GERNALIZATION OF NEW SPEECH CONTRASTS TRAINED USING THE FADING TECHNIQUE

Donald G. Jamieson and April E. Moore

University of Western Ontario, London, Ontario, Canada

ABSTRACT

Subsequent to training with synthetic speech using a fading technique, we found substantial improvements in the ability of unilingual francophone adults to identify voiced and voiceless English "th" sounds, presented in a vowel-consonant-vowel (VCV) format. Improvements were seen for the tokens used in training, for natural utterances by two speakers, and when listening in noise as well as in quiet. Smaller improvements occurred for target sounds in other word positions (ie., VC or CV context) or in other vowel environments.

1. INTRODUCTION

While adults may develop a sophisticated command of a new language, difficulties often persist with the perception and pronunciation of phonemes which are foreign to the person's first language. A common example of this phenomenon is seen in Canadian Francophone adults, who have difficulty distinguishing and pronouncing the English voiced (V^+) and voiceless (V^-) fricatives /θ/ and /ð/, often substituting them with /t/ and /d/, which are present in the phonemic repertoire of French.

The present study sought to examine some of the limits of a method by which adults can be trained to perceive non-native contrasts [1][5]. Subjects' abilities to identify and discriminate the synthetic voiced and voiceless English "th" sounds improved significantly after only 90 minutes of training, and generalized from the synthetic stimuli used in training to *natural* speech counterparts of /ð/ and /θ/ in CV syllables recorded by male and female talkers. Training effects transferred across different voices, but did not generalize to the same contrasts presented in different positions within the word, nor to the /ð/-/d/ contrast.

The present work explored the potential to increase generalization through a minor procedural modification: training with a VCV continuum, rather than a CV continuum, to provide acoustic cues associated with the formant transitions for *both* the preceding vowel and the following vowel. It was hypothesized that training would transfer to trained and non-trained word position and possibly endure when vowel context was altered.

2. METHOD

Our training paradigm employs the perceptual fading technique introduced by Terrace [6]. The goal is to train a perceptual contrast with a minimum amount of difficulty and few errors, by beginning training with an exaggerated exemplar of the feature being trained -- in this case, V^+ or V^- frication -- and providing immediate feedback. The distinction being trained is perceptually salient initially, but becomes progressively more subtle as training progresses.

2.1 Stimuli

2.1.1 Training Stimuli. Training used 8 VCV speech segments, synthesized at 20 kHz using an implementation of Klatt's [4] cascade/parallel speech synthesizer for IBM AT computers [3].

The 8 stimuli formed a VCV continuum with the consonant varying from the voiced interdental fricative /ð/ to the voiceless interdental fricative /θ/. The neutral vowel, /ʌ/, was used in both the initial and final positions in all training stimuli. The parameter values used to generate these sounds were based on those used in [1].

Vowel duration was fixed at 135 ms and 210 ms for initial and final positions, respectively; the duration of frication varied from 360 ms in stimuli 1 and 8, and to 90 ms in stimuli 4 and 5, respectively, decreasing by 90 ms for each consecutive stimulus.

2.1.2 Test Stimuli. Three sets of test stimuli were used in pretesting and posttesting. The *synthetic VCV syllables* were the same stimuli used in training. The *natural speech nonsense syllables* included 8 VCVs, 8 VCs, and 8 CVs. Within each subset, each of the English sounds /t/, /d/, /θ/, and /ð/ appeared once with the vowel /ʌ/, (as in training), and once with the vowel /i/ (where formant transitions were dissimilar). All tokens were spoken by 2 native speakers of Canadian English, 1 male and 1 female, and recorded, edited, and stored on disk using the CSRE software [3].

The 12 minimal-pair *word* stimuli contrasted /ð/ with /θ/; /ð/ with /d/; /θ/ with /t/; and /t/ with /d/ -- each, in -initial,-medial and word-final position. The word pairs used were: either-ether; riding-writhing; pity-pithy; wading-waiting; loathe-loath; fraught-froth; bade-bathe; bud-but; this'll-thistle; tinker-thinker; den-then; and doe-toe.

2.2 Subjects

Twenty-one unilingual Francophone college and university students (8 males and 13 females), participating in an English-language summer immersion program at the University of Western Ontario, were paid to serve as subjects. All placed in the lowest third of their class on an English language placement test and an oral language interview, and all passed a hearing screening at 20 dB HTL.

Pretest scores were used to construct one control and one experimental group, approximately matched in both pre-training perception skills, and gender.

2.3 Procedure

Pretesting was conducted one week, with training during the second week, and posttesting during the third week. Instructions were given in French, and only when it was clear that the task was understood did testing proceed.

2.3.1 Pretesting. Subjects were tested individually in a sound-attenuating booth, being seated at a small table facing a monitor on which were displayed response alternatives. Subject's listened to a signal over headphones, then identify the consonant portion as being either /d/, /t/, /θ/, or /ð/. On each trial, 4 words appeared on the monitor: THE, THING, DOG, TIME -- each containing one of the target sounds.

2.3.2 Training. Subjects were trained individually using the configuration described for pretesting. The task was to listen to a signal (one of the 8 synthetic sounds), then identify the consonant as being either /θ/ or /ð/ by selecting the word containing the target (THE, or THING). Immediately follow-

ing each response, the correct choice was illuminated to provide feedback. The initial stages of training were very easy, with stimuli selected from the opposite ends of the continuum. Progress from one stage to the next required 90% correct performance, on three consecutive blocks of trials. As training progressed, more medial stimuli from the continuum were used, so that the task became more difficult. During the final two blocks of training sounds, were presented sounds in a background of speech babble, to simulate many real life listening situations. Subjects were tested for one hour periods with breaks between blocks. Only two subjects failed to complete training within the alloted 4 hours.

2.3.3 Posttesting. Posttesting stimuli and procedures were identical to those used in pretesting.

3. RESULTS

3.1 Synthetic Tokens

Data were first reduced to the proportion of each identification response made by each listener for each stimulus under each test condition. There was a clear bias on the part of all subjects to choose *voiced* responses ($t = -2.353$, $df = 19$, $p = .014$) for pre-test, ($t = -2.958$, $df = 19$, $p = .004$) scores) for posttest. To allow performance to be measured independently of such biases, identification responses were converted to A' scores, using each subject's hit rate with a given stimulus type (e.g., "*TH*ING" responses when /θ/ stimuli were presented), in combination with that subject's overall error rate on all stimuli of the opposite type (e.g. "/θ/" responses to presentations of voiced stimuli) as the False Alarm rate. Pairing was effective in terms of equating the two groups at pretest ($t = -.843$, $df = 38$, $p = .404$).

A' pretest scores were subtracted from posttest scores to arrive at an A' difference score indicating the change from pretest to posttest. The A' difference provided the basis for further analysis. Posttest scores were significantly higher than the pretest scores for the trained group, ($t = 3.814$, $df = 9$, $p = .002$, one-tailed) and for the control group ($t = 2.396$, $df = 9$, $p = .020$, one-tailed), but the trained group improved more than the control group from pretest to posttest ($t = 1.935$, $df = 9$, $p = .042$, one-tailed).

3.2 Natural Nonsense Syllables

Nonsense syllables varied in their similarity to the synthetic training stimuli by syllable structure (VC, CV, or VCV), vowel environment (/ʌ/ vs. /i/), consonant (/ð, θ, t, d/), and talker (male vs. female). The pretest scores of the 2 groups did not differ ($t = .417$, $df = 18$, $p = .681$).

Analyses on subsets of the nonsense syllables demonstrated substantially different degrees of generalization for different types of stimuli. Training transferred directly to the natural /ʌðʌ/ and /ʌθʌ/ tokens, with an overall improvement in performance for both talkers, under both noisy and quiet conditions ($t = 2.68$, $df = 9$, $p = .013$ for the trained group and $t = 2.06$, $df = 9$, $p = .034$ for the control group). A greater improvement occurred for the trained group ($F(1,9) = 4.83$, $p = .055$).

Less transfer occurred when syllable structure changed (ie., with /θʌ, ðʌ, ʌθ, and ʌðʃ) and in other vowel contexts (ie., with /iðɪ/ and /iθi/). Scores for the /ʌtʌ/ and /ʌdʌ/ tokens failed to show an effect of training ($F(1,9) = 2.11$, $p = .1805$), reflecting the good pretest performance which produced a ceiling effect, reducing the *possibility* for a training effect.

3.3 Natural Word Pairs

The 48 natural words varied in terms

of syllable structure (VC, CV, or VCV), vowel environment, consonant (/ð, θ, t, d/), and talker (male vs. female). For example, for the "either, ether" word pair (same syllable position, but in a different vowel context from that used in training), training generalized and trained subjects showed somewhat better performance than did control groups ($F(1,9) = 2.26$, $p = .0928$). However, with the word pair "loath, loathe" (different word-position and vowel environment), training did not generalize, and trained subjects did not differ from control subjects ($F(1,9) = .33$, $p = .570$).

4. DISCUSSION

The present findings display an orderly pattern of results. A' difference scores improved most for the identification tasks involving the syllables /ʌθʌ/ and /ʌðʌ/, which are identical both in structure and in phonemic content to the synthetic training stimuli, next for nonsense syllables and words in which the syllable structure and consonant were held constant while the vowel environment differed from the training stimuli, and least for conditions involving altered syllable structures (CV and VC) and non-trained homorganic phonemes (/t/ and /d/). Consistent with previous research [2], speaker sex did not affect listeners' ability to perceive the non-native phoneme contrasts on which they were being trained.

5. REFERENCES

[1] JAMIESON, D.G., & MOROSAN, D. (1986), "Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones", *Perception & Psychophysics, 40*, 205-215.

[2] JAMIESON, D.G., & MOROSAN, D. (1989), "Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques", *Canadian Journal of Psychology, 43*, 88-96.

[3] JAMIESON, D.G., NEAREY, T., & RAMJI, K. (1989), "CSRE: The Canadian Speech Reasearch Environment", *Canadian Acoustics, 17*, 23-35.

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

[5] MOROSAN, D. & JAMIESON, D.G. (1989), "Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word-position or task", *Journal of Speech and Hearing Research, 32*, 501-511.

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

ACKNOWLEDGEMENTS

We are grateful to M.F. Cheesman, J. Booth, K. Ramji and W. Allsop for advice and assistance. The project was supported, in part, by grants from the NSERC, URIF, and from Unitron Industries Ltd. Address correspondence to Dr. D.G. Jamieson, Hearing Health Care Research Unit, University of Western Ontario, London, ON, CANADA, N6G 1H1.