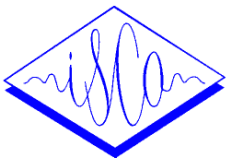

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

ACQUISITION OF THE FRENCH VOT CONTRASTS BY ADULT SPEAKERS OF MANDARIN CHINESE

Bernard L. Rochet* and Fangxin Chen**

*Department of Romance Languages
**Department of Linguistics
The University of Alberta, Edmonton, Alberta, Canada, T6G 2E6

ABSTRACT

The present study was designed to investigate whether perceptual training in the form of structured identification tasks with synthetic stimuli could have an effect on the perception and the production of voicing contrasts in the stop consonants of Standard French by native speakers of Mandarin Chinese. Twelve adult subjects took part in the experiment, which consisted of a pretest, 3 hours of training in six half-hour sessions, and a posttest. The pretest and the posttest consisted of identifying and imitating natural stimuli, and of identifying synthetic CV stimuli from several continua in which the voice onset time (VOT) duration of the initial stops varied from -60 ms to +130 ms in 10 ms steps. Training materials were 7 sets of synthetic stimuli consisting of a labial stop followed by the vowel /u/, sequenced according to a technique based on perceptual fading. After three hours of training, a) the subjects' identification functions were closer to those of native speakers of French for the trained continuum; b) there had been transfer of the training effect to the other places of consonantal articulation, and to the vowels [a] and [i] for the labial stops; and c) improved performance was also observed at the production level (imitation task). These findings suggest that adults may learn to perceive and produce non-native speech contrasts with limited but structured perceptual training, and that training effect may transfer to phonetic contexts other than those included in the training set. This last result is interpreted as reflecting the fact that although voicing distinctions are actualized at different locations on the VOT continuum in different languages, variations in VOT duration in terms of phonetic context exhibit some well-defined universal tendencies.

10.21437/ICSLP.1992-87

I. INTRODUCTION

The use of auditory training as a prerequisite to the development of acceptable pronunciation of a second-language (L2) has been advocated by many researchers and practitioners. It is generally agreed that auditory training leads to improved phonetic perception, and some claim that perception of L2 contrasts is not only a necessary, but also a sufficient condition for their acceptable production. Others, however, contend that auditory training has little to contribute to the acquisition of correct pronunciation habits. The evidence gathered so far through a wide array of experiments is ambiguous [1]. While some have shown that improved perceptual performance carried over in some measure to production, the results are often unconvincing, in that training appears to be effective for some pronunciation problems but not others [2], or only some of the observed trends reach statistical significance [3], or do so only for some scorers [4]. In addition, a positive relationship between auditory training and articulation is usually found in studies of younger learners and rarely found for adult learners [5], a finding which is in keeping with the common observation that "late" L2 learners rarely achieve a near-native pronunciation in the target language [6].

Recent studies combining technological advances and new methodologies have shown that short periods of training with carefully structured training sets can result in improved perceptual performance by adults learning non-native contrasts [7, 8, 9]. The question of carryover or transfer of training is raised in most of these studies. It is important to establish whether, and under what conditions, auditory training on one specific pair of speech sounds (or syllables) contrasting in a single feature may facilitate the discrimination and identification of other pairs participating in the same contrast. Some results suggest that training from a small set of stimuli may carry over to untrained stimuli [10], and to the perception of natural stimuli [8]. Little has been done, however, to establish whether improvement in perceptual performance carried over to production.

The purpose of this paper is contribute to the present understanding of the value of auditory training in the teaching/learning of L2 pronunciation. The experiment reported below attempted to determine whether perceptual training in the form of structured identification tasks with synthetic stimuli could have an effect on the perception and the production of voicing contrasts in the stop consonants of Standard French by native speakers of Mandarin Chinese. In particular, it aimed at establishing whether training with synthetic stimuli consisting of one pair of consonants (voiced and voiceless dental stops) in front of a single vowel (/u/) could facilitate the perception of the voicing contrast for consonants with a different place of articulation, and followed by different vowels.

II. EXPERIMENT

Twelve native speakers of Mandarin Chinese (5 males and 7 females), ranging in age between 28 and 37, and reporting no hearing deficiencies, took part in the experiment. All had been living in Canada for 2 to 4 years. All subjects were paid for their participation.

The experiment consisted of a pretest, followed by 3 hours of training in six half-hour sessions, and a postest. The pretest and the posttest consisted of an imitation task and a perceptual task. In the imitation task, subjects were asked to repeat words recorded by a native speaker of Standard French and containing voiced and voiceless stops (labials, dentals and velars) before the vowels /u/ for the 3 series and /a, i/ for the labials, in initial position. These natural speech tokens were recorded by a native speaker of Standard French and digitized at a sampling rate of 10kHz. The digitized stimuli were presented 10 times each in randomized order to each subject for imitation. They were played free field in a sound attenuated room at a comfortable listening level. Subjects' imitations of these signals were recorded on a TEAC V-437C cassette recorder by means of a Sennheiser MD421N microphone. At the same time, subjects had to identify the initial consonant of each stimulus by clicking with a mouse on the appropriate box of a computer display.

In the perceptual task, subjects were asked to identify as voiced or voiceless the initial consonant C of CV synthetic stimuli varying in VOT from -60 to +130 ms in 10 ms steps, and in which C was a labial followed by the vowels /u, a , i/, or a dental or velar followed by the vowel /u/. The vowel duration was kept constant at 250 ms. These stimuli were produced in cascade with the Klatt cascade/parallel synthesizer [11] at a sampling rate of 10 kHz, using Canadian Speech Research Environment software [12]. The stimuli were presented 10 times in randomized order, and delivered to subjects through a loud speaker in the sound attenuating booth which was used for the imitation task.

The training materials were 7 sets of synthetic stimuli consisting of a labial stop followed by the vowel /u/, starting with maximally differentiated exemplars, and gradually narrowing the gap between the two categories. This 7-step sequence was designed to implement a modified fading technique [8]. In the first five sets, each category was

represented by 3 tokens which differed from each other by 10 ms of VOT duration. Thus, the first set consisted of six syllables with VOT values of -60, -50, -40 ms for the voiced category, and +70, +80, and +90 ms for the voiceless category. In the second set, the gap between the two categories was narrowed, and the available tokens were characterized by VOT durations of -50, -40, -30, and 60, 70 and 80 ms. More medial stimuli on the continuum were introduced in the same way with each set, until the VOT values in set #5 were -20, -10, 0, and 30, 40, and 50 ms. In sets #6 and #7, the number of tokens for the voiced category was increased to include all the stimuli from -60 to 0 ms (in steps of 10 ms), while the voiceless category was represented by tokens with 30, 40, and 50 ms of VOT. This aspect of the fading technique, which consists of gradually reducing the magnitude of the perceptual contrast, was supplemented by a modification of the mode of presentation of the stimuli in the last two sets. While the stimuli were delivered by means of good quality earphones in sets #1 to #7, they were played over a loud speaker in sets #8 and #9. In addition, in set #9, a background of cafeteria noise was introduced.

For each training set, subjects listened to voiced and voiceless examples until they were satisfied that they could tell the difference between the two categories. This was accomplished by clicking with a mouse on two boxes--labelled "p" and "b" respectively--on a computer screen. Clicking on a given box triggered the playing, in random fashion, of one of the several tokens corresponding to the category represented in that box, and available for that particular set. Subjects were free to click on each box in any order and as many times as they wished. When they felt that they had no difficulty with the sound/label associations presented to them by the computer during that training phase, they proceeded to a test in which the same sounds were presented to them for identification in randomized fashion for a total of 120 per test. As each sound was heard, subjects clicked on one of two boxes on the computer screen, labelled "p" and "b" respectively. For each answer, subjects were given immediate feedback: A message appeared on the screen and told them whether their answer was correct or incorrect, and encouraged them to replay, as many times as they wished, the test sound they had just heard and attempted to identify, until they were satisfied that what they heard corresponded to the correct answer.

Subjects worked at their own pace, for 6 thirty-minute sessions. They were allowed to proceed to the next unit when they had completed a set with a success rate of 95%. Every day, they started the session by repeating the last set they completed successfully on the previous day. All subjects completed the 9 sets within the 6 thirty-minute training sessions.

III. RESULTS

The purpose of the experiment was to answer the following questions:

1. Does training result in modification of the perception of the stop continuum used in the training set (labial + /u/)?
2. If any training effect is observed for the labial + /u/ series, does it transfer to the other consonantal series for the same vowel?
3. Does training transfer to the other vowels (/a, i/) in the labial series?
4. Does training transfer to the perception of natural stimuli?
5. If training has an effect at the perceptual level, is that effect accompanied by a similar effect on production?

Perception was assessed in two ways: a) by obtaining identification functions and crossover boundaries for each of the synthetic continua; and b) by checking the accuracy with which subjects identified the natural tokens presented to them in the imitation task.

Production was also assessed in two ways: a) each item produced by each subject was identified as voiced or voiceless by 3 native speakers of French, and the number of correct identifications vs. mistakes was tallied; and b) each item was measured for VOT values using the Alligator waveform editing program [13].

Results of the Pretest

For the perceptual task with natural stimuli, 41% of all the stimuli--i.e., all the tokens containing a voiced or a voiceless initial stop, in the vocalic environments described above--were misidentified. Of the faulty responses, 87% were for voiceless stimuli, which were perceived as voiced, and 13% were for voiced stimuli, which were perceived as voiceless. Focusing on the patterns of misidentification within each voicing category, 11% of the voiced stops were misperceived (as voiceless), and 72% of the voiceless stops were misperceived (as voiced).

On the other hand, in the imitation task, 38% of the stimuli were repeated incorrectly, i.e., with the wrong voicing characteristics. Of these faulty repetitions, 27% were for voiceless stimuli which were perceived as voiced by the francophone judges, and 83% were for voiced stimuli, which were perceived as voiceless by the francophone judges. Stated differently, 52% of the voiced stops were repeated as voiceless, and 29% of the voiceless stops were repeated as voiced. The lack of parallelism between the perception and the production scores can be understood if we consider the identification functions and crossover boundary values obtained via the perceptual task with synthetic stimuli. The mean pretest VOT value for the location of the boundary between voiced and voiceless stops was 39 ms. Because French voiceless stops are produced with VOT values which range between 15 and 60 ms depending on consonant place and vowel context [14], it is understandable that a significant number of French stops--which are articulated with less than 39 ms of VOT--can be perceived as voiced. On the other hand, the low percentage of French voiced stops perceived as voiceless is understandable because they are always articulated with less than 39 ms of VOT.[1] In order to understand the pattern of mismatching at the production level, it is necessary to keep in mind the results of the identification task performed by the subjects during the imitation task, their L1 articulatory habits, and the perceptual expectations of the French listeners. Chinese voiceless aspirated stops are produced with long lag VOT (with mean values between 95 and 115 ms), and their unaspirated counterparts are produced with short lag VOT (with mean values between 10 and 35 ms).[2] The mean boundary location between voiced and voiceless stops for the French judges was 18 ms. Thus, it is clear that stops identified as voiceless (or voiceless aspirated) by Chinese speakers were likely to be repeated in such a way as to be definitely perceived as voiceless by French listeners. On the other hand, stops perceived as voiced by Chinese speakers could be repeated by them in such a way as to fall on either side of the French boundary, and therefore be perceived as voiced or voiceless by francophones. In this way, a certain percentage of French voiceless stops mistakenly identified as voiced could still be produced with enough lag VOT to be perceived as voiceless by francophones, i.e., as if they had been identified and repeated correctly. This explains why the large number of misidentifications of voiceless stops does not translate into an equally large percentage of mispronunciations of the same stops. At the same time, the possible VOT range for Chinese unaspirated stops, and its relationship to the French voiced/voiceless boundary explains why the percentage of mispronounced voiced stops is higher than the percentage of misperceived voiced stops.

Results of the Posttest

Effect of training on perception. The results of the perceptual test (with synthesized stimuli) are summarized in Fig. 1, where the mean posttest values are compared with the mean pretest values and with the mean French values for the syllables considered. These results reveal that training led to a modification of the perception of the /bu-pu/ continuum: Whereas the mean pretest VOT boundary between the /bu/ and /pu/ syllables was located at approximately 40 ms, the posttest boundary is located at 28 ms, i.e., closer to the French boundary which is at 20 ms. These differences are significant at the 0.01 level (t=6,813, 252 d.f.).

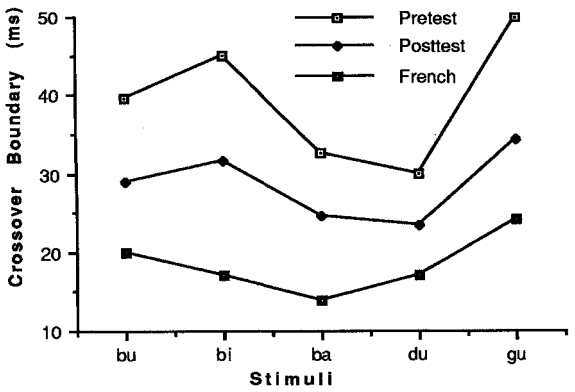

Fig. 1: Crossover Boundaries

The results also indicate that the effect of training (which was restricted to the syllables /bu/ and /pu/) was carried over to the other syllables starting with a bilabial, but followed by a vowel other than /u/: from pretest values of 45 and 33 ms in front of the vowels /i/ and /a/, respectively, the posttest boundaries are located at 32 and 25 ms, i.e., closer to the French target values of 18 and 14 ms. Similarly, the effect of training has transferred to the syllables starting with dental and velar stops, from pretest values of 30 and 50 ms to posttest values of 23 and 34 ms, respectively, i.e., closer to the French values of 17 and 23 ms.

Effect of training also transferred to the perception of voiceless natural stimuli. A smaller percentage (40% vs. 72%) of voiceless tokens were perceived as voiced during the posttest (t=15,492, 718 d.f., p<0.01). On the other hand, no significant change took place between the pretest and the posttest in the perception of voiced natural stimuli, which is not surprising because French voiced stops were rarely misidentified during the pretest.

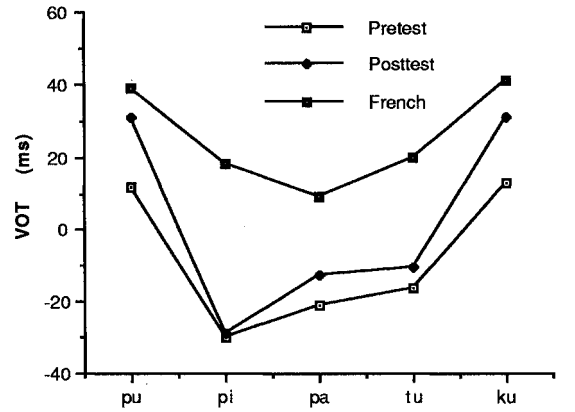

Fig. 2: Imitation of French Voiceless Stops (VOT values)

Effect of training on production. As shown in Fig. 2, the mean VOT durations of target voiceless stops increased between the pretest and the posttest, and became closer to the mean French values (pretest mean = -9, posttest mean = +0.5, t = -4.413, 718 d.f., p<0.01). This probably reflects the fact that a smaller number of voiceless tokens was identified as voiced in the posttest (see above).

Similarly, Fig. 3 shows an increase in the mean amount of prevoicing between the pretest and the posttest (from -36 to -57 ms, t=7.095, 719 d.f., p<0.01). It was noted above that the subjects had shown no significant improvement in their identification of voiced natural stimuli. It appears, therefore, that the increase in mean lead VOT duration was due to an increase in the extent of prevoicing in individual tokens identified as voiced.

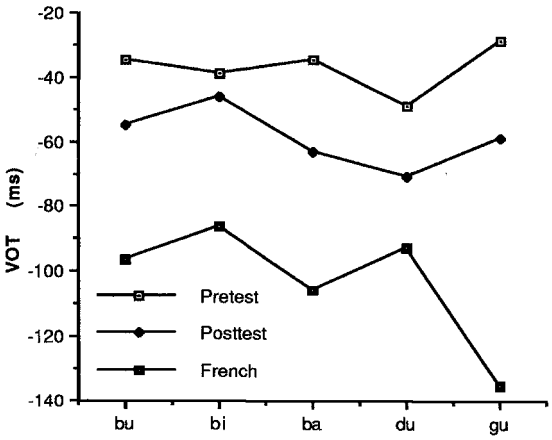

Fig. 3: Imitation of French Voiced Stops (VOT values)

Finally, assessment of the subjects' production revealed that a smaller number of tokens were mispronounced in the posttest, both for the voiced and the voiceless categories, as assessed by the French judges. While the improvement in the production of voiceless stops (from 30% to 19% of incorrect pronunciations) is significant at the 0.01 level (t=7.938, 719 d.f.), the change noted for voiced stops fails to reach significance (from 56% to 53%, t=1.598, 719 d.f., p=0.0552).

IV. DISCUSSION

The results of the present study show that transfer of training can take place across consonant place of articulation, and vocalic context. It would be premature, however, to conclude that such transfer is always possible, and that it is enough to train for one category/environment in order to obtain improvement for all the categories/environments which share the trained feature. For such an assumption to be valid, it would be necessary to have established that the same phonological category, or the same phonological contrast is actualized (in articulatory and acoustic terms) and perceived in the same way, in different phonological units and in different phonetic contexts. In the case of voicing contrasts, this means that such contrasts should be actualized and cued by the same physical property or properties for all consonantal types and in all phonetic contexts. That such is not the case has been well documented by Kohler [16].

The transfer of training to other phonological categories and environments (in perception), and to production performance, which was observed in the present study, may reflect the restricted scope of the investigation (the only environment considered is that of word-initial stops) and the possibility that voicing contrasts are actualized in parallel fashion (albeit with different VOT values) in French and in Mandarin Chinese stop consonants. There is evidence that VOT durations in word-initial stops vary in a strikingly parallel fashion in Standard French and in Mandarin Chinese [14, 15] with the consonantal place of articulation, and with the nature of the following vowel. In both languages, VOT durations are longer for velars than for dentals or labials, and in the presence of a following high vowel. The fact that, in Mandarin Chinese, the contextual variation observed in production is paralleled in perception [15] suggests that speakers of Mandarin Chinese are likely to include information about the context when assigning stop consonants to the "voiced" or "voiceless" category on the basis of VOT.

Contextual effects of an intrasegmental nature (consonantal place of articulation), and a transsegmental nature (height of the following vowel) on VOT durations have been observed in several languages [14, 17], and it has been suggested that the observed contextually-determined variation is attributable to aeromechanical factors, and can be considered universal [17]. This does not mean, however, that

all languages can be expected to exhibit parallel VOT variation. The universal tendencies mentioned above may be obscured or counteracted by language specific characteristics, with the result that two given languages may exhibit different patterns of VOT variation. In addition, the "voicing" contrast may not be actualized in all languages in the same way, and if it appears that VOT can be viewed as the primary cue for this contrast [18] it is equally clear that it is rarely the only cue [16, 19].

V. CONCLUSION

The results of the present study confirm previous studies' findings that brief treatment with structured sets of synthetic stimuli can lead to improvement in perception performance [7, 8, 9], with carryover to perception of natural stimuli [8]. They also suggest that improvement in perception performance can in turn translate into improvement in production performance, and that transfer of training from one specific context to other environments must be viewed as possible when contextual dependencies for the feature(s) to be acquired/modified are parallel in L1 and L2, as appears to be the case with VOT, a feature that behaves in remarkably uniform fashion across languages. But care must be taken not to use VOT as a cover term which would merely replace other cover terms such as voiced/voiceless. It is also essential to establish that VOT is the main acoustic cue or relevant articulatory feature, and to verify whether the voiced/voiceless distinction is cued by VOT in all environments. Training on VOT cannot be expected to transfer to environments where the voicing contrast is not cued by VOT but by other characteristics, such as vowel or consonant closure duration.

This underscores the need to conduct thorough contrastive analyses of the L1 and L2 materials, in all the possible contexts, and from a perceptual point of view as well as a production focus [16]. Only by producing such analyses, and providing detailed specifications of allophonic rules, not in terms of general/generic categories, but of the perceptual and articulatory behaviour of the speaker/listener, can we hope to achieve some degree of success in our attempts to predict what mistakes L2 learners are likely to make, and what corrective steps are most likely to be efficient.

Notes

[1] We would in fact expect that no French voiced stop would be perceived as voiceless by Chinese speakers. The fact that 11% of them were may be attributable to random mistakes and to the Chinese speakers' lack of familiarity with long lead VOT stops.

[2] These are mean values, and the range of actual VOT durations is larger. These durations vary with consonant place and with the nature of the following vowel, with the larger values occurring for velars and before high vowels [15].

References

[1] B. L. Rochet. "Training non-native speech contrasts on the Macintosh." In M.-L. Craven, R. Sinyor, and D. Paramskas, eds., *CALL: Papers and reports*. La Jolla, CA: Athelstan , pp. 119-126, 1990.

[2] P. Pimsleur. "Discrimination Training in the Teaching of French Pronunciation." *Modern Language Journal*, vol. 47, no.5, pp. 199-203, 1963.

[3] T. H. Mueller, and H. Niedzielski. "The Influence of Discrimination Training on Pronunciation." *Modern Language Journal*, vol. 52, no.7 , pp. 410-416, 1968.

[4] W. A. Henning. "Discrimination Training and Self-Evaluation in the Teaching of Pronunciation." *IRAL*, vol. 4, no. 1, pp. 7-17, 1966.

[5] P. S. Weiner, "Auditory Discrimination and Articulation." *Journal of Speech and Hearing Disorders*, vol. 32, no.1, pp. 19-28, 1967.

[6] J. E. Flege. "The Production and Perception of Foreign Language Speech Sounds." In H. Winitz, ed., *Human Communication and its Disorders, a Review 1988*. Norwood, New Jersey: Alex Publishing.

[7] D. B. Pisoni, R. N. Aslin, A. J. Perey, and B. L. Hennessy. "Some Effects of Laboratory Training on Identification and Discrimination of Voicing Contrasts in Stop Consonants." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, pp. 297-314, 1982.

[8] D. G. Jamieson, and D. E. Morosan. "Training Non-native Speech Contrasts in Adults: Acquisition of the English /θ/-/ð/ by Francophones." *Perception and Psychophysics*, vol. 40, no. 4, pp. 205-215, 1986.

[9] D. G. Jamieson, and D. E. Morosan. "Training New, Nonnative Speech Contrasts: A Comparison of the Prototype and Perceptual Fading Techniques." *Canadian Journal of Psychology*, vol. 43, no. 1, pp. 88-96, 1989.

[10] C. McClaskey, D. B. Pisoni, and T. D. Carrell. "Transfer of Training of a New Linguistic Contrast in Voicing." *Perception and Psychophysics*, vol. 34, pp. 323-330, 1983.

[11] D. H. Klatt. "Software for a Cascade/Parallel Formant Synthesizer." *Journal of the Acoustical Society of America*, vol. 67, no 3, pp. 971-995, 1980.

[12] D. G. Jamieson, T. M. Nearey, K. V. Ramji, and T. A. Baxter. *Canadian Speech Research Environment* (version 3.0). Dept. of Communicative Disorders, University of Western Ontario, and Dept. of Linguistics, University of Alberta, 1990.

[13] A. J. Roszypal. "Wavelet Speech Synthesizer." *Canadian Acoustics*, pp. 62-67, 1987.

[14] B. L. Rochet, T. M. Nearey, and M. J. Munro. "Effects of Voicing, Place and Vowel Context on VOT for French and English Stops." *Journal of the Acoustical Society of America*, vol. 81 (Supplement), p. S65, 1987. (Abstract).

[15] B. L. Rochet, and Y. Fei. "Effect of Consonant and Vowel Context on Mandarin Chinese VOT: Production and Perception." *Canadian Acoustics*, vol. 19, no. 4, pp. 105-106, 1991.

[16] K. J. Kohler. "Contrastive Phonology and the Acquisition of Phonetic Skills." *Phonetica*, vol. 38, pp. 213-226, 1981.

[17] J. J. Ohala. "Articulatory Constraints on the Cognitive Representation of Speech." In T. Myers, J. Laver, and J. Anderson, eds., *The Cognitive Representation of Speech*. Amsterdam: North Holland, pp. 111-127, 1981.

[18] L. Lisker. "In Qualified Defense of VOT." *Language and Speech*, vol. 21, no. 4, pp. 375-383, 1978.

[19] L. Santerre, and C. Y. Suen. "Why look for a Single Feature to Distinguish Stop Cognates?" *Journal of Phonetics*, vol. 9, pp. 163-174, 1981.