Published in final edited form as:
J Acoust Soc Am. 1993 August ; 94(2 Pt 1): 1148–1151.

Training listeners to perceive novel phonetic categories: How do we know what is learned?

John S. Logan,
Department of Psychology, Carleton University, Ottawa, Ontario K1S 5B6, Canada

Scott E. Lively, and
Department of Psychology, Indiana University. Bloomington, Indiana 47405

David B. Pisoni
Department of Psychology, Indiana University. Bloomington, Indiana 47405

Abstract

Logan *et al.* [J. Acoust. Soc. Am. **89**, 874–886 (1991)] carried out a perceptual learning experiment in which Japanese listeners were trained to identify English words containing /r/ and /l/. Pruitt [J. Acoust. Soc. Am. **94**, 1146–1147 (1993)] has criticized several aspects of the procedures and conclusions. First, he argues that the lack of appropriate control groups make interpretation of the results problematic. Second, he asserts that the generalization test was methodologically flawed. Although Pruitt raises some important issues that are worth pursuing in future research, it is argued that the methodology that was employed and the conclusions that were drawn were valid.

I. THE EFFECT OF TRAINING ON PRETEST/ POSTTEST PERFORMANCE

Pruitt contends that the design of our experiment did not allow us to disentangle a number of factors that may have contributed to the success of our training regimen. We trained native Japanese listeners in an identification task and evaluated their performance using a pretest/ posttest design. During training, subjects were presented words which contained /r/ and /l/ in five different phonetic contexts. Each of the words was produced by five different talkers. In addition, extensive feedback was provided. Feedback not only included information about whether subjects' responses were correct or incorrect, but subjects were also re-presented any stimulus on which they made an error. Logan *et al.* demonstrated that such a training procedure produced a statistically reliable improvement in identification performance for naturally produced English words containing /r/ and /l/. This finding contrasts sharply with the result obtained by Strange and Dittmann (1984), who found very little improvement on the natural speech tokens that were used during the pretest-posttest phase of their experiment. Moreover, we also found that performance varied systematically as a function of phonetic environment and talker.

Pruitt (1993) takes issue with the fact that we did not systematically explore the relative contribution of each of these factors to the improved identification performance we observed. His complaint, however, neglects several important issues. First, it was never our intention in this initial study to examine every possible permutation of stimulus and procedural variables. Instead, our goal was to see if we could develop a laboratory-based training procedure that would facilitate the identification performance of Japanese listeners presented with naturally produced English words containing /r/ and /l/ (Logan *et al,* 1991, p. 874). Based on an analysis of previous efforts (Jamieson and Morosan, 1986, 1989; Strange and Dittmann, 1984), we determined that sufficiently variable stimuli combined with an appropriate training task were critical for the formation of robust new phonetic categories (cf. Posner and Keele, 1968). Our results convincingly demonstrated both of these points.

Second, we ourselves acknowledged that substantial work remained to determine the extent to which factors such as talker variability and phonetic variability contributed to the improved performance we obtained (see Logan *et al.,* 1991, p. 883). To this end, we recently have completed several experiments that examined the separate contribution of phonetic variability and talker variability (see Lively *et al.,* 1993a). In one experiment, we trained Japanese listeners with tokens from the three most difficult phonetic environments for Japanese listeners (initial singleton, initial consonant clusters, and intervocalic positions). Tokens containing /r/ and /l/ from these environments were produced by five different talkers. We replicated our original results. Improvements in identification performance were obtained from the pretest to the posttest and during training. In addition, generalization performance was not reliably different when listeners responded to tokens produced by an old talker compared to tokens produced by a new talker; performance with both generalization talkers was equivalent to performance at the conclusion of training. In a second experiment, we trained another group of listeners with a single talker who produced the /r/–/l/ contrast in all five phonetic environments. Improved performance in some phonetic environments was observed both in the posttest and during training. However, the limitation of this training strategy was revealed by the results of the tests of generalization. Subjects were significantly more accurate in identifying tokens from a familiar talker. However, their mean level of performance with the old talker was only equivalent to performance during weeks 1 and 2 of training. Overall, the results of these new training experiments suggest that talker variability plays a larger role than phonetic variability, thus addressing one of the major concerns raised by Pruitt in his criticism of our original study.

Pruitt suggests that the improvements in identification performance obtained by Logan *et al.* in the posttest could be due to "…subjects' aptitudes, the testing conditions, or the quality of the stimuli…" First, our subjects were not so homogeneous in terms of "aptitude" that this characteristic alone caused them to perform in an identical manner during the experiment. As shown in Table A II of Logan *et al.,* average pretest performance ranged from 58.3% to 95.8% correct, not at all what one would expect if initial performance was equivalent across subjects. Moreover, the fact that all of our subjects showed a consistent improvement from pretest to posttest can be taken as evidence that the effects of training were real and not due to characteristics specific to the group of subjects we tested.

We are in complete agreement with Pruitt when he suggests that "testing conditions" and "the quality of the stimuli" were responsible for the improvements we observed. After all, providing the appropriate "testing conditions" and ensuring "quality of the stimuli" was one of the major goals of our experiment to begin with! It is important to note here that our pretest-posttest materials were the same words that were used by Strange and Dittmann, who failed to obtain any reliable changes from the pretest to the posttest. The implications of our results is that we used an effective training program whereas previous efforts did not.

Finally, Pruitt notes that for words containing /r/ and /l/ in initial position, the pretest performance for the subjects tested in our experiment was "quite good when compared to other studies with native Japanese speakers and naturally produced stimuli (e.g., Strange and Dittmann, 1984, showed only 64% correct identification on word-initial singletons at pretest)." The point of this statement appears to be that our subjects' initial level of performance was solely responsible for the improvements we observed. We take issue with this on the grounds that the level of word-initial singleton performance obtained in Strange and Dittmann's pretest is not representative of all Japanese subjects. For example, Mochizuki (1981) reported a figure of approximately 88% correct for seven Japanese subjects presented words containing /r/ and /l/ in word-initial singleton positions, a value higher than the 80% we observed for word-initial singletons in the pretest phase of our experiment. Furthermore, we have recently replicated our earlier findings with a group of monolingual Japanese speakers (*N*=19) from Kyoto, Japan (Lively *et al.,* 1993b). Mean performance on the pretest across all phonetic environments was only 63% for these subjects and was as low as 52% for contrasts in initial consonant clusters. In short, there is no systematic evidence to back Pruitt's claims that initial levels of performance were responsible for the results obtained in Logan *et al.* We conclude that our training procedure was effective, even for subjects whose initial performance may have been very poor.

II. GENERALIZATION PERFORMANCE

Pruitt's second major point concerns the tests of generalization that we used to assess the listeners' ability to transfer what they had learned during training to novel words and novel talkers. The tests of generalization were administered to subjects after they had completed training and the posttest. It consisted of two parts, novel words produced by a novel talker and novel words produced by a familiar talker whom subjects had heard during training. Pruitt argues that generalization cannot be assessed using our procedure for two reasons. First, he states that there was no control for the intelligibility of the two talkers. Second, he states that there was no mention of controlling the phonetic contexts across the two parts of the test. Finally, he argues that we "…overstated the results of the test of generalization."

Designing a test of generalization to assess what listeners have learned from a training task poses several difficulties. Pruitt correctly points out that the inherent intelligibility of talkers is one factor that could play a significant role in determining listeners' performance. However, the complexity of this issue is illustrated by the following example. In our experiment we pretested all of the stimuli used in the generalization tests with native speakers of English to insure that all of the stimuli were equated for intelligibility. Yet, when the Japanese listeners were presented with the same stimuli, their performance varied as a function of talker and phonetic context. The obvious point of this example is that intelligibility depends upon whether the listener is a native speaker or nonnative speaker. The less obvious point relates to how stimuli can be equated for intelligibility when they are presented to nonnative listeners. If one were to attempt to equate the intelligibility of talkers for the Japanese listeners, what criteria would be used to demonstrate equivalent intelligibility? In the case of English listeners, all of the tokens selected for use in our experiment produced no /r/ or /l/ errors. Unfortunately, given the variability in performance due to phonetic context for the Japanese listeners, it is unclear how talkers could be best matched for intelligibility.

Instead, it seemed to us to be more productive to deal directly with variability in intelligibility and its effect on generalization. Thus, we compared the most intelligible talker in training to a talker that our subjects had never heard before. We tested three Japanese listeners and found that their performance was marginally better for the familiar talker than for the unfamiliar talker. Although it is possible that this effect was due to differences in the inherent intelligibility of the talkers, it is also consistent with a growing body of evidence suggesting that listeners retain quite detailed information about unique perceptual episodes, such as the specific talker producing a specific word (e.g., Craik and Kirsner, 1974; Goldinger, 1992; Palmeri *et al.,* 1993; Pisoni, 1992; Schacter and Church, 1992). In addition, Lively *et al.'s* (1993a) recent training results demonstrate that generalization performance appears to be governed by the composition of the training items, rather than inherent differences in intelligibility. These findings indicate that we should consider the contribution of different sources of variability to phonetic category acquisition before trying to identify an intelligibility metric for nonnative listeners.

Pruitt's second criticism of the test of generalization used in Logan *et al.* was that it was not clear whether the words produced by the two talkers were comparable in terms of phonetic context. The stimuli in our tests of generalization were minimal pairs of English words that contrasted /r/ and /l/ in the same five phonetic environments used during training. The majority of these words contained /r/ or /l/ in initial singleton position or initial consonant clusters. Thus, the distribution of items used during training and generalization tests were similar across phonetic environments. Unfortunately, syllabic and phonetic contexts could not be controlled precisely across stimulus sets due to the limited number of English words that contrast /r/ and /l/ in each potential context. The fact that none of the words used in the pretest and posttest or in training were repeated in either of the tests of generalization imposed further constraints on the distribution of words according to context. In short, although the tokens used in these three sets of words virtually exhaust the set of /r/–/l/ minimal pairs available in English, the composition of words in each set was comparable.

The final criticism Pruitt makes regarding our test of generalization is that we overstated our results. We disagree. Pruitt's argument is based, in part, on the marginal difference that we observed when we compared generalization performance for words produced by an old talker and words produced by a new talker. The most convincing evidence against Pruitt's position is that Lively *et al.* (1993b) have replicated the generalization results obtained by Logan *et al.* In their replication, Lively *et al.* used the same stimuli and procedures used in our original experiment to test Japanese listeners living in Japan. In addition to obtaining a reliable improvement between pretest and posttest as a function of training, the listeners also reliably identified novel words produced by a familiar talker more accurately than novel words produced by an unfamiliar talker. Given this replication of our earlier findings, we do not think that we overstated our original claims.

As a final comment on the issue of generalization, Pruitt's statements suggest that any difference in performance between familiar and unfamiliar talkers is at odds with our claims regarding the robust nature of our training procedure—"… it is paradoxical that Logan *et al.* insist that this nonsignificant trend indicates a lack of generalization considering their claims regarding stimulus variability and robust learning." We do not think that any of our statements were paradoxical. Instead, Pruitt's statements suggest that he considers generalization to be an all or none phenomenon. While such a position would immeasurably simplify the evaluation of training methods, it is unlikely that this is the case. As our data demonstrate, generalization depends on the degree of similarity between the training stimuli and the test stimuli. As the test stimuli become more similar to the training stimuli, they will be identified more accurately. Similarly, as test stimuli diverge from training stimuli, they will be identified less accurately (cf. Strange and Dittmann, 1984). Thus, discrepancies between pretest/posttest and generalization performance are not paradoxical unless one relies on a narrow definition of what generalization means.

III. CONCLUSION

It is rare in science for one experiment to address each and every variable that may affect the outcome of an experiment. If scientists delayed publication until they had meticulously examined every possible combination of experimental variables that might affect the outcome of an experiment, the progress of science would be ill-served. Logan *et al.* described an initial effort to examine the role of stimulus variability in training listeners to perceive non-native phonetic categories. We have taken this initial effort and used it as a starting point for systematically exploring some of the variables that may have contributed to the effects we obtained (e.g., Lively *et al.,* 1993a; Lively *et al.,* 1993b). Contrary to Pruitt, we believe that this study has advanced our understanding of some very basic processes involved in learning novel phonetic categories. Our results address several important issues in speech perception and, more generally, in the field of perceptual categorization. First, we have demonstrated an effective means for modifying phonetic perception in nonnative speakers of English. Second, our findings suggest that listeners develop context-sensitive representations for new phonetic categories, rather than abstract, idealized canonical representations. Talker variability and variability due to phonetic environment appear to be important factors in acquiring new phonetic categories. Finally, our results demonstrate that selective attention to the contrastive cues of novel phonetic categories can be quickly and robustly modified in the laboratory with relatively simple training procedures. The important role of stimulus variability in acquiring new phonetic categories contrasts sharply with the traditional assumption that listeners develop abstract context-invariant prototypes for nonnative speech sounds and that these units are acquired by focusing listeners' attention only on the contrastive or criterial cues used by native speakers of a language.

References

Craik F, Kirsner K. The effect of speaker's voice on word recognition. Q J Exp Psychol. 1974; 26:274–284.

Goldinger, SD. Research on speech perception: Tech Rep No 7. Speech Research Laboratory, Indiana University; Bloomington, IN: 1992. Words and voices: Implicit and explicit memory for spoken words.

Jamieson D, Morosan D. Training non-native speech contrasts in adults: Acquisition of the English / ð/–/θ/ 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]

Lively SE, Logan JS, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning perceptual categories. J Acoust Soc Am. 1993a (in press).

Lively SE, Pisoni DB, Tohkura Y, Yamada R. Training Japanese listeners to identify English /r/ and / l/: A replication and extension. J Acoust Soc Am. 1993b submitted to.

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]

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

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. J Exp Psychol: Learning, Memory Cognit. 1993; 19:309–328.

Pisoni, DB. Some comments on invariance, variability and perceptual normalization in speech perception. International Conference on Spoken Language Processing; Banff, Alberta, Canada. 1992.

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

Pruitt JS. Comments on 'Training Japanese listeners to identify /r/ and /l/: A first report" [J S Logan, S E Lively, and D B Pisoni, J Acoust Soc Am 89, 874–886 (1991)],'. J Acoust Soc Am. 1993; 94:1146–1147. [PubMed: 8240541]

Schacter DL, Church B. Auditory priming: Implicit and explicit memory for words and voices. J Exp Psychol: Learning, Memory Cognition. 1992; 18:915–930.

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]