify English /r/ and /l/ under several conditions. Performance was found to depend on the phonetic environment in which the /r/–/l/ contrast occurred and the voice of the talker producing the contrast. Generalization was also dependent upon the composition of the training set: Listeners trained with high variability stimulus sets generalized well to new tokens produced by a familiar talker and to novel tokens produced by an unfamiliar talker. In contrast, listeners trained with only a single talker showed little evidence of generalization to new tokens or new talkers. Taken together with our earlier findings, the present results are inconsistent with predictions from a simple prototype model of categorization. In contrast, an exemplar-based model with a selective attention mechanism (Jusczyk, 1989, 1993a,b; Nosofsky, 1986, 1987) appears to be able to account for the findings of the present study and to accommodate a wide variety of other results related to the role of stimulus variability in speech perception and spoken word recognition.

Acknowledgments

This research was supported, in part, by NIH Research Grant No. DC-00111-15 and, in part by NIDCD Research Training Grant No. DC00012-14 to Indiana University in Bloomington, IN. The authors wish to thank Terrance Neary, Patrice Beddor, and an anonymous reviewer for their helpful comments on a previous version of this manuscript.

References

Atkinson RC. Optimizing the learning of a second language vocabulary. J. Exp. Psychol. 1972; 96:124–129.

Best, CT. The emergence of language-specific influences in infant speech perception. In: Nusbaum, HC.; Goodman, J., editors. Development of Speech Perception: The Transition from Recognizing Speech Sounds to Spoken Words. Cambridge, MA: MIT; 1993.

Best CT, Strange W. Effects of phonological and phonetic factors on cross-language perception of approximants. J. Phonet. 1992; 20:305–320.

Carney A, Widin G, Viemeister N. Noncategorical perception of stop consonants varying in VOT. J. Acoust. Soc. Am. 1977; 62:961–970. [PubMed: 908791]

Dissosway-Huff, P.; Port, R.; Pisoni, DB. Research on Speech Perception Progress Report No. 8. Bloomington, IN: Speech Research Laboratory, Indiana University; 1982. Context effects in the perception of /r/ and /l/ by Japanese.

Flege, J. The production and perception of foreign language speech sounds. In: Winitz, H., editor. Human Communication and Its Disorders: A Review. Norwood, NJ: Ablex; 1989a. p. 224-401.

Flege J. Chinese subjects' perception of the word-final English /t/–/d/ contrast: Performance before and after training. J. Acoust. Soc. Am. 1989b; 86:1684–1697. [PubMed: 2808918]

Flege, J. Perception and production: The relevance of phonetic input to L2 phonological learning. In: Ferguson, C.; Huebner, T., editors. Crosscurrents in Second Language Acquisition and Linguistic Theories. Philadelphia, PA: John Benjamins; 1990.

Flege J, Wang C. Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/–/d/ contrast. J. Phonet. 1989; 17:299–315.

Garner, W. The Processing of Information and Structure. Potomac, MD: LEA; 1974.

Gillette S. Contextual variation in the perception of L and R by Japanese and Korean speakers. Minn. Papers Ling. Philos. Lang. 1980; 6:59–72.

Goldinger, SD. Research on Speech Perception Technical Report No. 7. Bloomington, IN: Speech Research Laboratory, Indiana University; 1992. Words and voices: Implicit and explicit memory for spoken words.

Goldinger SD, Pisoni DB, Logan JS. On the locus of talker variability effects in recall of spoken word lists. J. Exp. Psychol. Learn. Mem. Cogn. 1991; 17:152–162. [PubMed: 1826729]

Goto H. Auditory perception by normal Japanese adults of the sounds 'L' and 'R'. Neuropsychol. 1971; 9:317–323.

Henly E, Sheldon A. Duration and context effects on the perception of English /r/ and /l/: A comparison of Cantonese and Japanese speakers. Lang. Learn. 1986; 36:505–521.

Hintzman DL. Schema abstraction in a multiple trace memory model. Psychol. Rev. 1986; 93:411–428.

Jamieson D, Morosan D. Training non-native speech contrasts in adults: Acquisition of the English /θ/–/ð/ contrast by francophones. Percept. Psychophys. 1986; 40:205–215. [PubMed: 3580034]

Jamieson D, Morosan D. Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. Can. J. Psychol. 1989; 43:88–96. [PubMed: 2819599]

Jusczyk, P. Developing phonological categories from the speech signal. paper presented at The International Conference on Phonological Development, Stanford University; 1989.

Jusczyk, P. Infant speech perception and the development of the mental lexicon. In: Nusbaum, HC.; Goodman, JC., editors. The Transition from Speech Sounds to Spoken Words: The Development of Speech Perception. Cambridge, MA: MIT; 1993a.

Jusczyk P. From general to language-specific capacities: The WRAPSA model of how speech perception develops. J. Phonet. 1993b; 21:3–28.

Klatt DH. Speech perception: A model of acoustic-phonetic analysis and lexical access. J. Phonet. 1979; 7:279–312.

Kuhl PK. Perception of auditory equivalence classes for speech in early infancy. Infant Behav. Dev. 1983; 6:263–285.

Kuhl PK. Human adults and human infants show a 'perceptual magnet effect' for the prototype of speech categories, monkeys do not. Percept. Psychophys. 1991a; 50:93–107. [PubMed: 1945741]

Kuhl, PK. Speech prototypes: Studies on the nature, function, ontogeny and phylogeny of the 'centers' of speech categories. In: Tohkura, Y.; Vatikiotis-Bateson, E.; Sagisaka, Y., editors. Speech Perception, Production and Linguistic Structure. Tokyo: OHM; 1991b. p. 239-264.

Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B. Linguistic experience alters phonetic perception in infants by 6 months of age. Science. 1992; 255:606–608. [PubMed: 1736364]

Kruschke J. ALCOVE: An exemplar-based connectionist model of category learning. Psychol. Rev. 1992; 90:22–44. [PubMed: 1546117]

Lane H. The motor theory of speech perception: A critical review. Psychol. Rev. 1965; 7:275–309. [PubMed: 14348425]

Lane, H. A behavioral basis for the polarity principle in linguistics. In: Salzinger, K.; Salzinger, S., editors. Research in Verbal Behavior and some Neurological Implications. New York: Academic; 1969. p. 79-98.

Lehiste I. Acoustic characteristics of selected English consonants. Int. J. Am. Ling. 1964; 30:10–115.

Liberman AM, Harris KS, Kinney JA, Lane HL. The discrimination of relative onset-time of the components of certain speech and non-speech patterns. J. Exp. Psychol. 1961; 61:379–388. [PubMed: 13761868]

Lively, SE.; Pisoni, DB.; Logan, JS. Some effects of training Japanese listeners to identify English /r/ and /l/. In: Tohkura, Y.; Vatikiotis-Bateson, E.; Sagisaka, Y., editors. Speech Perception, Production and Linguistic Structure. Tokyo: OHM; 1991. p. 175-196.

Lively, SE.; Pisoni, DB.; Yamada, RA.; Tohkura, Y.; Yamada, T. Research on Speech Perception Progress Report No. 18. Bloomington, IN: Speech Research Laboratory, Indiana University; 1993. Training Japanese listeners to identify English /r/ & /l/: III. Listeners show long-term retention of new phonetic contrasts.

Logan JS, Lively SE, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/: A first report. J. Acoust. Soc. Am. 1991; 89:874–886. [PubMed: 2016438]

Mann V. Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English 'l' and 'r'. Cognition. 1986; 24:169–196. [PubMed: 3816123]

McClaskey C, Pisoni D, Carrell T. Transfer of training to a new linguistic contrast in voicing. Percept. Psychophys. 1983; 34:323–330. [PubMed: 6657433]

Miller JL. Properties of feature detectors for VOT: The voiceless channel of analysis. J. Acoust. Soc. Am. 1977; 62:641–648. [PubMed: 903506]

Miller JL, Connine CM, Schermer TM, Kluender KR. A possible auditory basis for internal structure of phonetic categories. J. Acoust. Soc. Am. 1983; 73:2124–2133. [PubMed: 6875098]

Miller JL, Volaitis LE. Effect of speaking rate on the perceptual structure of a phonetic category. Percept. Psychophys. 1989; 46:505–512. [PubMed: 2587179]

Mochizuki M. The identification of /r/ and /l/ in natural and synthesized speech. J. Phonet. 1981; 9:283–303.

Mullennix J, Pisoni D. Stimulus variability and processing dependencies in speech perception. Percept. Psychophys. 1990; 47:379–390. [PubMed: 2345691]

Nosofsky RM. Attention, similarity, and the identification-categorization relationship. J. Exp. Psychol. Gen. 1986; 115:39–57. [PubMed: 2937873]

Nosofsky RM. Attention and learning processes in the identification and categorization of integral stimuli. J. Exp. Psychol. Learn. Mem. Cogn. 1987; 15:700–708.

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. J. Exp. Psychol. Learn. Mem. Cogn. 1993; 19:309–328. [PubMed: 8454963]

Pisoni DB. Auditory and phonetic codes in the discrimination of consonants and vowels. Percept. Psychophys. 1973; 13:253–260.

Pisoni DB. Some comments on invariance, variability, and perceptual normalization in speech perception. Proc. ICSLP. 1992; 92:587–590.

Pisoni DB, Aslin RN, Perey AJ, Hennessy BL. Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. J. Exp. Psychol. Hum. Percept. Perform. 1982; 8:297–314. [PubMed: 6461723]

Pisoni D, Nusbaum H, Greene B. Perception of synthetic speech generated by rule. Proc. IEEE. 1985; 11:1665–1676.

Pisoni DB, Tash J. Reaction times to comparisons within and across phonetic categories. Percept. Psychophys. 1975; 15:285–290.

Polka L. Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. J. Acoust. Soc. Am. 1991; 89:2961–2977. [PubMed: 1918634]

Polka L. Characterizing the influence of native language experience on adult speech perception. Percept. Psychophys. 1992; 52:37–52. [PubMed: 1635856]

Posner M, Keele S. On the genesis of abstract ideas. J. Exp. Psychol. 1968; 77:353–363. [PubMed: 5665566]

Posner M, Keele S. Retention of abstract ideas. J. Exp. Psychol. 1970; 83:304–308.

Price, JP. unpublished doctoral dissertation. University of Pennsylvania; 1981. A cross-linguistic study of flaps in Japanese and in American English.

Repp BH. Dichotic competition of speech sounds: The role of acoustic stimulus structure. J. Exp. Psychol. Hum. Percept. Perform. 1976; 3:37–50.

Rosch E. Cognitive reference points. Cogn. Psychol. 1975a; 7:532–547.

Rosch E. The nature of mental codes for color categories. J. Exp. Psychol. Hum. Percept. Perform. 1975b; 1:303–322.

Rosch E, Mervis CB. Family resemblances: Studies in the internal structure of categories. Cogn. Psychol. 1975; 7:573–605.

Samuel AG. Phonetic prototypes. Percept. Psychophys. 1982; 31:307–314. [PubMed: 7110883]

Sawusch, JR. Auditory and phonetic coding of speech. In: Schwab, EC.; Nusbaum, HC., editors. Pattern Recognition by Humans and Machines: Vol. 1, Speech Perception. New York: Academic; 1986. p. 51-88.

Schacter DL, Church BA. Auditory priming: Implicit and explicit memory for words and voices. J. Exp. Psychol. Learn. Mem. Cogn. 1992; 18:915–930. [PubMed: 1402716]

Sheldon A, Strange W. The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. Appl. Psycholing. 1982; 3:243–261.

Smith, E.; Medin, D. Categories & Concepts. Cambridge, MA: Harvard U.P.; 1981.

Strange W, Dittmann S. Effects of discrimination training on the perception of /r/–/l/ by Japanese adults learning English. Percept. Psychophys. 1984; 36:131–145. [PubMed: 6514522]

Strange, W.; Jenkins, J. Role of linguistic experience in perception of speech. In: Walk, RD.; Pick, HL., editors. Perception and Experience. New York: Plenum; 1978. p. 125-169.

Terbeek, D. Working Papers in Phonetics. Vol. Vol. 37. Los Angeles: UCLA; 1977. A cross-language multidimensional scaling study of vowel perception.

Underbakke M, Polka L, Gottfried TL, Strange W. Trading relations in the perception of /r/–/l/ by Japanese learners of English. J. Acoust. Soc. Am. 1988; 84:90–100. [PubMed: 3411058]

Volaitis LE, Miller JL. Phonetic prototypes: Influence of place of articulation and speaking rate in the internal structure of voicing categories. J. Acoust. Soc. Am. 1992; 92:723–735. [PubMed: 1506527]

Werker J, Logan J. Cross-language evidence for three factors in speech perception. Percept. Psychophys. 1985; 37:35–44. [PubMed: 3991316]

Werker J, Tees R. Phonemic and phonetic factors in adult cross-language speech perception. J. Acoust. Soc. Am. 1984; 75:1866–1878. [PubMed: 6747097]

Yamada, RA.; Tohkura, Y. Perception of American English /r/ and /l/ by native speakers of Japanese. In: Tohkura, Y.; Vatikiotis-Bateson, E.; Sagisaka, Y., editors. Speech Perception, Production and Linguistic Structure. Tokyo: OHM; 1991. p. 155-174.

Yamada R, Tohkura Y. The effects of experimental variables in the perception of American English /r,l/ by Japanese listeners. Percept. Psychophys. 1992; 52:376–392. [PubMed: 1437471]

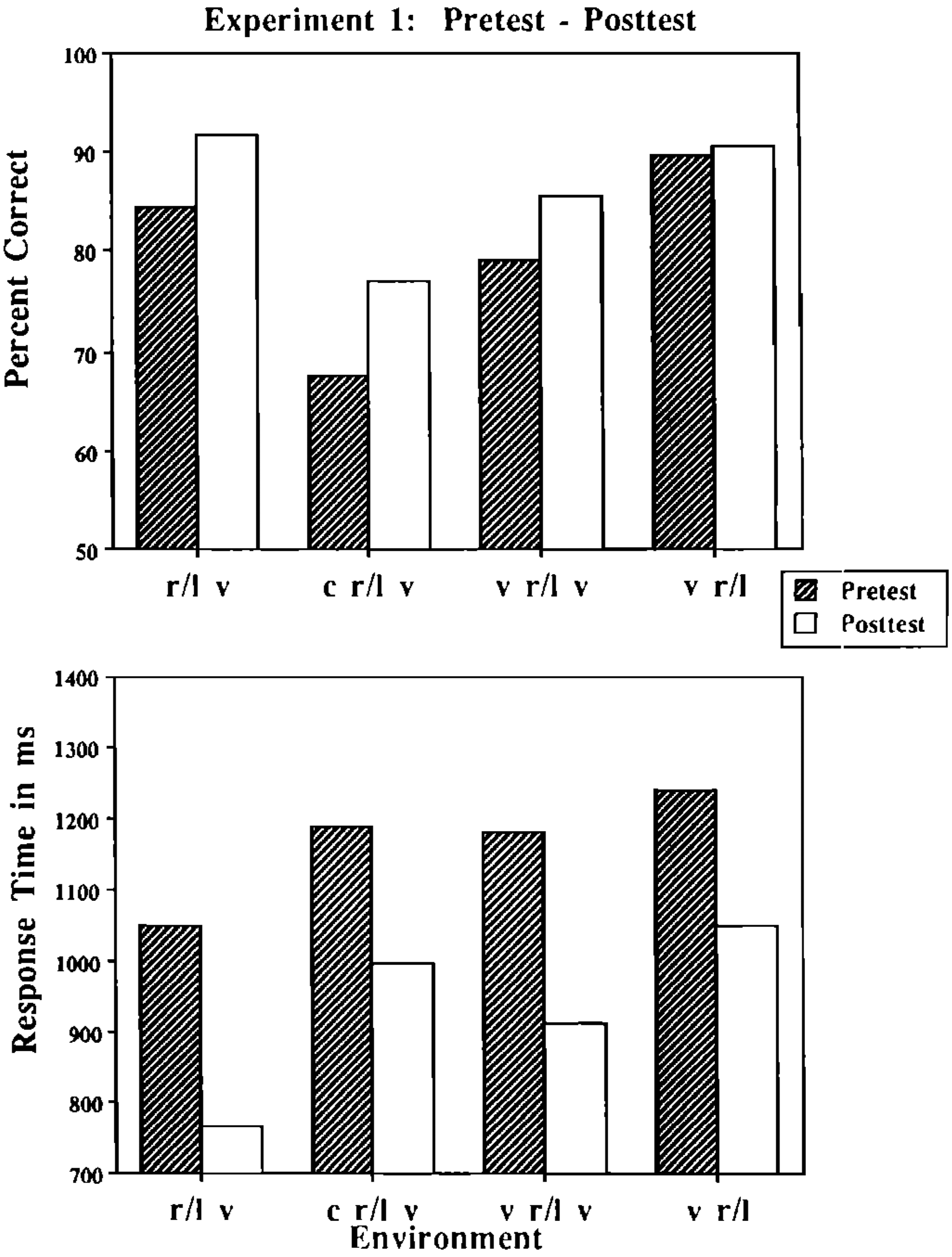

FIG. 1.

Upper panel, subjects' accuracy in experiment 1 in the pretest and posttest as a function of phonetic environment. Lower panel, response latencies.

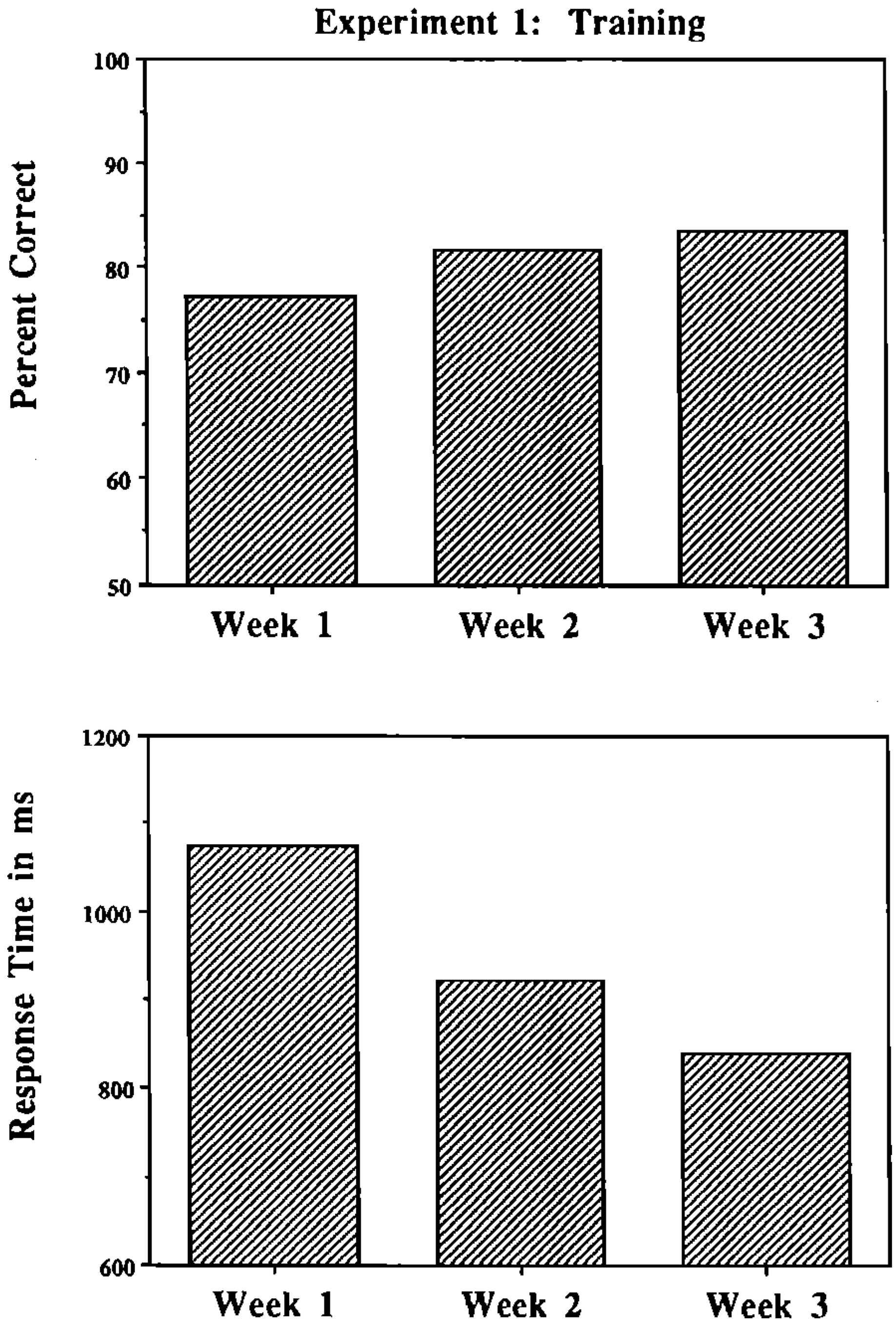

FIG. 2.

Upper panel, subjects' accuracy in experiment 1 during training as a function of week of training. Lower panel, response latencies.

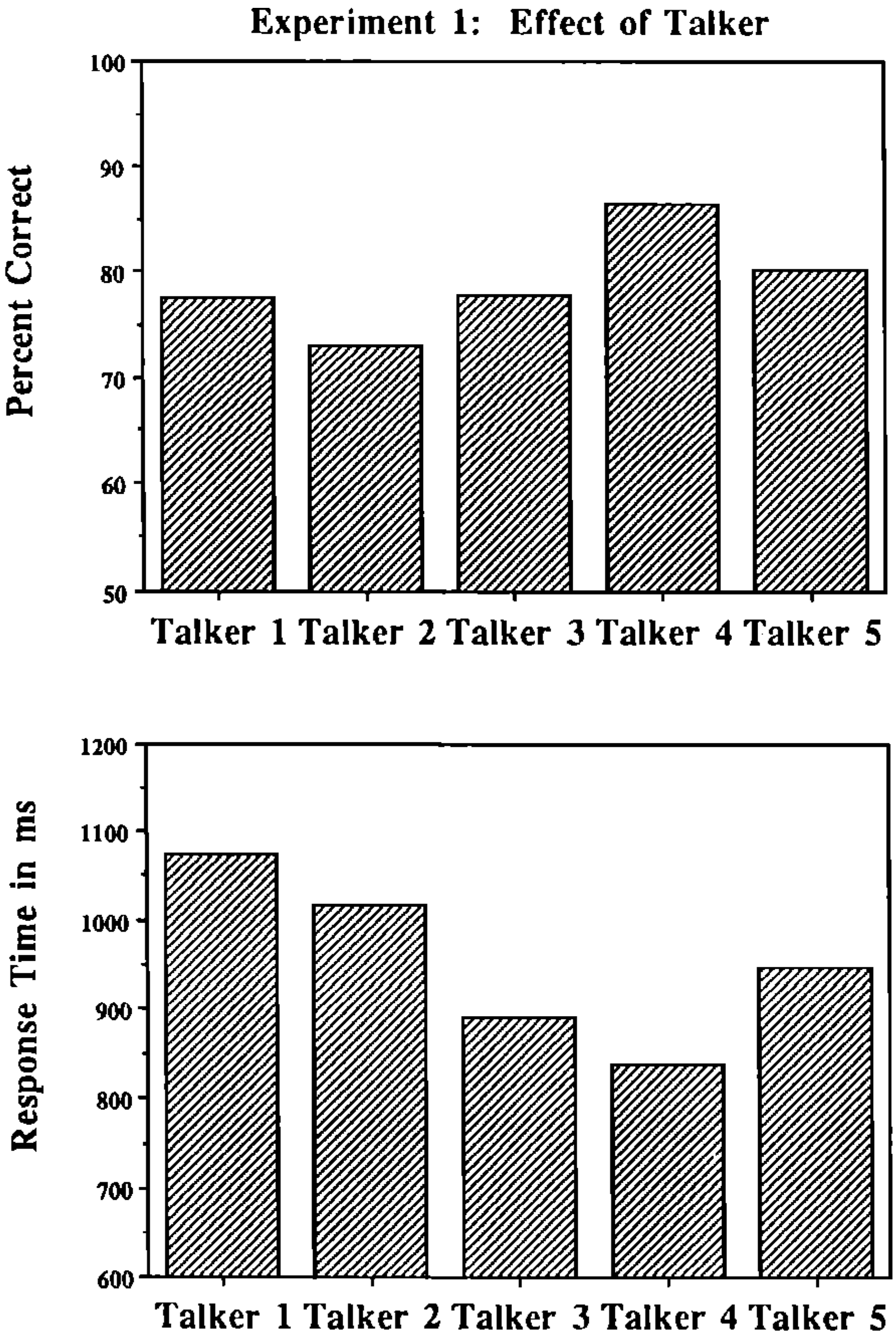

FIG. 3.

Upper panel, accuracy during training in experiment 1 as a function of talker. Lower panel, response latencies.

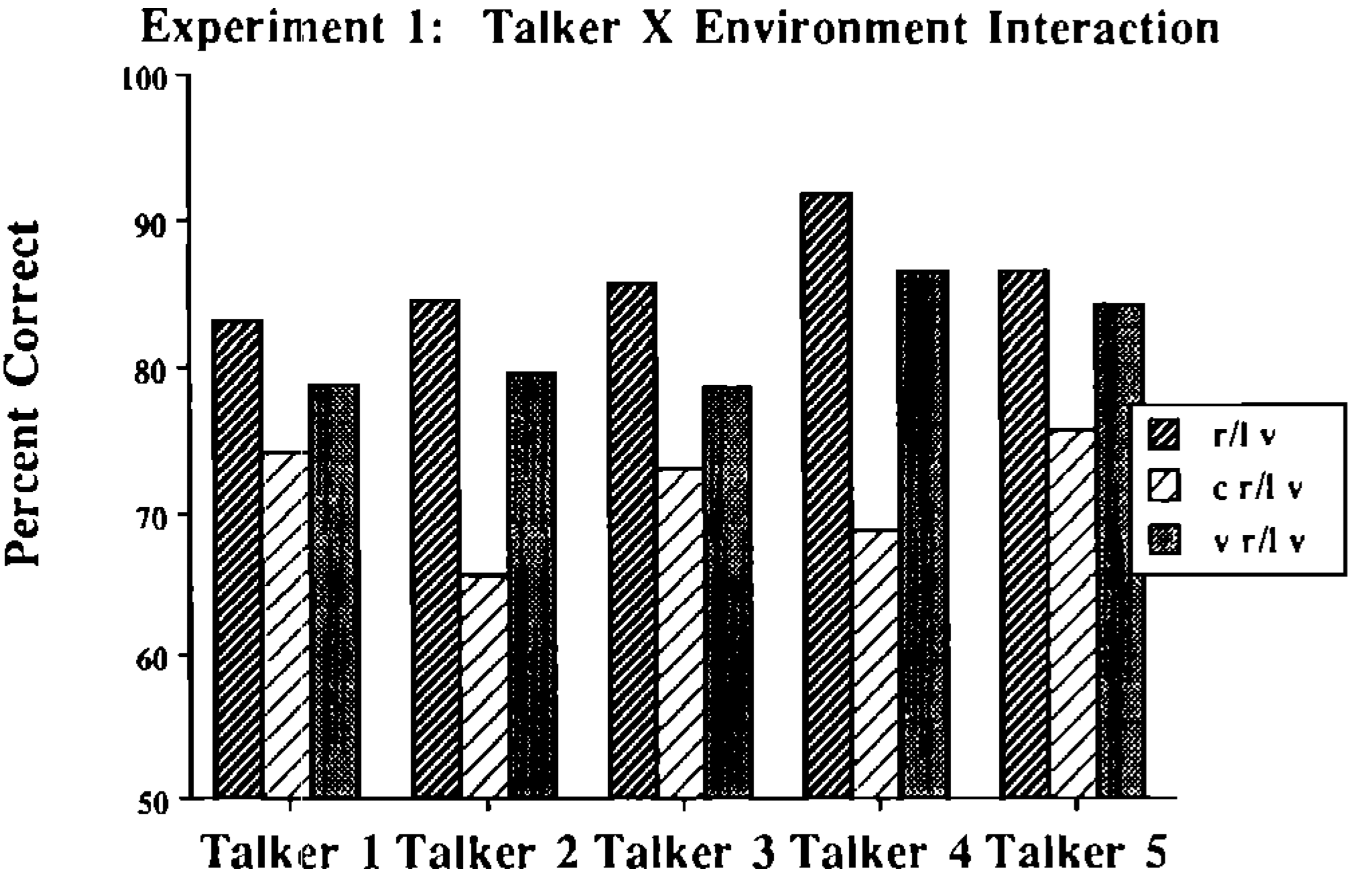

FIG. 4.
Interaction of talker and phonetic environment during the training phase of experiment 1.

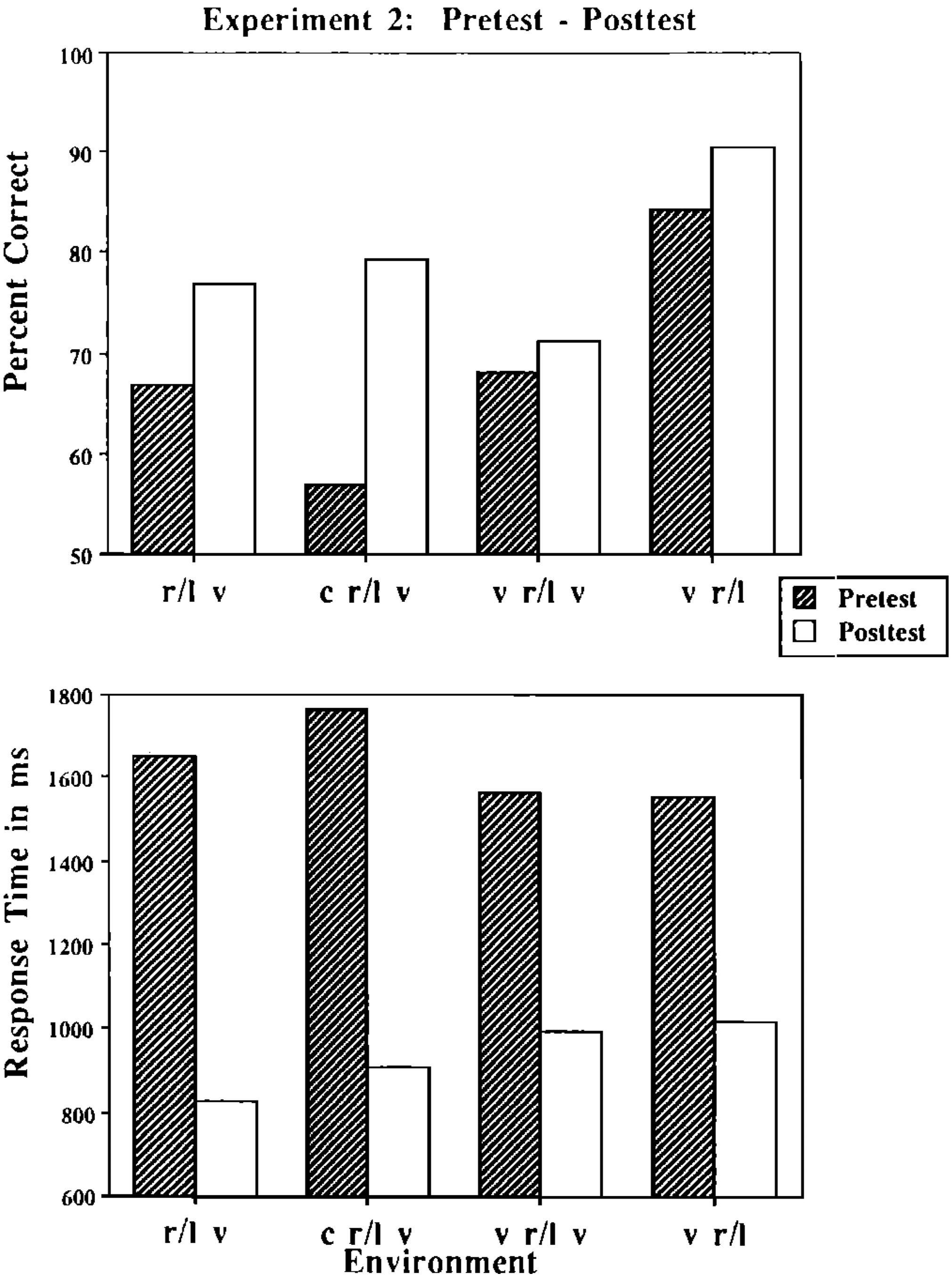

FIG. 5.

Upper panel, subjects' accuracy in experiment 2 in the pretest and posttest as a function of phonetic environment. Lower panel, response latencies.

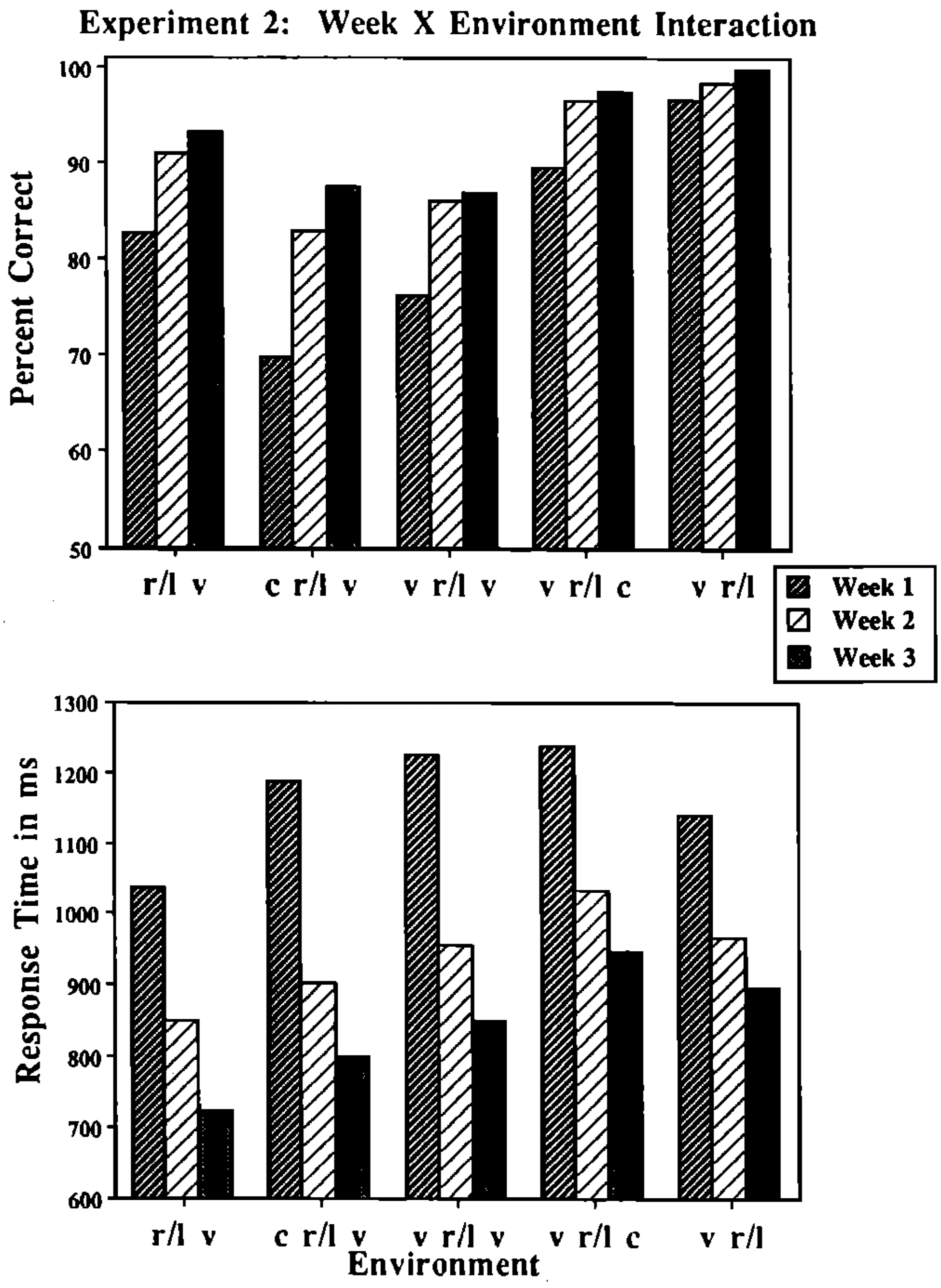

FIG. 6.

Upper panel, the interaction of week of training with phonetic environment during training in the accuracy data of experiment 2. Lower panel, a similar interaction in the response time data.

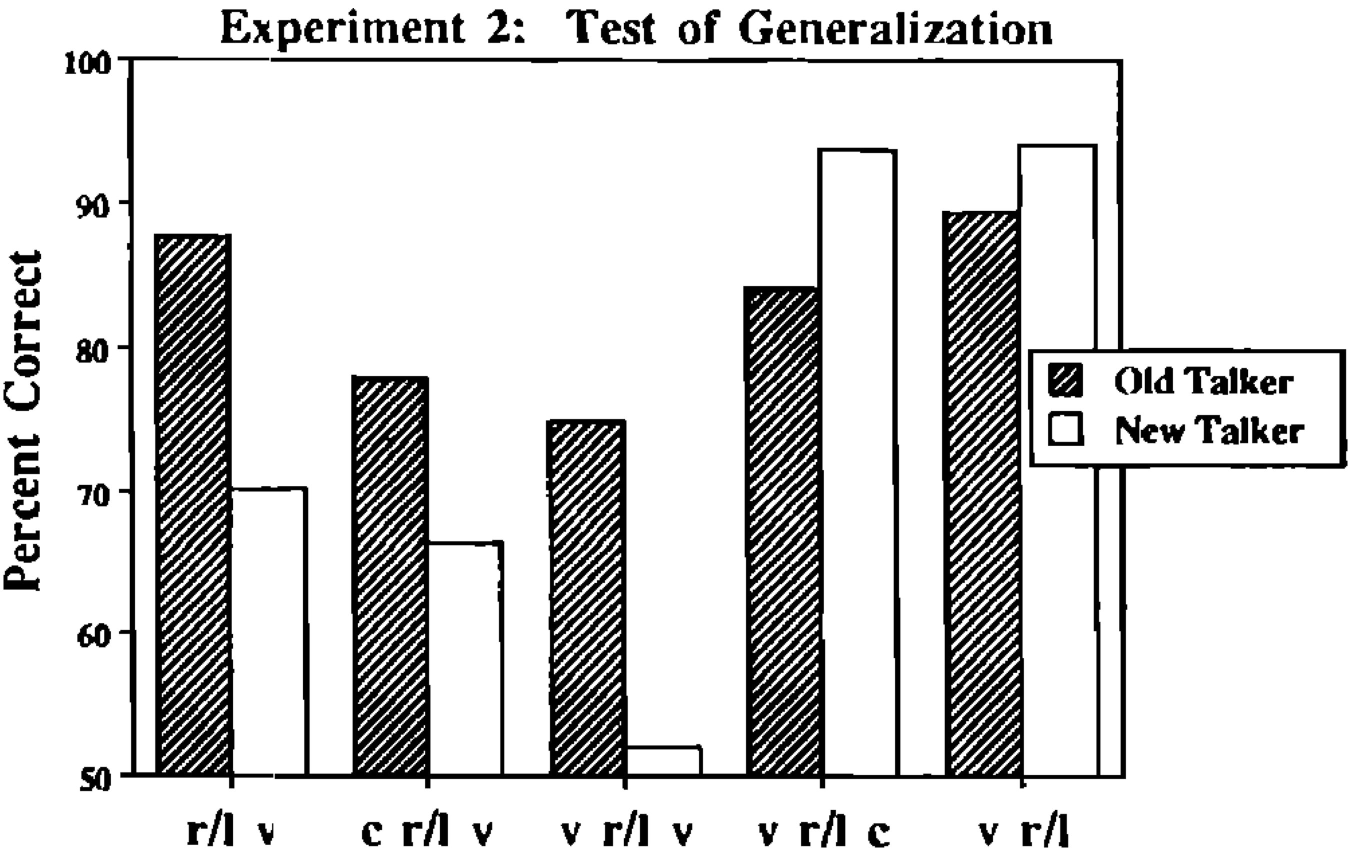

FIG. 7.

A significant interaction between talker and phonetic environment in the test of generalization from experiment 2.