Two procedures for training a novel second language phonetic contrast

JAMES EMIL FLEGE
University of Alabama, Birmingham

ADDRESS FOR CORRESPONDENCE
Department of Biocommunication, University of Alabama at Birmingham, VH 503,
Birmingham, AL 35294-0019

ABSTRACT
Native speakers of Mandarin who learn English as a second language (L2) frequently misidentify unreleased tokens of /t/ and /d/ in the final position of English words. The purpose of this study was to evaluate two training methods that might be used to increase Mandarin adults' accuracy in identifying such stops. Subjects were assigned to receive training using a two-alternative forced-choice procedure or a categorical same/different discrimination procedure. Small but significant increases in the percentage of correct identifications of /t/ and /d/ tokens were obtained for both groups of native Mandarin subjects, although their gains did not differ significantly. These gains were still evident for both groups 2 months after completing the training. The effect of training for both groups generalized to words that were not used in training. Contrary to the hypothesis that identification training promotes more robust long-term memory representations than same/different training, the magnitude of generalization observed for the two groups did not differ significantly. However, subjects seemed to maintain the effects of the same/different training better than the effects of the identification training. Taken together, the results obtained here challenge the view that identification training is superior to same/different training as a means for training novel phonetic contrasts.

Adults are language-specific perceivers of speech as the result of their perceptual attunement to the phonetic characteristics of the vowels and consonants of their native language (L1). For example, having heard /p t k/ produced with long-lag voice onset time (VOT) values, native English adults require longer VOT values to label synthetic stops as voiceless than native Spanish adults who have been exposed to short-lag realizations of /p t k/ (Flege & Eefting, 1986). Perceptual attunement to L1 phonetic norms facilitates the processing of L1 speech. It may have quite a different effect, however, for adults who are learning a second language (L2). Spanish learners of English, for example, tend to show a crossover from predominantly voiced to voiceless stop responses at shorter VOT values than English monolinguals (e.g., Williams, 1980). When asked to decide which member of a /ba/-to-/pa/ VOT continuum represents the “best” exemplar of the /p/ category, Spanish–English bilinguals generally choose stimuli
having a shorter VOT value than those chosen by English monolinguals (Flege & Schmidt, in press).

Differences between nonnative and native speakers in segmental phonetic perception may slow and/or impair nonnatives' comprehension of the L2 – especially synthetic versions of the L2, or L2 speech that is noncontextualized or heard in nonideal listening conditions (Bond & Moore, 1994; Bott, 1993; Greene, Pisoni, & Gradman, 1985; Mack, 1988; Mack, Tierney, & Boyle, 1990; McAllister, 1990; Nabelek & Donahue, 1984; Oyama, 1978; Ozawa & Logan, 1989). Despite the clear imprint of the L1 phonetic system on one's perception of phonetic segments in an L2, there is evidence that segmental phonetic perception remains malleable in adulthood, at least to some extent. Adults' perception of L2 vowels and consonants becomes more accurate (i.e., more nativelike) during the naturalistic acquisition of an L2 (e.g., Bohn & Flege, 1990; Flege & Bohn, 1989; Flege & Hillenbrand, 1985; Gottfried & Beddor, 1988; Mack, 1989; Williams, 1979; Yamada & Tohkura, 1992). However, it appears that few individuals who are first massively exposed to an L2 in adulthood learn to perceive L2 phonetic segments just like monolingual speakers of the target L2.

Short-term laboratory training experiments have been carried out to train nonnative adults on phonetic contrasts not found in their L1 (e.g., Flege, 1989; Flege & Wang, 1989; Jamieson & Morosan 1986; Lively, Yamada, Tohkura, & Yamada, 1994; Logan, Lively, & Pisoni, 1991; Pisoni, Lively, & Logan, 1993; Polka, 1989, 1991; Pruitt, Strange, Polka, & Aquilar, 1990; Strange & Dittman, 1984; Strange, Polka, & Aquilar, 1989). In a typical study, about 800 training trials are administered during the course of several sessions. For most nonnative subjects, the ability to identify differentially realizations of novel L2 phonetic categories improves significantly as the result of such training. Gains of about 10% are common. The effects of perceptual feedback training have been shown to generalize to stimuli not used in training, as well as to new speakers, and to persist for up to 6 months beyond the completion of training. However, as for naturalistic (i.e., unaided) L2 acquisition, nonnative adults who have received laboratory training usually do not identify L2 phonetic segments as accurately as do monolingual speakers of the target L2.

The size of the gains obtained in previous training studies has varied considerably, owing to a wide range of differences among the studies. These differences include the characteristics of the nonnative subjects studied, the relation of the L2 phonetic categories to those in the L1, and the nature and range of the stimuli. Another important difference among the studies is the procedure used in the training (Strange, 1992). The stimuli in most of these studies consist of multiple realizations of minimally paired words that exemplify the phonetic contrast in question (e.g., right vs. light). Subjects who receive identification training hear a single stimulus on each trial. Their task is to identify each stimulus in terms of two categories (/r/ or /l/ in the example given). Subjects gradually learn the properties of the two categories as they are told the correct identity of each stimulus immediately after they have labeled it.

Subjects who receive same/different discrimination training hear two
stimuli on each trial. Their task is to decide whether each pair contains two instances of one category or instances of two different categories. In the categorical same/different procedure evaluated in the present study, the stimuli in each pair were always physically nonidentical. Thus, the subject's task was not to decide whether one stimulus had been presented twice in succession (and thus are physically identical), but to decide whether two stimuli with “name identity” (i.e., two different realizations of a single phonetic category) had been presented.

It is widely believed that a nonnative's initial difficulty in perceiving L2 phonetic segments is not the result of irreversible sensory loss. It is also generally held that the gains yielded by short-term laboratory training (or those occurring during naturalistic L2 acquisition) derive from modifications in selective attention and/or a reweighting of the multiple properties that define such phonetic categories (see, e.g., Strange, 1994). Some controversy exists, however, as to the most effective method for inducing such changes in segmental phonetic perception.

Lively et al. (1994) enumerated the advantages of a two-alternative forced-choice identification procedure. When many different tokens of two categories are randomly presented along with immediate feedback, this training procedure may promote the formation of new phonetic categories that are “robust” — that is, they are not adversely affected by variations in stimulus properties that are irrelevant to phonetic identity (e.g., variations in speaking rate or idiosyncrasies of individual talkers) (see also Pisoni et al., 1993). Logan et al. (1991) suggested that identification training may be more effective than same/different training because it encourages subjects to develop and use phonetic codes in long-term memory rather than to rely on rapidly fading information in sensory memory. Jamieson and Morosan (1986, 1989) suggested that, because same/different training encourages attention to within-category acoustic differences, it may fail to increase listeners' sensitivity to the core properties that define the L2 categories. Fewer reservations have been expressed concerning the identification training procedure. However, Polka (1992) observed that subjects who receive identification training may learn to respond correctly by detecting any property that might be used to identify differentially the two categories — not necessarily the properties used by native speakers.

According to Burnham, Earnshaw, and Quinn (1987), identification and same/different discrimination exploit similar underlying processes, but Strange (1992) drew a distinction between a traditional same/different (AX) discrimination task and a categorical (or “categorial”; W. Strange, personal communication, 1995) same/different training procedure. Same/different training may encourage listeners to ignore within-category acoustic variation while attending to the phonetically relevant differences that define the two categories. According to Polka (1992), a categorical same/different procedure requires a degree of perceptual constancy because the multiple tokens of each category are likely to encompass a wide range of acoustic variants. This may encourage subjects to acquire new “equivalence classes” for L2 phonetic segments (Strange, 1992).
Given the limited amount of time that adult learners of an L2 might be willing to devote to "fine-tuning" their perception of L2 phonetic segments, it is of practical importance (and also theoretical interest) to assess the relative efficacy of the identification and categorical same/different procedures. No previous study, to our knowledge, has directly compared these two procedures. Native Mandarin subjects in the present study were trained to distinguish /t/ and /d/ in the final position of English words. (Mandarin speakers often have difficulty identifying unreleased word-final stops because their L1 does not permit obstruents to occur in word-final position.) Half of the subjects were assigned to receive identification training; the other half received same/different training. The words used in training were spoken by adult, female native speakers of English. One question of interest was, would the two groups of native Mandarin subjects differ in terms of the gains resulting from such training? Also, would the effects of identification and same/different training generalize equally to words not used in training.

**METHOD**

**Subjects**

Participating in this study were 20 adult native speakers of Mandarin. The subjects had arrived in the United States at an average age of 27 years (range, 23–36) and had been living in the United States for an average of 1.0 year at the time of testing (range, 0.1–3.5). The native Mandarin subjects reported having studied English in school for an average of 7 years prior to their arrival (range, 6–9). Their estimated use of English on a daily basis was 26% of the time (range, 0%–47%).

The subjects (5 female, 5 male) were randomly assigned to receive identification or categorical same/different feedback training on the contrast between English word-final /t/ and /d/. The identification and same/different subjects differed little in their average age upon arrival (26 vs. 28 years) or their length of residence in the United States (0.8 vs. 1.1 years), amount of formal instruction in English (7 vs. 8 years), or self-estimated daily use of English (23% vs. 34%). The identification and same/different subjects also gave comparable evaluations of their own ability to pronounce English using a 7-point rating scale (4.3 vs. 4.1).

A control group (5 females, 5 males) of monolingual native speakers of English also participated. Their average age was the same as that of the native Mandarin subjects (viz., 28 years). Six native English subjects were from Alabama, and one each was from Mississippi, Kentucky, Louisiana, Ohio, and Ontario. Like the Mandarin subjects, the native English subjects were all recruited on the campus of the University of Alabama at Birmingham.

**Stimuli**

Mandarin does not permit a word to end in an obstruent consonant; thus, the phonetic contrast between the /t/ and /d/ in English words like *beat* and *bead* might be regarded as a "novel" phonetic contrast for Mandarin
speakers who learn English as an L2. However, previous work has shown that native Mandarin speakers of English living in the United States have relatively little difficulty identifying tokens of word-final English /t/ and /d/ that have audible release bursts (Flege, 1989; Flege & Wang, 1989; see also Hsia, 1992). This is consistent with the more general finding that novel voicing contrasts are easier to learn than are novel place of articulation contrasts (Strange, 1992; Werker & Tees, 1984). However, native Mandarin subjects have great difficulty identifying unreleased tokens of /t/ and /d/. This is important because word-final stops are often unreleased in conversational English. In the present study, then, the Mandarin subjects were trained to identify word-final /t/ and /d/ tokens whose final release burst had been removed digitally.

The decision was made to use naturally produced words rather than synthetic stimuli because synthetic speech typically does not represent the full range of acoustic properties that specify phonetic categories (e.g., Pisoni, Nusbaum, & Greene, 1985). The native speakers of English read a list of words containing four minimally paired consonant–vowel–consonant words (viz., beat–bead, bit–bid, bet–bed, bat–bad). The final stops in words spoken by one male and one female speaker were sometimes unreleased and so were eliminated from the corpus. One token of the minimally paired words spoken by each of the remaining eight speakers was digitized at 10 kHz with 12-bit amplitude resolution. All 64 digitized words had an audible final release burst, and all of the /d/-final words had voicing (i.e., glottal pulsing) that extended through some part or all of the closure interval.

The CVC stimuli were normalized for peak amplitude. Copies of the 64 CVCs were then edited, as described in Flege (1989). Everything after the complete constriction of the word-final stop (signaled by changes in waveform shape and intensity) was removed digitally. This eliminated three important cues to the /t/–/d/ contrast from the “edited” CVC stimuli: (a) the presence versus absence of closure voicing in the stop closure interval, (b) the duration of the stop closure interval, and (c) the amplitude, frequency, and temporal differences residing in the /t/ and /d/ release bursts. The editing procedure did not affect other cues to the /t/–/d/ contrast, such as vowel duration, vowel quality, and F1 offset frequency. Cues such as these provide sufficient information for native English subjects to identify /t/ and /d/ tokens correctly.

The intelligibility of stops in both the edited and unedited CVC stimuli was assessed in a preliminary experiment. The 128 stimuli (8 Speakers × 8 Words × 2 Editing Conditions) were randomly presented five times each to seven native English-speaking listeners. The words produced by the one male and one female speaker whose stops were misidentified most often were eliminated from the corpus. Stops produced by the remaining six speakers were identified correctly at high rates, both in the unedited CVCs (M = 99% correct) and in the edited CVCs stimuli (M = 96%). The correct identification scores were comparable for /d/ and /t/ (99% and 96%, respectively) and for stops produced in the context of all four vowels (/i/: 99%, /i/: 99%, /ε/: 96%, /æ/: 95%).
The stops produced by the six speakers were not equally intelligible. The correct identification scores for the stops in the unedited CVCs ranged from 93% for one male speaker to 99.5% for two female speakers. The extent to which editing lowered the correct identification scores ranged from 0.3% for one female speaker to 7.8% for one male speaker. An examination of the results for individual stimuli revealed that most of the misidentifications of the edited CVC stimuli were for two tokens of bat, both produced by male speakers. Inspection of the waveforms for these stimuli revealed that they had been edited incorrectly: in both instances, one too many glottal pulses had been removed. When the two stimuli were reedited, the final /t/s in these CVCs were clearly identifiable to the author as /t/. The final stimulus set consisted of 96 CVC words (6 Speakers \(\times\) 8 Words \(\times\) 2 Editing Conditions). Words spoken by the three female speakers were used as the training stimuli; words spoken by the three male speakers were used to test for generalization of training.

**Procedure**

The native Mandarin subjects participated in 10 sessions over a 3-month period. They were told that the purpose of the study was to increase their ability to identify word-final English stops. They were informed that they would receive feedback training in 7 sessions over a 3-week period, and that their ability to identify English /t/ and /d/ would be assessed in a pretest, immediately after each training session, and in two posttests held after the completion of training. The native English subjects participated in just two sessions (Day 1, Day 9) separated by about 5 weeks. Unlike the native Mandarin subjects, they received no feedback training.

All of the subjects were tested individually in a sound booth. The stimuli were presented binaurally via headphones at a comfortable level. The subjects responded by pushing one of two buttons on a response box (marked either “t” vs. “d” or “same” vs. “different”).

Table 1 indicates the stimuli presented during sessions held over the 10 days of the study and the number of random presentations. The two aims of the identification pretest (administered on Day 1) were to provide a baseline for the Mandarin subjects and to replicate the effects of editing obtained in previous studies. The unedited CVCs spoken by the male and female speakers were presented for identification without feedback (the order was counterbalanced) to all 30 subjects on Day 1, followed by the edited version of the same CVCs. We expected the native English subjects to identify the word-final /t/ and /d/ tokens at near-perfect rates in both the unedited and edited CVCs, and this was indeed the case. Thus, as planned, only the edited CVCs were used in the remainder of the study.

Feedback training was administered on Days 2 to 8. Only the CVCs produced by the female speakers were used in the training. The subjects’ ability to identify correctly, without feedback, the word-final stops in the CVCs was assessed immediately after the training on Days 2 through 8. Despite the fact that only 24 edited CVCs were used in the training (3
Female Speakers × 8 Words), the Mandarin subjects continued to make identification errors, although the number of errors decreased gradually. After each identification test the subjects were told how many correct responses they had given in order to sustain their motivation.

The edited CVCs spoken by the female and the male speakers were presented for identification without feedback on Days 9 and 10. Data for the male speakers provided a test of generalization because these stimuli had been heard only once before, during the pretest on Day 1. It is widely believed (e.g., Morosan & Jamieson, 1989) that the use of multiple tokens for each category and/or the inclusion of multiple speakers will increase the extent to which the effects of the feedback training will generalize to untrained stimuli. Stimulus variability is thought to encourage subjects to focus on general rather than token-specific properties of the stimuli (e.g., Logan et al., 1991; see also Posner & Keele, 1968). Morosan and Jamieson (1989) found that training on synthetic stimuli patterned after an adult male voice generalized equally to words spoken naturally by adult males and females. However, it appears that, when training is administered on naturally produced words, subjects are better able to identify phonetic segments in words spoken by known speakers (i.e., those who produced the training stimuli) than in words spoken by an unfamiliar speaker (Lively et
One question of interest, then, was whether the two procedures used here for training stops in the words spoken by the female speakers would yield different amounts of generalization to the CVCs produced by the male speakers.

The data from Day 9 and Day 10, which were held 2 weeks and 2 months, respectively, after the completion of the training, were also used to assess the persistence of training effects. Tees and Werker (1984) reported that training-induced gains in native English subjects' ability to discriminate a contrast between unfamiliar syllable-initial Hindi stops was no longer evident several weeks after completing the training. Lively et al. (1994) trained Japanese adults to identify English /ʃ/ and /θ/ in several phonetic contexts. The significant gains on word-initial and word-final singleton liquids, the result of identification training, were still evident 3 months after the training, but not 6 months afterwards. Thus, another question of interest was whether the effects of identification and same/different training on the Mandarin subjects' identification of burstless word-final stops would persist for 2 months and, if so, to what extent.

The pretest, feedback training, posttraining identification tests, and posttests were all run on-line using a PDP 11/73 computer. The CVC stimuli were converted from digital to analog form, low-pass filtered (4 kHz), and presented binaurally to the subjects. In sessions that required an identification response, the subjects were told to push a button marked “t” or “d” after hearing each CVC stimulus. During same/different training, the subjects were told to push a button marked “same” or “different.” The subjects were required to respond to each trial and were told to guess if uncertain. Each new trial was presented 1 sec after a response was given to the preceding trial. During the feedback sessions, a light was illuminated over the correct button immediately after a response was given.

Two stimuli per trial were presented to the subjects in the same/different training, whereas only one stimulus per trial was presented in the identification training. This raised the question – how can we balance out the differences in the amount of training provided by these procedures? The solution was to give feedback to the identification subjects on just the first 120 (of 240) randomly presented stimuli. This was done to ensure that the subjects in the two groups heard the same number of stimuli during training and received the same amount of information in regard to the accuracy of their responses.

The same/different subjects received feedback on all 120 pairs in each training session. Of these, 60 were “same” and 60 were “different” pairs. The two CVCs in each pair were always spoken by different speakers. The same pairs consisted of two physically different CVCs that ended in either /t/ or /d/, whereas the different pairs consisted of one word ending in /t/ and one word ending in /d/. The interstimulus interval (ISI) between the members of each pair was fixed at 250 msec. This relatively short ISI was chosen to maximize the difference between the same/different and identification procedures. Short ISIs encourage subjects to rely on sensory memory, whereas, given the rapid deterioration of information in sensory
memory, longer ISIs encourage subjects to rely on the phonetic codes generated for each stimulus (e.g., Werker & Logan, 1985).

RESULTS

Pretraining baseline

Four mean percentage correct identification scores were computed for each subject, based on his or her identification of stops on Day 1—one each for the edited and unedited CVCs produced by male and female speakers. Each mean was based on 40 forced-choice responses (4 Minimally Paired Words \( \times \) 5 Random Presentations). As expected (Flege, 1989; Flege & Wang, 1990), the native English subjects identified /t/ and /d/ almost equally well in the unedited and edited CVCs (\( M = 99\% \) and 96\% correct, respectively). Also as expected, the native Mandarin subjects correctly identified word-final stops far more often in the unedited CVCs (identification subjects: 92\%; same/different subjects: 94\%) than in the edited CVCs (66\% correct for both groups).

Because there were substantial differences in variance across the three groups, the percentage correct identification scores were transformed using an arcsine transformation (Kirk, 1968). These transformed scores were then submitted to a Group (3) \( \times \) Editing Condition (2) \( \times \) Speaker Gender (2) ANOVA, with repeated measures on the last two factors. The main effect of group, \( F(2, 27) = 136.0, p < .05 \), and the main effect of editing, \( F(1, 27) = 320.2, p < .05 \), were significant, but the gender main effect was nonsignificant, \( F(1, 27) = 1.36, p > .10 \).

A significant Group \( \times \) Editing Condition interaction was obtained, \( F(2, 27) = 30.5, p < .05 \). Paired t tests revealed that the percentage correct scores were significantly higher in the unedited than in the edited CVC stimuli for all three groups (\( p < .05 \)). The interaction seemed to have occurred because the size of the editing effect was much larger for the Mandarin subjects who received identification and same/different training (27\%, 28\%), than for the native English subjects, who showed a mean decrease of only 3\% due to the editing.

The Group \( \times \) Speaker Gender interaction also reached significance, \( F(2, 27) = 5.63, p < .05 \). Paired t tests revealed that the native English subjects identified stops significantly more often in the CVCs produced by female than male speakers (98.5\% vs. 97.0\%, \( p < .05 \)). However, the difference between the CVCs produced by the male and female speakers was nonsignificant for both groups of native Mandarin subjects (\( p > .05 \)). Thus, any difference between these CVCs observed later in the study for the native Mandarin subjects could be attributed to the training administered on the CVCs spoken by the female speakers.

The simple main effect of group was significant for the CVCs produced by both the female and male speakers, \( F(2, 27) = 100.1 \) and 95.8, respectively (\( p < .05 \)). Post-hoc tests (Tukey’s HSD) revealed that, for the CVCs produced by both genders, the native English subjects identified
Figure 1. Mean percentage correct identification scores for word-final tokens of /t/ and /d/ in the edited CVCs spoken by female speakers. Identification (ID) and same/different (SD) subjects received identification and categorical same/different training, respectively. The scores obtained on Days 2 to 8 were obtained immediately after feedback training was administered; Days 9 and 10 occurred 2 weeks and 2 months, respectively, after the completion of training. The brackets enclose ±1.0 standard error.

stops significantly more often than did both groups of native Mandarin subjects ($p < .05$). Importantly, the subjects assigned to receive the identification or same/different training did not differ significantly from one another ($p > .05$).

Training results

Figure 1 shows the rate at which /t/ and /d/ tokens in the CVCs spoken by female subjects were identified correctly on the 10 days of the study. On Day 1, both groups of native Mandarin subjects identified the stops correctly in $66\%$ of instances. The percentage correct scores increased as the same CVCs were presented (without feedback) after the training administered on Days 2 to 8. Compared with the Day 1 scores, the Day 8 scores were $17\%$ higher for the subjects who received identification training, and $11\%$ higher for the subjects who received same/different training. An examination of the scores obtained on Days 9 to 10 revealed that the effect of the training persisted for up to 2 months afterwards. For both the identification and the same/different subjects, the percentage correct scores remained higher on Day 9 ($83\%, 77\%$) and on Day 10 ($80\%, 77\%$) than on Day 1 ($66\%$ for both groups).
Although the two native Mandarin groups did not differ on Day 1, scores were higher for the identification than the same/different subjects following the training administered on Day 2 and on all subsequent days of the study. To determine if the two groups differed reliably, the percentage correct identification scores were submitted to a Group (2) x Day (10) ANOVA, with repeated measures on the day factor. The day effect was significant, $F(9, 162) = 10.8, p < .05$, the group factor was not, $F(1, 18) = 2.5, p > .10$, nor was the two-way interaction, $F(9, 162) = 1.04, p > .10$. The lack of a significant two-way interaction suggests that the benefits derived from the identification and the same/different training did not differ significantly.

To determine how much training was needed to increase significantly the native Mandarin subjects' ability to identify edited word-final stops, the scores obtained from the identification and same/different subjects were averaged. A Tukey's test revealed that the scores obtained on the Day 1 pretest (66% correct) and after the first training session on Day 2 (71%) did not differ significantly ($p > .10$). However, the scores obtained after the second training session on Day 3 (73% correct) and all subsequent days were significantly higher than the Day 1 scores ($p < .05$). The Tukey's test also revealed that the percentage correct scores obtained on Days 6 and 7 were higher than the Day 2 scores, and the Day 8 scores were higher than the Day 2 and Day 3 scores ($p < .05$). No other differences reached significance.

A series of Group x Day ANOVAs were carried out to examine the scores obtained before and after training (i.e., the percentage correct scores obtained on Day 1 vs. Day 9, Day 1 vs. Day 10, and Day 2 vs. Day 8). All three analyses indicated that /t/ and /d/ were identified significantly more often following the training ($F$ values ranging from 19.1 to 34.1, $p < .05$). None of these analyses yielded a significant main effect of the group factor ($F$ values ranging from 0.28 to 1.19, $p > .10$) nor a significant two-way interaction ($F$ values ranging from 0.16 to 1.32, $p > .10$). These results also suggested that the identification subjects did not benefit significantly more from their training than did those subjects who received the same/different training.

Generalization

An important measure of the efficacy of a training procedure is the extent to which its effects generalize to stimuli not used in training. A high degree of generalization suggests that a training procedure has engendered the formation of a long-term memory representation that is more abstract than the sum total of the physical properties encountered in the training stimuli.

To assess generalization, we compared the percentage of correct identifications obtained for the CVC words spoken by the female speakers, which were used in training, and for the CVCs spoken by the male speakers, which were not. Scores obtained for the CVCs produced by the male and female speakers on Day 1 (prior to training), Day 9 (2 weeks after training), and
Figure 2. Mean correct identification scores for /t/ and /d/ tokens in words used in training ("female stimuli") and words not used in training ("male stimuli") prior to feedback training (D1) and after training (D9, D10). The native Mandarin subjects received either identification (ID) or same/different (SD) training. The brackets enclose ± 1.0 standard error.

Day 10 (2 months after training) were compared. As in the foregoing analyses, the percentage of times each subject correctly identified the /t/ and /d/ tokens in four minimally paired words was calculated.

The native English subjects were tested on edited final stops in the CVCs spoken by male and female speakers on Day 1 and Day 9 only. Their overall rate of correct identifications on those two days (96% vs. 97%) did not differ significantly, $F(1, 9) = 2.2, p > 0.10$, nor did stimulus familiarity (i.e., female vs. male talkers) interact significantly with the day of testing, $F(1, 9) = 2.64, p > .10$.

Figure 2 shows the mean percentage correct identification scores obtained in the pretest (Day 1) and in the two posttests (Days 9 and 10) for the /t/ and /d/ tokens spoken by the male and female speakers. The identification and same/different subjects' pretest scores were quite similar ($M = 66\%$ for both groups). The overall rate of correct identifications obtained for the identification and same/different subjects was also similar (74% vs. 72%). The accuracy with which the subjects in both groups identified English stops in both sets of CVC stimuli increased as the result of the training. The percentage correct scores on Day 9 and Day 10 averaged 80% and 77%, respectively, for the identification subjects. They averaged 74% and 76% for the same/different subjects. The increase from the Day 1 pretest
to the two posttests was somewhat larger for the CVCs produced by the female than the male speakers.

The percentage correct scores obtained for the native Mandarin subjects were submitted to a Group (2) × Day of Testing (3) × Speaker Gender (2) ANOVA, with repeated measures on the day and talker factors. The main effect of group was nonsignificant, $F(1, 18) = 0.68, p > .10$, but the main effect of day was significant, $F(2, 36) = 22.7, p < .05$. The scores obtained for all 20 Mandarin subjects were examined using a post-hoc test; this revealed that the Day 9 and Day 10 scores, which did not differ significantly from one another, were significantly higher than the Day 1 scores. If identification training is more effective than same/different training, then one might have expected a larger difference for pre-versus posttest scores for the subjects who received the identification training. However, the Group × Day interaction did not reach significance, $F(2, 36) = 1.38, p > .10$.

A significant Day × Speaker Gender interaction was obtained, $F(2, 36) = 8.07, p < .05$. Tests of simple main effects revealed that the percentage correct identification scores obtained for all 20 native Mandarin subjects on Day 1 for the CVCs produced by the female and male speakers (66% for both genders) did not differ significantly ($p > .10$). However, the native Mandarin subjects identified stops significantly more often in the CVCs produced by female than male speakers on Day 9 (80% vs. 74% correct) and on Day 10 (79% vs. 74%) ($p < .05$). This finding suggests that the subjects retained information pertinent to the specific tokens used during the training. If identification training promotes more general representations in long-term memory than does the same/different training, then one might have expected to see larger female versus male differences in the scores obtained on the posttest for the same/different than the identification subjects. However, neither the Group × Speaker Gender interaction, $F(1, 18) = 0.38, p > .10$, nor the Group × Speaker Gender × Day interaction, $F(2, 36) = 0.17, p > .10$, reached significance.

Inspection of Figure 2 reveals that the scores obtained for the identification subjects decreased slightly from Day 9 (2 weeks after the training) to Day 10 (2 months after the training), whereas the same/different subjects showed a small increase. To determine if the effects of the two types of training persisted to differing extents, the percentage correct scores obtained on Days 9 and 10 were submitted to a Group (2) × Day (2) × Speaker Gender (2) ANOVA. This analysis yielded a significant Group × Day interaction, $F(1, 18) = 8.15, p < .05$, but a nonsignificant three-way interaction, $F(1, 18) = 0.72, p > .10$. Tests of simple main effects were carried out to examine the average scores obtained for the CVCs produced by the female and male speakers. These tests revealed that the identification subjects identified stops significantly less often on Day 10 than Day 9 (77% vs. 80%) ($p < .05$), whereas the same/different subjects showed a nonsignificant tendency to identify stops correctly more often on Day 10 than Day 9 (76% vs. 74%), ($p > .05$). The two groups did not differ on either Day 9 or Day 10, however ($p < .05$).
DISCUSSION

The results obtained on the Day 1 pretest replicated findings obtained in two previous studies examining the identification of word-final stops without release bursts (Flege, 1989; Flege & Wang, 1990). Adult native speakers of English managed to identify at near-perfect rates /t/ and /d/ tokens in minimally paired words which had a final release burst, and also tokens in edited copies of the same words after their final release bursts and closure voicing (if any) were removed digitally. The 20 native Mandarin subjects, who arrived in the United States in adulthood, had little difficulty identifying /t/ and /d/ in the unedited words. However, they frequently misidentified /t/ and /d/ tokens found in the edited CVC stimuli. It appears that they were unable to exploit the cues that remained after editing (e.g., the duration and quality of the vowel, F1 offset frequency) as effectively as did the native English subjects. In conversational English, /b d g/ and /p t k/ tokens are often unreleased in word-final position, and tokens of /b d g/ may be produced without closure voicing. The native Mandarin subjects' difficulty in identifying final stops in the edited CVC stimuli may have been indicative of a difficulty with the phonetic perception of conversational English.

The Mandarin subjects were randomly assigned to receive one of two different kinds of feedback training on the edited CVC stimuli. One training procedure made use of a two-alternative forced-choice procedure; the other made use of a categorical (or “categorial”) same/different discrimination procedure. Both groups of subjects heard the same small set of stimuli (edited CVC words spoken by female speakers) in sessions distributed over a 3-week period. The identification of /t/ and /d/ was assessed in a two-alternative forced-choice test (without feedback) after the training was administered in each of the 7 feedback training sessions.

The rate at which /t/ and /d/ tokens were identified correctly was assessed after each training session. The percentage correct identification scores obtained for both native Mandarin groups increased steadily. Compared with the scores observed on the pretest (Day 1), the correct identification scores increased by 17% for the identification subjects and by 11% for the same/different subjects. However, the difference between the two training groups did not reach significance. Moreover, the slightly greater gains obtained for the identification subjects might be attributed in part to the similarity of the method used to evaluate the effects of the training (viz., an identification test) and the method used to train the identification subjects (a two-alternative forced-choice test with feedback).

A number of investigators have suggested that an identification training procedure is more effective than a same/different procedure. A possible basis for the putative difference is that a same/different procedure may cause subjects to focus on phonetically nonrelevant, within-category differences rather than on acoustic phonetic attributes that are essential to the identity of the phonetic category (Jamieson & Morosan, 1986, 1989). It has also been suggested that identification training promotes the development
of more robust or general representations in long-term memory than does same/different training (Logan et al., 1991; Pisoni et al., 1993).

Generalization was assessed by comparing the percentage correct identification scores obtained for CVC stimuli spoken by female speakers (which served as the training stimuli) and CVC stimuli produced by male speakers. The scores were higher 2 weeks and 2 months after completing the training than before it, both for the female- and male-produced CVCs. However, all of the subjects, whether they received the identification or the same/different training, identified the stops more often in the training CVCs than in nontraining stimuli. This finding, which is similar to that reported in Lively et al. (1994), suggests that the Mandarin subjects retained information pertinent to the specific stimuli on which they received feedback training. This finding does not in itself disconfirm the hypothesis that identification training promotes the formation of more general long-term memory representations for L2 phonetic categories than does same/different training. Still, the similarity of the difference between trained and nontrained stimuli for the subjects who received the two kinds of training challenges this hypothesis. Had the subjects who received the identification training established a more general representation for English word-final stops, one might have expected them to show less of a difference between the CVC stimuli produced by the female and male speakers than subjects who received the same/different training.

The results obtained here support the view that, like identification training, categorical same/different training encourages subjects to establish equivalence classes based on phonetically relevant properties rather than prompting them to focus exclusively on token-specific information (Polka, 1992; Strange, 1992). The hypothesis that same/different training encourages attention to within-category differences and a reliance on information in sensory memory may need to be restricted to tasks in which subjects must decide, in regard to a limited number of stimuli, whether or not the members of a pair are physically identical. Moreover, the study provided preliminary evidence of one potentially important advantage of same/different training. The identification subjects showed a slight decline from Day 9 (2 weeks after the completion of training) to Day 10 (2 months after training). The same/different subjects, on the other hand, seem to have consolidated what they had learned over time; they identified stops correctly more often on Day 10 than on Day 9. The size of the increase was small, however.

Although the two kinds of training yielded generally comparable outcomes, an incidental finding of the study provided a reason for favoring identification over same/different training. The native Mandarin subjects were queried on Day 8 concerning their personal reactions to the training study. Those who received identification training responded more positively than those who received same/different training. The identification subjects felt that they benefited more from the training than did the same/different subjects ($M = 3.0$ vs. $3.9$, on a 7-point scale ranging from very much to very little). The identification subjects also said that they enjoyed their training experience more than did the same/different subjects ($M =$
1.0 vs. 3.3). In addition, the identification subjects seemed more willing than the same/different subjects to receive additional training ($M = 1.5$ vs. $2.4$, on a scale ranging from definitely yes to definitely no).

In summary, this study assessed two methods for training a novel second language (L2) phonetic contrast: a two-alternative forced-choice identification procedure and a categorical (or categorial) same/different discrimination procedure. The identification training procedure yielded non-significantly greater gains than did the same/different training, and both procedures yielded a similar amount of generalization to nontrained stimuli. The study provided evidence that the effects of same/different training may persist to a greater extent than those of identification training. Taken together, the results challenge the view that identification training is greatly superior to categorical same/different training. However, the subjects seemed to have responded more positively, in subjective terms, to the identification training than to the same/different training that was administered. This may be the strongest reason at present for favoring an identification procedure over a same/different procedure when attempting to train a novel L2 phonetic contrast.

**ACKNOWLEDGMENTS**

This research was supported by NIH grant DC00257. The author thanks Laurie Skelton and Chipin Wang for their important contributions to the study, and Anna Schmidt for comments on a previous version of this article.

**NOTES**

1. The same pattern of results was also obtained when only the scores obtained for Days 1 through 8 were examined, as well as when the scores for Days 1 through 10 were transformed using an arcsine transformation (Kirk, 1968). The same results were also obtained when the responses obtained on all 10 days were converted to $A'$ scores. Such scores represent a bias-free measure of perceptual sensitivity. They were derived from the proportion of correct identifications of /t/ ("hits") and the proportion of incorrect identifications of /d/ as /t/ ("false alarm") (see Grier, 1971).

2. The same results were obtained when the raw percentage correct scores were transformed using the arcsine transformation recommended by Kirk (1968) and also when $A'$ values were examined.

**REFERENCES**


contrasts: Generalization across voices, but not word-position or task. *Journal of Speech and Hearing Research*, 32, 501-511.


