Perception & Psychophysics
1986, 40 (4), 205-215

Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones

DONALD G. JAMIESON and DAVID E. MOROSAN
University of Calgary, Calgary, Alberta, Canada

Speech perception abilities are modified by linguistic experience to maximize sensitivity to acoustic contrasts that are important for one's linguistic community, while reducing sensitivity to other acoustic cues. Although some of these changes may be irreversible, in other cases adults may learn to perceive non-native speech sounds in a linguistically meaningful manner with limited perceptual training. The present study investigates the possibility of using a technique based on perceptual fading to train Canadian francophone adults to distinguish the voiced and voiceless "th" sounds of English: /ð/, as in "the," versus /θ/, as in "theta." Following a pretest to measure identification and discrimination performance with both natural and synthetic speech tokens, 10 subjects were trained using synthetic stimuli. Approximately 90 min of this training improved performance with both natural and synthetic tokens relative to that of untrained control subjects. The results suggest that there is a much higher degree of plasticity in these acoustic/linguistic categories than would be inferred from the normal performance of Canadian francophones who learn English as adults. The nature of the training technique is discussed in relation to other training paradigms.

Linguistic experience produces major and very durable changes in the perception of some speech sounds. For example, English speakers place the phonetic boundary separating /b/ and /p/ at approximately +25 msec voice-onset time (VOT; see Lisker & Abramson, 1970; Williams, 1977, 1979), whereas Spanish speakers place this boundary at approximately 0 to −5 msec VOT (see Williams, 1977). Similarly, speakers of Canadian French place the voiced/voiceless categorical boundary at a shorter VOT value than do unilingual English or bilingual French/English speakers (see Carmazza, Yeni-Komshian, Zurif, & Carbone, 1973). One consequence of such subtle language-specific distinctions is that adults who learn a second language often have special difficulty with the perception and production of sounds that are distinct in the new language but allophonic in their native language. The example considered in this paper is one aspect of the voiced/voiceless distinction between the English consonants /ð/ and /θ/. Francophones who learn English late in life demonstrate a special difficulty with these sounds, and also confuse /ð/ with /d/ and /θ/ with /t/. The present paper reports an experimental examination of the possibility of efficiently training such distinctions.

At least some speech discrimination abilities are known to change during the first year of life in response to linguistic environment. For example, Werker and her associates demonstrated that young (6–8 months) infants from English-speaking homes can discriminate non-English (Hindi and Thompson[1]) speech phones as well as native-speaking adults, whereas English-speaking adults and older (10–12 months) English infants cannot (Werker, Gilbert, Humphrey, & Tees, 1981; Werker & Tees, 1983, 1984a). Lasky, Syrdal-Lasky, and Klein (1975) demonstrated that young (4–6.5 months) infants from Spanish-speaking homes in Guatemala could discriminate voiced/voiceless contrasts (+20 msec vs. +60 msec) as well as prevoiced/voiced contrasts (−20 msec vs. −60 msec) but could not discriminate speech sounds centered on the normal Spanish boundary of 0 msec (−20 msec vs. +20 msec). Spanish infants, like English infants, must therefore undergo a perceptual reorganization early in their language acquisition.

In the only studies to date of infant sensitivity to a voiced/voiceless fricative distinction, Eilers, Wilson, and Moore (1977) reported that infants 6 months of age could distinguish /sa/ from /za/, whereas Eilers (1977) found that 3-month-olds could not distinguish /sa/ from /za/. However these younger infants could distinguish /as/ from /az/. Although Eimas and Tartter (1979) have suggested that Eiler's research may underestimate infant sensitiv-

This work was supported by grants from the Alberta Heritage Foundation for Medical Research, the National Health Research Development Program of Health and Welfare Canada, and the Natural Sciences and Engineering Research Council to D.G.J. and by an AHFMR Scholarship to D.E.M. We are grateful to Eleanor Rogers for providing testing facilities in Kingston and for helping to obtain subjects, to Carol McDermid for her assistance in providing subjects at Calgary, and to Fred Wightman and Terry Dolan for their hospitality at the Waisman Center, University of Wisconsin, where this work was completed while D.G.J. was a Visiting Fellow. Special thanks are due to Meg Cheesman, Terry Nearey, Curtis Ponton, and Mike Procter for advice and assistance throughout this project. Requests for reprints should be directed to Donald G. Jamieson, Speech and Audition Laboratory, Department of Psychology, University of Calgary, 2500 University Dr. N.W., Calgary, Alberta, T2N 1N4, Canada.

ity, these results at least suggest that the voiced/voiceless fricative distinction may be acquired within the first few months of life, presumably as a consequence of the child's linguistic environment.

Apparently, when individuals are fluently bilingual they can make speech distinctions that are appropriate to either language. Elman, Diehl, and Buchwald (1977) showed that fluent bilingual Spanish/English speakers classified natural /ba/ and /pa/ tokens differently, depending on the language in which the precursor phrase "please write the word" was spoken. More /b/ responses were given for intermediate VOT tokens (+15, +19, and +26 msec, respectively) following the phrase given in English than following the phrase given in Spanish. More typically, however, perceptual difficulties remain even after an individual becomes skilled in a new, non-native language. As one example, Florentine (1985) reported that listeners whose native language was other than English had particular difficulty perceiving English speech in noise, longer after they had become fluent English speakers. As another example, Flege and Hillenbrand (1986) reported that native speakers of English differ from native speakers of Swedish or Finnish in their ability to integrate multiple cues to voicing in the syllable-final fricatives /s/ and /z/. For example, Finnish speakers who were experienced speakers of English and had lived for some time in an English-speaking country were unable to use a decrease in the duration of frication as a cue to voicing of the final fricative.

Since some non-native speech contrasts appear to be remarkably difficult for adults to learn, it is not too surprising that there is a lengthy list of studies reporting failures to train non-native contrasts. Werker and Tees (1984b) initially found that the poor categorization of natural Thompson ejective (glottalized) velar and uvular stop sounds (/ki/ and /qi/) by unilingual English speakers was not improved by training these subjects with the acoustic cues which distinguish the sounds (the burst and initial transitions of the stimuli) in isolation from the remainder of the syllable. In a subsequent experiment, Werker and Tees (1984b) demonstrated that English adults could easily categorize both the initial, ejective portions of the Thompson /ki/-/qi/ sounds and the final, vocalic portions of these sounds, and that they could discriminate the sounds in a linguistically meaningful fashion, provided the sounds were presented with an interstimulus interval (ISI) of 500 msec. However, these listeners could not discriminate the sounds with an ISI of 1,500 msec and could not "categorize" the sounds by responding to a change from a sequence of repetitions of one of the full /ki/-/qi/ sounds to a sequence of the other (i.e., /ki/, /ki/, /ki/, . . . , /ki/, /qi/, . . . , /qi/; or the converse), when the sounds were presented with an ISI of 1,500 msec.

These difficulties with non-native speech contrasts may indicate that certain distinctions are extremely difficult for adults to learn, or even that adults cannot learn to make certain distinctions in a linguistically meaningful manner. Alternatively, and more optimistically, the difficulties may indicate that inadequate training techniques have been used. The latter view is encouraged by several studies that have successfully trained speech contrasts. In the first of these studies, Lane and Moore (1962) reestablished a /t/-/d/ voicing discrimination in an aphasic adult, using progressively more difficult stimuli throughout approximately 15 min of identification training, followed by a final few minutes of ABX discrimination training.

Carney, Widin, and Viemeister (1977) demonstrated that English-language listeners could be trained to classify synthetic speech sounds that differed in VOT, arbitrarily (and contrary to the classification used within their own language)—for example, as very prevoiced versus prevoiced/voiced, or to make intracategory VOT discriminations as well as intercategory discriminations at a consistently high level. Carney et al. used feedback with testing under conditions of minimum stimulus uncertainty—for example, in a fixed-standard AX paradigm with a brief (500-msec) ISI and the systematic presentation of comparison stimuli throughout a complete block of discrimination trials.

Repp (1981) demonstrated that training with isolated cues changes how familiar speech cues to fricative identity are processed by native speakers. He used truncated /s/ and /z/ fricatives matched with contradictory subsequent formant transitions (e.g., /s/ with the transitions from a /za/ sound). Initially, subjects tended to classify these sounds according to the identity of the formant transition information. After a period of discrimination training with isolated fricative information, the subjects could classify synthetic sounds on the basis of their fricative identity.

Pisoni, Aslin, Perey, and Hennessy (1982) demonstrated that simple exposure to phonetic prototypes could be used to establish three categories on a synthetic VOT continuum representing labial stops ranging in VOT from −70 msec to +70 msec. Giving subjects a simple labeling opportunity while they listened to three speech tokens typical of the three phonetic categories prevoiced (VOT = −70 msec), voiced (VOT = 0 msec), and voiceless (VOT = +70 msec) resulted in categorical discrimination performance and in reliable identification functions indicating the information of three phonetic categories. McClaskey, Pisoni, & Carrell (1983) also used identification training with exemplars of prevoiced, voiced, and voiceless stops to enhance a three-category classification of a synthetic voicing continuum in English-speaking subjects. This enhancement transferred to sounds synthesized at a second place of articulation which had not been trained in those subjects.

Most of the preceding studies did not examine whether training transferred to natural speech sounds. While none of the studies found successful transfer to natural speech, it is clear that, within the restricted set of synthesized or edited tokens, training can produce clear changes in the categorization and discrimination of non-native speech contrasts by adult listeners. It seems likely that at least some of the failures to train non-native contrasts reflect

deficiencies in the specific training method. We believe that three training principles are involved: (1) *Acoustic context*—Training should ensure that the relevant speech cues are presented in an acoustic context that is appropriate for normal speech, rather than in isolation. (2) *Identification training*—Since the desired outcome is to improve the listener's ability to classify speech sounds into the categories that are relevant for the new language, the listener's training task should involve identification (with feedback). Practice at discrimination, on the other hand, is likely to have the undesirable effect of enhancing sensitivity to within-category acoustic differences. (3) *Acoustic uncertainty*—Training should begin by focusing attention on the critically relevant cues and then introduce a range of acoustic variability in the signals. Each of these principles is discussed in turn.

The Importance of Acoustic Context

Because of the variety of intrasignal interactions, a particular acoustic cue often sounds very different when it is presented within a speech context from when it is presented in isolation. One expression of this notion is the demonstration that functions relating discrimination and identification with the isolated initial portions of consonant-vowel speech sounds to changes in the first 50 msec of a speech sound may be quite distinct from those observed when the same acoustic changes are followed by an unvarying (steady-state) vowel (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). To go further, Jusczyk, Smith, and Murphy (1981) found that their subjects could classify signals consisting of the initial 30 msec of a set of synthesized consonant-vowel (CV), stimuli as /b/ or /d/ consonants, as consistently as subjects who classified the entire CV stimuli. However, listeners performed poorly when discriminating between isolated formants or when presented with the initial portion of the first and second formants alone. Similarly, Werker and Tees (1984b) found that although their listeners could discriminate the initial portions of their Thompson /ki/-/qi/ distinction, they could not discriminate the full syllables. Moreover, Miyawaki et al. (1975) found that the isolated third formant differences which formed the basis for a synthetic, English /r/-/l/ distinction were discriminated in a similar fashion by both Japanese and English-speaking American subjects. However, when invariant first and second formants were added to the F_3 stimuli to approximate complete /r/ or /l/ tokens, only the English speakers could discriminate and categorize these same stimuli. Such results suggest that attempts to increase the salience of certain acoustic cues to a phonetic distinction by removing them from the speech context (e.g., Werker & Tees, 1984b) may, in fact, change the way the cue is heard. Consequently, although performance may be excellent when the cue is presented in isolation, performance may remain poor when the listener tries to *identify* the speech sounds. Speech training should therefore ensure that the acoustic signals presented during training are appropriate to the objective of improving performance on the task of accurately classifying the targeted sounds.

The Selection of an Appropriate Training Task

Some of the failures to improve listeners' identification of non-native speech sounds have used discrimination tasks with feedback to focus the subjects' attention on the acoustic differences between phonemes. Edman (1980) reported success with discrimination training in attempts to boost subjects' sensitivities to intraphonemic differences among synthetic stimuli on VOT continua and place of articulation continua. Using an AX discrimination paradigm, Edman reported general intraphonemic sensitivity gains and consistent transfer of this training effect to intraphonemic discrimination performances on nontrained continua. Although subjects often learn such tasks, discrimination training rarely improves the categorization of nonnative speech contrasts, even when those contrasts are modifiable by other techniques. Strange (1972) reported that training a prevoiced-voiced VOT discrimination using a randomized fixed-step oddity task failed to improve identifications of the same sounds. Also, Strange and Dittmann (1984) used an all-step, fixed-standard, AX discrimination procedure with immediate feedback to train /r/-/l/ categorizations by Japanese subjects. They characterized the improvement seen in discrimination and identification scores as "slow and effortful." The subjects showed limited transfer to a second synthetic minimal pair but failed to show transfer to natural tokens. The doubtful utility of discrimination training seems explainable in light of Samuel's (1977, 1982) work on phonetic prototypes. Samuel has argued that speech recognition is performed through a process whereby stimuli are matched to phonetic prototypes focused at particular points along the salient acoustic dimensions. Incoming stimuli are categorized according to their relative distances from alternative prototypes. One source of support for this position is that subjects trained on a VOT continuum using an ABX task show increased sensitivity between phonemic categories and within categories near the phonetic boundary, but that VOT differences at values near the phonetic prototype (i.e., category center) remain undetectable (Samuel, 1977).[2] In this view, acoustic differences near the prototypical values are perceptually assimilated by the dominating prototype construct at these values. It should be noted, however, that discrimination training tends to increase intraphonemic sensitivities, which in normal phonetic development may necessarily be processed as perceptually irrelevant. Such increased within-category sensitivity could diminish attempts to form a new phonetic category.

Role of Uncertainty

In contrast to the work that preceded theirs, Carney et al.'s (1977) research demonstrated the ease with which stop consonant VOT discrimination and identification

could be modified in listeners by simple training. These authors contrasted their training, in which uncertainty was reduced for listeners through the use of fixed standards and systematic presentation of stimuli, with the earlier failures which used random sequences of speech sounds. Similarly, Pisoni et al. (1982) trained listeners without introducing uncertainty into the training set. However, when the objective of training is to improve performance with natural speech, listeners must, at some point, learn to ignore the within-category acoustic variability that occurs in natural speech, while attending to the relevant between-category variability. Such learning can occur through training with multiple natural tokens, as in Tees and Werker (1984), or through training with a variety of synthetic tokens, as here. However, it is unlikely that training a difficult speech contrast without the inclusion of such variability will prepare the listener to deal with natural speech tokens.

The combination of an initial reduction in uncertainty, through the systematic sequencing of stimuli, followed by increased stimulus uncertainty in later stages of training is at the root of the perceptual fading technique introduced by Terrace (1963). This technique attempts to train a perceptual contrast, without subject errors, by beginning with clearly discriminable stimuli which may exaggerate the normal perceptual differences or add other salient features. Progress in training is made by slowly reducing the magnitude of the perceptual contrast, in small steps, so that the task never becomes too difficult and errors remain infrequent. Using this technique, a high level of identification and discrimination performance can be attained in a short interval of time, without frustrating the subject.

The current work focuses on the training of the /ð/-/θ/ contrast for francophone adults. This contrast is not used in French, and it appears to be remarkably difficult for many francophones to acquire. We sought specifically to determine whether training with a continuum in which stimuli varied only in the duration of voiced or voiceless frication would improve performance. We used a variation of the fading technique, with synthetic prototypes as exemplars. Our training began by requiring the identification of two very distinct exemplars—one voiceless and one voiced fricative. During training, we then systematically increased the amount of intraphonemic variation by adding new stimuli, one at a time. These new sounds were progressively less salient exemplars of the voiced and voiceless tokens.

METHOD

Stimuli

Twenty-four consonant-vowel syllables (CVs) were used throughout the experiment. Sixteen were natural tokens, spoken by a single male talker; eight of these tokens were voiced /ð/ and eight were voiceless /θ/ tokens. The remaining eight stimuli were synthesized in cascade at a 10-kHz sample rate, using Klatt's (1980) cascade/parallel speech synthesizer (Kewley-Port, 1978) implemented on a Vax 11/730 computer system, manufactured by the Digital Equipment Corporation. The four voiced, /ð/, CVs began with a fricative sound generated at formant center frequencies of 295, 1220, and 2540 Hz using source generator settings of AV=30, AS=45, and AF=30. The four stimuli differed in the duration of frication, decreasing from 140 to 35 msec of frication in 35-msec steps, for stimuli designated numbers 8 through 5, respectively. At the termination of frication, there was a voiced 35-msec transition to a 175-msec /ʌ/ vowel, which was synthesized using formant center frequencies of 620, 1220, and 2550 Hz. This combination yielded voiced fricative stimuli with total durations of 350, 315, 280, and 245 msec, respectively, for stimuli numbered 8, 7, 6, and 5.

The voiceless fricatives, /θ/, were synthesized with formant center frequencies at 295, 1290, and 2540 Hz and source settings at AF=50 and AH=30. Other aspects of the stimuli, including frication durations, were also identical, yielding voiceless fricative stimuli of 350, 315, 280, and 245 msec (stimuli numbered 1 through 4, respectively). Figure 1 presents oscillograms and spectrograms displaying the variations of the critical acoustic cues for the synthetic tokens.

Natural speech tokens were recorded in a double-walled, IAC sound-attenuating chamber, using an AKG C451 EB condenser microphone and a Revox B710 MkII recorder. Sixteen tokens were selected from a larger number of stimuli, produced by the same speaker, to ensure a substantial range of variability in frication duration. Figure 2 presents spectrograms of these natural tokens. All tokens were consistently identified by three Canadian anglophones, and each was judged, by those listeners, to be a good exemplar of that particular fricative category. The natural sounds were lowpass filtered at 4800 Hz and digitized at 10 kHz using a 12-bit analog-to-digital converter installed within the LPA subsystem of a VAX 11/730 computer and stored on disk. Stimuli were output in the desired orders, lowpass filtered at 4800 Hz, amplified (Crown D-75), and recorded on tapes for use in the experiment.

Stimulus tapes were produced on a Revox B710 MkII recorder using Maxell XLII tapes. Background "cafeteria" noise was recorded on a separate channel for later mixing. Subjects reported that they realized that this noise contained human voices, but they were unable to hear any message. Where noise was used, the noise level was 57 dB SPL. All stimuli were presented binaurally to subjects at a level of 70 dB SPL using a Sony TC PB5 playback system to drive AKG 240 headphones.

Subjects

Twenty Canadian francophone subjects were selected from approximately 180 students participating in the Queen's University's summer school of English. All met the following criteria: (1) They scored below the school's 50th percentile (<74) on the English Placement Test (English Language Institute, 1972); (2) they declared French to be their mother tongue; and (3) they held current residency in the province of Quebec. Ten male and 10 female subjects, aged 18 to 32, participated.

All subjects were paid for participation. They were initially asked to participate for 2 h. "Pretesting" was completed during the first hour, after which the 20 students were matched into 10 pairs of subjects on the basis of their error rates in identifying the natural tokens of the voiced and voiceless stimuli. Students assigned to the treatment group were then asked whether they would participate for a total of 4 h (i.e., for an additional 2 h) at the same rate of pay. In no case did a subject refuse this offer. Only at this point was any subject informed of the study's objective of examining language learning.

Procedure

On the first day of testing, all subjects received a pretest consisting of an identification task followed by a discrimination task. The subjects were rank-ordered by their scores on this pretest and then

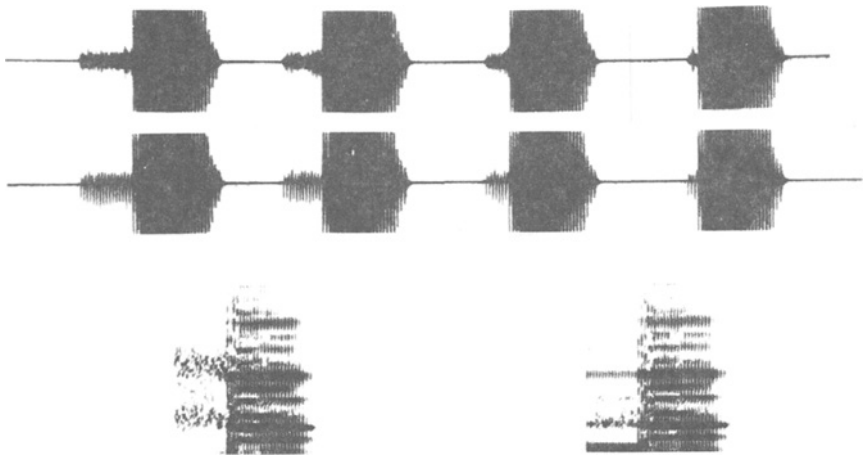

Figure 1. Acoustic representations of the synthetic tokens used for training and for testing. Waveforms for the four voiceless stimuli, displayed in the top row, show the reduction of the duration of voiceless frication from Stimulus 1 to Stimulus 4. Waveforms for the four voiced stimuli, displayed in the second row, show the reduction in the duration of voiced frication from Stimulus 8 to Stimulus 5, respectively. The bottom-left portion of the figure displays a spectrogram of Stimulus 1—the synthetic voiceless token having the greatest duration of voiceless frication. The bottom-right portion of the figure displays a spectrogram of Stimulus 8—the synthetic voiced token having the greatest duration of voiced frication.

assigned to the control or training group, in random alternation, to form two groups with equivalent pretest scores on the task involving the identification of natural tokens. On 4 subsequent days, subjects in the training group received two training sessions followed by a posttest of the same form as the pretest. Control group subjects received the pretest and posttest, using identical procedures and stimulus tapes, at intervals equivalent to their trained, matched counterparts, but they received no experimental training. All subjects continued to participate in their English immersion course throughout the experiment. All testing and training occurred within a 20-day period. Control subjects were never posttested more than 1 school day before their matched, trained counterparts, and they were normally posttested 1 or more days after their counterparts. Pretests were completed between Day 1 and Day 4. The two training sessions were completed between Days 7 and 10 and Days 9 and 13, respectively, and posttesting took place between Days 11 and 20. At least 24 h separated any two sessions for a subject.

Pretesting. The pretest comprised an identification task and a discrimination task. In the identification task, a subject was presented with a randomized sequence of the 24 stimulus tokens (8 synthetic sounds and 16 natural sounds). Following each presentation, the subjects indicated whether the sound was voiced or unvoiced by circling the word "the" (for voiced) or the word "teeth" (for unvoiced). During each of the pretest and posttest sessions, each of the 24 stimuli was presented 12 times—three times in each of four blocks of 72 trials. Within blocks, items were presented in a completely randomized order, at a rate of one every 4 sec. The entire identification sequence required approximately 25 min. The subjects were tested in a quiet room, either individually or in pairs.

For the discrimination task, the subjects were then told that pairs of the computer sounds would be presented with a cafeteria-noise background. The subjects were encouraged to listen for very small differences between the stimulus pairs and to respond "same" only if the stimuli were exactly identical. Written responses of "same" or "different" were made by circling "=" or "≠" for each stimulus pair. Fifteen stimulus pairs were used: eight identical pairs (each of the eight stimuli from the synthetic continuum, presented twice in succession) and seven different pairs (each of the stimuli 1 to 7, followed by the next stimulus in the sequence (i.e., the next most voiced, producing pairs 1-2, 2-3, 3-4, 4-5, 5-6, 6-7, and 7-8).

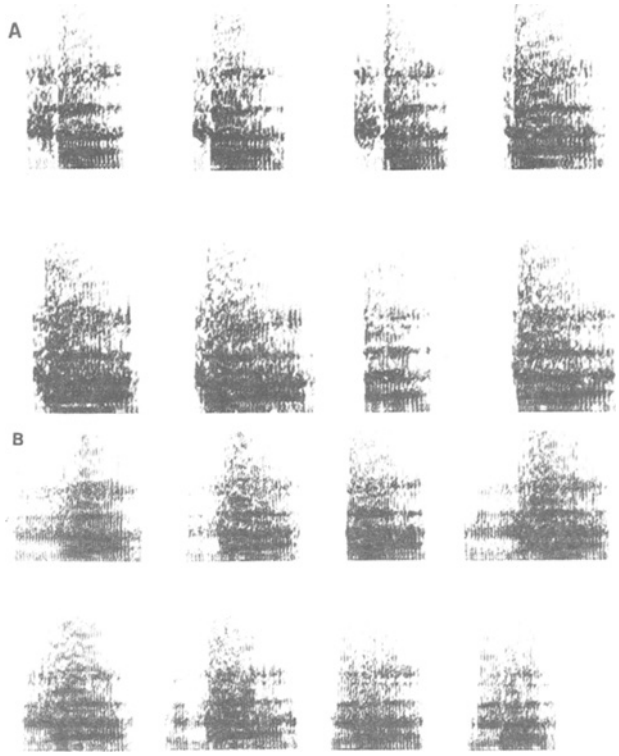

Figure 2. Spectrograms of the 16 natural speech tokens ordered as in Table 2. Figure 2A displays the eight voiceless /θ/ tokens, ordered from left to right within rows, by the frequency with which they were correctly identified. The four tokens most often correctly identified as "voiceless" are displayed in the upper panel; those that were less well identified are displayed in the lower panel. Figure 2B displays the eight voiced /ð/ tokens ordered from right to left within rows by the frequency with which they were correctly identified. The four tokens most often correctly identified as "voiced" are displayed in the lower panel; those that were less well identified are displayed in the upper panel.

During each of the pretest and posttest sessions, each of the 15 stimulus pairs was presented 12 times—3 times in each of four blocks of 45 trials each. Within blocks, the pairs were presented in a completely randomized order, at a rate of 1 pair every 6 sec. Stimuli within a pair were separated by an onset asynchrony of 850 msec. The entire discrimination test, which followed the identification test described above, required approximately 20 min.

Testing was preceded by English verbal instructions as to how the subjects were to make their responses. Differences in language competence introduced some variability in the extent to which the task was explained to subjects, beyond the standard instructions. However, the task did not proceed until the experimenter was satisfied that each subject understood the experiment and the task.

Identification training. Identification training consisted of a series of trials in which one synthetic token was presented, after which the subject was required to identify the stimulus by pressing one of two microswitches labeled "teeth" and "the," respectively. Feedback was then provided to subjects through the immediate illumination of a small white light situated 7 cm above the response buttons whenever an incorrect response was made.

A series of 12 training tapes were used, each consisting of a sequence of identification trials. Trials were grouped into blocks of 20, with stimuli presented once every 4 sec within a block. The 12 tapes formed a sequence for a modified fading technique for perceptual training (cf. Terrace, 1963), beginning with the relatively easy task of identifying the most extreme stimuli of the continuum (i.e., 1 and 8), on Tape 1. For subsequent tapes, the task became more difficult as more medial stimuli from the continuum were introduced into the set of stimuli presented for identification. Thus, the number of stimuli included on each tape increased from just the two most extreme stimuli (1 and 8) to the entire set of eight sounds over the first 10 tapes. These training tapes contained the synthesized stimuli alone, without the cafeteria-noise background distraction. The cafeteria-noise background was introduced in Tapes 11 and 12, with four stimuli (1, 2, 7, and 8) and six stimuli (1, 2, 3, 6, 7, and 8), respectively.

The stimuli presented on each tape were randomized within the 20-trial blocks in unequal proportions so that at least 50% of the trials were drawn from the two most medial stimuli present. Each tape was used for at least three blocks of trials with a subject (a maximum of eight blocks were available on each tape); subjects advanced to the next more difficult tape when they had completed three consecutive blocks of a tape (i.e., 60 trials) with not more than one error per block.

Training continued for approximately 45 min on Day 2 and concluded on Day 3 when all 12 tapes had been mastered. Only 1 of the 10 subjects failed to reach criterion before the end of the second session. This subject was therefore the only to receive the full 90 min of training.

Posttesting. Identification and discrimination posttesting was completed on Day 4. Testing used a different tape, with different randomization sequences, for both identification and discrimination tasks. Other aspects of the procedure were identical to those used during pretesting.

RESULTS AND DISCUSSION

Separate analyses were performed, in turn, for the identification task with the synthetic and natural stimuli, respectively, and for the discrimination task. The results of these analyses are discussed in sequence below.

Identification of Synthesized Tokens

Identification data were first analyzed by calculating the proportion of correct identification responses given by each listener with each stimulus in each condition. Table 1 summarizes these data, separately for the pretest and posttest measures, for the control and training groups. As the entries in Table 1 show, subjects in both groups displayed a strong bias for "voiced" responses in their pretest responses. Post hoc tests confirmed that voiced stimuli were identified more accurately than voiceless stimuli [$F(1,72) = 16.22, p < .01$, and $F(1,72) = 56.49$, $p < .01$, for the trained and the control groups, respectively].

Table 1
Proportion of Correct Identifications for Each Synthetic Stimulus as a Function of Training Condition and Time of Test

		Stimulus							
Group	Test	1	2	3	4	5	6	7	8
Control	Pre	.45	.33	.40	.41	.83	.81	.80	.78
	Post	.51	.43	.37	.38	.89	.89	.85	.87
Trained	Pre	.48	.49	.49	.52	.71	.78	.72	.77
	Post	.96	.97	.86	.83	.88	.97	.94	.97

To allow changes in sensitivity to be measured independently of such biases, identification responses were converted to A′ scores (see McNichol, 1972) using each subject's hit rate with a given stimulus, in combination with that subject's overall error rate on all stimuli of the opposite type (e.g., "voiced" responses with voiceless stimuli) as the false-alarm rate. Figure 3 shows that training improved listeners' identification of both voiced and voiceless tokens, but that the control group did not improve from pretest to posttest. Figure 4 compares the improvement produced by training, with the nonsignificant improvement in identification accuracy found with control subjects, separately for each of the eight synthetic tokens.

A′ scores for the control and trained groups were submitted to separate 2×8 repeated measures analyses of variance to determine whether pretest and posttest scores differed for the eight synthetic stimuli. For the trained group, the ANOVA confirmed that posttest scores were higher than pretest scores [$F(1,9) = 32.00, p < .01$]. Stimuli did not differ in overall identifiability [$F(7,63) = .44, p > .05$], and there was no interaction between stimulus and time of test [$F(7,63) = .17, p > .05$]. For the control group, the ANOVA confirmed that posttest scores did not differ from pretest scores [$F(1,9) = 1.47$, $p > .05$], that no significant identifiability variance existed among stimuli [$F(7,63) = .70, p > .05$], and stimuli did not interact with time of test [$F(7,63) = .60, p > .05$]. Post hoc tests confirmed a significant improvement in performance with each stimulus for each subject in the trained group (minimum $t = 3.76$, $p < .0045$; Duncan's multiple range tests confirm $p < .05$ for all stimuli). For subjects in the control group, performance did not improve from pretest to posttest for any stimulus (maximum $t = 1.46, p > .176$; Duncan's tests show significant improvement at Stimulus 2 only: $W = 8.63, p > .05$).

Identification of Natural Tokens

As Table 2 shows, the 16 individual tokens of the natural stimulus set covered a substantial range of iden-

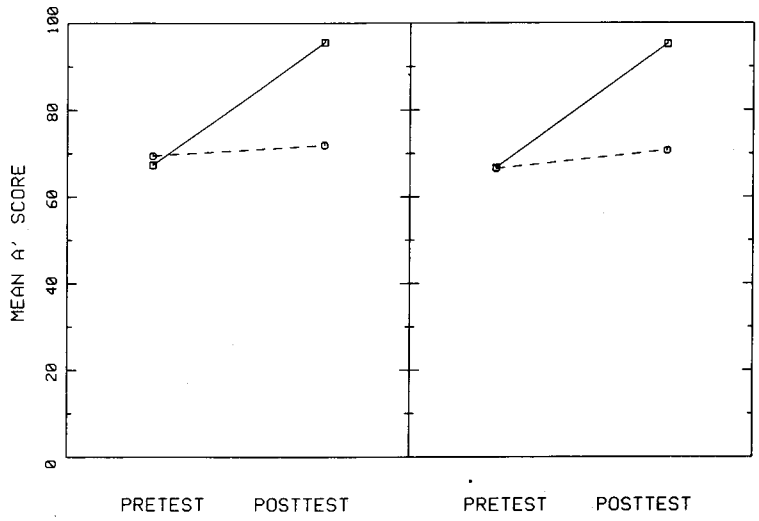

Figure 3. Comparison of pretest and posttest identification performance for voiced and voiceless synthesized sounds for listeners in the control group (broken line) and the training group (solid line), respectively. Each point represents the mean of 40 A′ scores, collapsed across the four tokens within a stimulus set and the 10 listeners within each group.

tifiability for the francophone listeners in both groups. Identification scores were converted to A′ scores to control for the effects of response bias prior to statistical analysis. Figure 5 compares the improvement produced by training with the nonsignificant improvement in identification accuracy found with control subjects, separately for each of the 16 natural tokens. Stimuli in Figure 5 are arranged by difficulty, with stimuli towards the extreme left (voiceless sounds) and extreme right (voiced sounds) being those that were best identified during pretesting, and those toward the center being least well identified. It is clear that although performance improves very generally with training, improvement is greatest for the items that were least well identified initially.

A′ scores were submitted to separate 2×16 repeated measures analyses of variance for each group to determine whether pretest and posttest scores differed for the 16 natural stimulus tokens. ANOVA confirmed that A′ increased after training [$F(1,9) = 8.44$, $p < .05$]. Individual tokens differed significantly in their identifiabil-

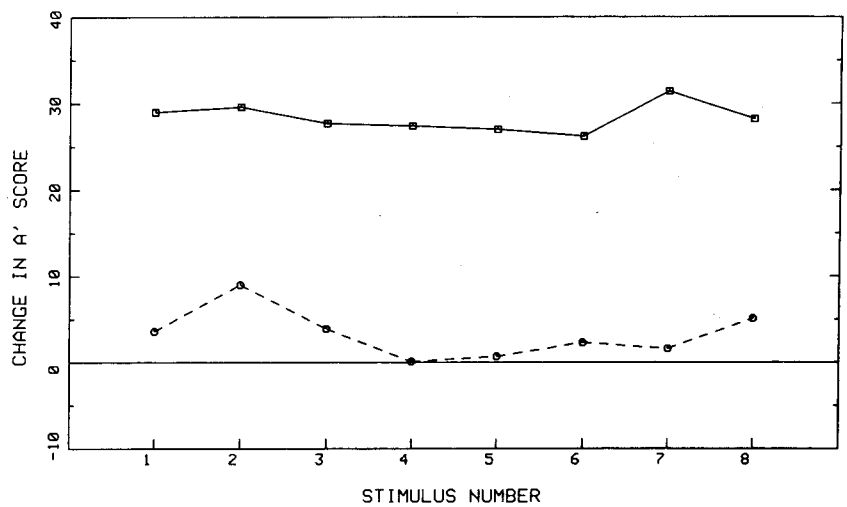

Figure 4. Change in the ability to identify individual synthetic speech tokens from pretest to posttest for listeners in the control group (broken line) and the training group (solid line), respectively. Each point represents the mean of 10 A′ difference scores (posttest A′ score minus pretest A′ score), collapsed across the listeners within each group.

Table 2
Proportion of Correct Identifications for Each Natural Token as a Function of Training Condition and Time of Test

		Tokens															
Group	Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Control	Pre	.77	.70	.73	.53	.46	.50	.43	.34	.68	.70	.75	.74	.79	.82	.80	.80
	Post	.87	.83	.83	.64	.66	.56	.52	.42	.76	.78	.57	.77	.83	.80	.81	.82
Trained	Pre	.76	.82	.74	.70	.58	.48	.48	.36	.73	.75	.72	.73	.70	.75	.82	.83
	Post	.89	.83	.87	.79	.74	.68	.59	.48	.89	.83	.58	.92	.93	.88	.88	.92

ity [$F(15,135) = 2.37, p < .05$] and in the amount by which identification performance improved [$F(15,135) = 2.82, p < .05$].

Pretest and posttest identification scores did not differ for control subjects [$F(1,9) = .74, p > .05$]. Control subjects showed different performance across stimuli [$F(15,135) = 3.59, p < .05$], but there was no interaction between stimuli and time of test [$F(15,135) = .87, p > .05$]. Post hoc tests confirmed a significant improvement in the identification of individual stimulus tokens by subjects in the trained group [t tests yielded $p < .05$ for all stimuli except Stimuli 2, 3, and 11; Duncan's tests show improvements for all stimuli ($p < .05$) except 2 and 11]. For subjects in the control group, performance improved from pretest to posttest for only two stimuli [t tests yielded no significant comparisons (maximum $t = 1.62, p > .07$; Duncan's tests showed improvements ($p < .05$) for Stimuli 5 and 10].

In sum, training with synthetic sounds, using a procedure that initially focused attention on a single acoustic distinction and subsequently introduced increasing amounts of "irrelevant" acoustic variability into the stimulus set, was successful. Identification performance improved with the synthetic stimuli that had been used during training and, more importantly, also improved with a variety of natural speech tokens. The training improved the subjects' ability to identify both voiced and voiceless fricatives in each case.

Discrimination of Synthesized Tokens

A′ scores were calculated for each subject for each pair of stimulus tokens, using, as the hit rate, the proportion of correct ("different") responses with a different stimulus pair and, as the false-alarm rate, the proportion of "different" responses when the first stimulus of that pair was repeated. Figure 6 displays the mean A′ values for each of the seven pairs of different stimuli obtained during pretesting and posttesting. Note that Stimuli 1 to 4 (hence pairs 1-2, 2-3, 3-4) are all voiceless stimuli with decreasing durations of the voiceless frication, respectively, while Stimuli 5 to 8 (hence pairs 5-6, 6-7, 7-8) are all voiced stimuli with increasing durations of the voiced frication, respectively. The pair 4-5, on the other hand, contains the stimulus with the least amount of voiceless frication (4) in combination with the stimulus with the least voiced frication (5).

We tested the hypothesis that training would increase linguistically relevant discrimination by comparing the pre- and posttest A′ values for each stimulus pair. A one-tailed dependent t test showed that the A′ score increased

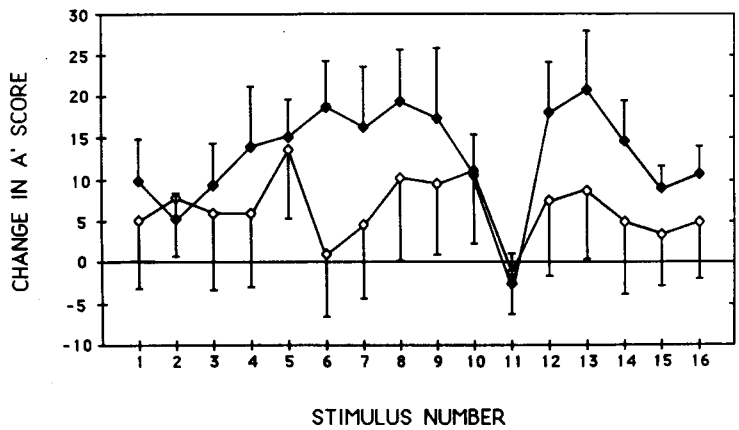

Figure 5. Change in the ability to identify individual natural speech tokens from pretest to posttest for listeners in the control group (empty symbols; lower curve) and the training group (filled symbols; upper curve), respectively. Each point represents the mean of 10 A′ difference scores (posttest A′ score minus pretest A′ score), collapsed across the listeners within each group. Stimuli 1 to 8 are voiceless tokens, ordered from the stimulus that was best identified during pretesting (i.e., Stimulus 1) to that which was most poorly identified (i.e., Stimulus 8). Stimuli 9 to 16 are voiced tokens, ordered from the stimulus that was most poorly identified during pretesting (i.e., Stimulus 9) to the one that was best identified (i.e., Stimulus 16). Error bars illustrate standard error values obtained in dependent t tests.

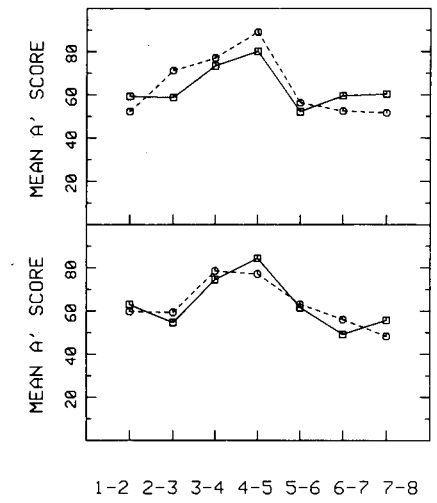

Figure 6. Comparison of listeners' ability to discriminate each of the seven pairs of different stimuli during the pretest (dashed line) and posttest (solid line), respectively. Each point represents the mean A' score, averaged across all 10 listeners in the control group (upper panel) or the trained group (lower panel). Note that pairs 1-2, 2-3-, 3-4, 5-6, 6-7, and 7-8 all require within-category discriminations based on differences in the duration of frication. The comparison of Stimulus 4 (voiceless) with Stimulus 5 (voiced) is quite different, since it requires the distinction between 35 msec of voiceless frication and 35 msec of voiced frication; this pair is thus the only one that requires a between-category comparison.

after training [$t(9) = 2.43, p < .02$ for the pair 4-5]. No other discrimination score, for either group, showed a significant increase from the pretest to the posttest, according to post hoc two-tailed t tests, even when a very liberal alpha-error level (.05) and compound alpha level (.49) were set. Thus, training was effective in improving performance on the most difficult discrimination between voiced and voiceless fricatives (i.e., the discrimination between Stimulus 5—the voiced token having the least amount of voiced frication—and Stimulus 4—the voiceless token having the least amount of voiceless frication). However, no similar change occurred for the control group. Importantly, training did not change within-category discrimination. Since accurate speech perception requires subjects to make accurate discriminations between phonemes on the basis of minimal cues while failing to discriminate between allophones on the basis of other (even large) nonphonemic acoustic differences, this result demonstrates that the training task was successful in improving discrimination in a linguistically relevant fashion.

GENERAL DISCUSSION

Our results are encouraging for attempts to train adult listeners to identify and discriminate non-native speech sounds. Just 90 min of practice in identifying synthetic speech tokens from a structured continuum improves the identification of both synthetic and natural speech sounds and intercategory discrimination while leaving intracategory discrimination unchanged. Most importantly, the training with synthetic tokens transferred strongly to the identification of natural tokens that had not been trained. This result therefore significantly extends the several previous demonstrations that certain non-native VOT identifications can be trained.

Identification performance with the synthetic continuum shows that the continuum endpoints (i.e., those with the longest duration of voiceless and voiced frication) were learned almost perfectly. For the more medial positions, with briefer frication, identification improved to a lesser level of accuracy. Performance improved for both voiced and voiceless stimulus types and for both synthetic and natural speech. The improvement shown in natural tokens is especially encouraging in light of the general paucity of previous successful attempts to achieve such transfer. The generalizability of this technique to natural speech can be examined in future studies which use tokens spoken by several different speakers, both male and female. Finally, since the control group showed remarkable consistency from pretest to posttest, the performance improvements shown by the trained subjects cannot be attributed to factors such as practice in the experimental conditions or the influence of generalized language training.

Given the frequency of common English words with an initial /th/ sound (e.g., they, this, there, their, that, then, the), it was possible that the brief incidental training which alerted listeners to the /ð/-/θ/ distinction in the pretest would lead to substantially improved performance for the control subjects. Clearly, however, neither this exposure nor the substantial daily exposure to spoken English in the immersion program improved performance with this specific non-native contrast; rather, more specific, directed intervention was required for the contrast to be acquired.

The fading technique begins by providing two distinct prototypes which highlight the acoustic factors for subjects. Once subjects are able to use these prototypes, other sounds are introduced in a systematic manner so that the listener learns to deal with the intraphonemic variability. Throughout this phase of training, the original prototype stimuli continued to be presented. Although Pisoni et al. (1982) have demonstrated that, for some contrasts, a very brief (< 15-min) orientation to prototypes, together with training using "prototypical" stimuli from the endpoints of a continuum, may be sufficient to produce good identification and discrimination throughout the synthetic continuum, a training set with additional acoustic variability may be required for learning to transfer to natural speech.

Discrimination training may fail to increase identification performance because the task causes listeners to focus on differences between stimuli, including differences within a phonetic group, rather than grouping stimuli in terms of their similarity to a prototype. Moreover, discrimination training normally presents different sounds from trial to trial. The use of a fixed standard from a single category for a block of trials, with comparison stimuli selected from throughout the alternate category, permits

the listener to develop prototypes and increases the tolerance of intraphonemic variability, since the comparison stimuli can vary from each other and from trial to trial on several dimensions at once. The systematic, organized presentation of training stimuli using a fixed standard in an all-step discrimination procedure reduces this problem and can be partially successful (e.g., Strange & Dittman, 1984), but this technique is slow and it does not seem to lead to reliable transfer to the identification task.

Identification training needs to guard against establishing a phonemic percept according to the idiosyncratic characteristics attended to during the initial exposure to the stimuli, as well as to ensure that acoustic variation is sufficient to permit the dimensions of acceptable intraphonemic variability to be "inferred." These problems become more tractable when the training technique uses stimuli that have been synthesized to emphasize these features.

Previously, we had attempted to train the /θ/-/ð/ distinction in Canadian francophone subjects using a continuum that contained the most extreme voiceless and voiced stimuli from the present experiment (1 and 8, respectively), but with the intermediate stimuli constructed by mixing both voiced and voiceless frication (Morosan & Jamieson, 1986). Training with this continuum failed to improve either identification or intercategory discrimination performance. We believe that the technique was not successful because it did not permit subjects to attend to the appropriate dimensions of variability, within and between phoneme categories. Thus, although subjects completed the training identification tasks above criterion levels, the systematic exposure to these stimuli did not develop a more categorical distinction.

This generalizability of the present study is limited by at least two factors. First, although the /ð/-/θ/ distinction is not phonemic in French, all Canadian francophones are exposed to these sounds through incidental exposure to English, especially in urban Quebec.[3] Perhaps as a consequence of such experience many of our listeners performed at above-chance levels during pretesting. Our success with the fading technique might not generalize to contrasts with which listeners have had no previous experience. Second, our voiced/voiceless distinction was quite straightforward, since the relevant variation lay along a single dimension. It will be a greater challenge to use these procedures to train a distinction for which several dimensions varied simultaneously.

One type of irrelevant acoustic variability used in this experiment was the duration of the frication portion of the signal. Within-category duration differences were used explicitly for training with the synthetic continuum, and, as an examination of Figure 2 reveals, the variation among the natural tokens selected for use in the experiment occurred primarily in the duration of the frication contained in each token and in the duration of the vowel subsequent to frication. In general, the voiced natural tokens used in the present study were longer than the voiceless natural tokens. While the duration of frication provides a cue to fricative voicing, in the present experiment duration did not provide a systematic cue to voicing category for the synthetic stimuli, either during training or during testing. Moreover, listeners do not appear to have used duration as a cue for the natural tokens, since duration does not change systematically when the stimuli are ordered either by their initial identifiability or by the amount of their posttraining identification gain scores. Second-language learners may be able to learn to make good use of such secondary voicing cues as duration of frication through explicit training procedures, but they do not appear to do so without such training. For example, the results reported by Flege and Hillenbrand (1986) demonstrate that native speakers of Finnish do not use fricative duration as a cue to voicing in syllable-final position, even after they have become experienced speakers of English. It thus remains for future work to determine whether such cues can be trained using procedures such as those described in the present paper.

CONCLUSIONS

The present study indicates that a brief, structured identification training program can be used to train adult francophones to identify both natural and synthetic tokens of English voiced, /ð/, and voiceless, /θ/, fricatives with a high degree of accuracy. This result contrasts with the difficulty that adult francophones have in acquiring these sounds either through participation in more typical language-learning programs, or even using another structured training sequence which mixed voice and voiceless frication within stimuli. It is argued that three principles offer a reliable guide to effective training: (1) Training should ensure that the relevant speech cues are presented in an acoustic context that is appropriate for normal speech, rather than in isolation; (2) the training task should involve identification with feedback rather than practice at discrimination; and (3) training should begin by focusing attention on the critically relevant cues, and then introduce a range of acoustic variability in the signals, to teach the listener to deal with within-category acoustic differences.

REFERENCES

CARMAZZA, A., YENI-KOMSHIAN, G. H., ZURIF, E. B., & CARBONE, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America*, **54**, 421-428.

CARNEY, A. E., WIDIN, G. P., & VIEMEISTER, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, **62**, 961-970.

EDMAN, T. R. (1980). *Learning of intra-phonemic discrimination for several synthetic speech continua*. Doctoral dissertation (University of Minnesota, University Microfilms No. 81-09, 419).

EILERS, R. E. (1977). Context sensitive perception of naturally produced stop and fricative consonants by infants. *Journal of the Acoustical Society of America*, **61**, 1321-1336.

EILERS, R. E., WILSON, W. R., & MOORE, J. M. (1977). Developmental changes in speech discrimination in infants. *Journal of Speech & Hearing Research*, **20**, 766-780.

EIMAS, P. D., & TARTTER, V. C. (1979). On the development of speech perception: Mechanisms and analogies. In L. P. Lipsitt & H. W. Reese (Eds.), *Advances in child development and behavior* (Vol. 13). New York: Academic Press.

ELMAN, J. L., DIEHL, R. L., & BUCHWALD, S. E. (1977). Perceptual switching in bilinguals. *Journal of the Acoustical Society of America*, **62**, 971-974.

ENGLISH LANGUAGE INSTITUTE (1972). *English Placement Test*. Ann Arbor, Michigan: University of Michigan, Test and Certification Division.

FLEGE, J. E., & HILLENBRAND, J. (1986). Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. *Journal of the Acoustical Society of America*, **79**, 508-517.

FLORENTINE, M. (1985). Speech perception in noise by fluent, non-native listeners. *Journal of the Acoustical Society of America*, **77**(Suppl. 1), S106.

JUSCZYK, P. W., SMITH, L. B., & MURPHY, C. (1981). The perceptual classification of speech. *Perception & Psychophysics*, **30**, 10-23.

KEWLEY-PORT, D. (1978). KLTEXC: Executive program to implement the Klatt software speech synthesizer (Progress Report 4, *Research on speech perception*). Bloomington: Indiana University.

KLATT, D. H. (1980). Software for a cascade-parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.

LANE, H. L., & MOORE, D. J. (1962). Reconditioning a consonant discrimination in an aphasic: An experimental case history. *Journal of Speech & Hearing Disorders*, **27**, 232-243.

LASKY, R. E., SYRDAL-LASKY, A., & KLEIN, R. E. (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, **20**, 215-225.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. G. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

LISKER, L., & ABRAMSON, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the Sixth International Congress of Phonetic Sciences, 1967*. Prague: Academia.

MCCLASKEY, C., PISONI, D. B., & CARRELL, T. D. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception & Psychophysics*, **34**, 323-330.

MCNICHOL, D. (1972). *A primer of signal detection theory*. London: George Allen & Unwin.

MIYAWAKI, K., STRANGE, W., VERBRUGGE, R., LIBERMAN, A. M., JENKINS, J. J., & FUJIMURA, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, **18**, 331-340.

MOROSAN, D. E., & JAMIESON, D. G. (1986). Efficient training of non-native speech discriminations with perceptual fading. *Proceedings of the 12th International Congress on Acoustics, 1986*. Canadian Acoustical Association.

PISONI, D. B., ASLIN, R. N., PEREY, A. J., & HENNESSY, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception & Performance*, **8**, 297-314.

REPP, B. H. (1981). Two strategies in fricative discrimination. *Perception & Psychphysics*, **30**, 217-227.

SAMUEL, A. G. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics*, **22**, 321-330.

SAMUEL, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307-314.

STRANGE, W. (1972). *The effects of training on the perception of synthetic speech sounds: Voice onset time*. Doctoral dissertation, University of Minnesota. (University Microfilms No. 73-01,066)

STRANGE, W., & DITTMANN, S. (1984). Effects of discrimination training in the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, **36**, 131-145.

TEES, R. C., & WERKER, J. F. (1984). Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology*, **38**, 579-590.

TERRACE, H. S. (1963). Discrimination learning with and without "errors." *Journal of Experimental Analysis of Behavior*, **6**, 1-27.

WERKER, J. F., GILBERT, J. H., HUMPHREY, K., & TEES, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, **52**, 349-355.

WERKER, J. F., & TEES, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology*, **37**, 278-286.

WERKER, J. F., & TEES, R. C. (1984a). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, **7**, 49-63.

WERKER, J. F., & TEES, R. C. (1984b). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, **75**, 1866-1878.

WILLIAMS, L. (1977). The perception of stop consonant voicing by Spanish-English bilinguals. *Perception & Psychophysics*, **21**, 289-297.

WILLIAMS, L. (1979). The modification of speech perception and production in second-language learning. *Perception & Psychophysics*, **26**, 95-104.

NOTES

1. Thompson is a northwest Canadian (interior Salish) language spoken in the interior of British Columbia.

2. An alternative interpretation of this result relates discriminability to the slope of the psychometric function relating identification performance to VOT.

3. However, our subjects came primarily from rural areas of Quebec, and had little knowledge of English.

(Manuscript received March 17, 1986;
revision accepted for publication June 30, 1986.)