Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training

Yue Wang[a)]
Department of Linguistics, University at Buffalo, State University of New York, Buffalo, New York 14260

Allard Jongman and Joan A. Sereno
Linguistics Department, University of Kansas, Lawrence, Kansas 66044

(Received 1 March 2002; accepted for publication 26 October 2002)

Training American listeners to perceive Mandarin tones has been shown to be effective, with trainees' identification improving by 21%. Improvement also generalized to new stimuli and new talkers, and was retained when tested six months after training [Y. Wang *et al.*, J. Acoust. Soc. Am. **106**, 3649–3658 (1999)]. The present study investigates whether the tone contrasts gained perceptually transferred to production. Before their perception pretest and after their post-test, the trainees were recorded producing a list of Mandarin words. Their productions were first judged by native Mandarin listeners in an identification task. Identification of trainees' post-test tone productions improved by 18% relative to their pretest productions, indicating significant tone production improvement after perceptual training. Acoustic analyses of the pre- and post-training productions further reveal the nature of the improvement, showing that post-training tone contours approximate native norms to a greater degree than pretraining tone contours. Furthermore, pitch height and pitch contour are not mastered in parallel, with the former being more resistant to improvement than the latter. These results are discussed in terms of the relationship between non-native tone perception and production as well as learning at the suprasegmental level. © *2003 Acoustical Society of America.* [DOI: 10.1121/1.1531176]

PACS numbers: 43.70.Kv [AL]

I. INTRODUCTION

Laboratory training of the discrimination of new phonetic contrasts is based on the assumption that the adult human perceptual system still has the capacity to change. Indeed, recent research has found that non-native segmental contrasts can be learned through auditory training. For example, there have been studies that trained American listeners with a three-way voice-onset time (VOT) distinction (e.g., Pisoni *et al.*, 1982; McClaskey, Pisoni, and Carrell, 1983), trained French listeners to identify the English /θ–ð/ contrast (e.g., Jamieson and Morosan, 1986, 1989), and trained Japanese listeners to identify English /r/ and /l/ (e.g., Logan, Lively, and Pisoni, 1991; Lively, Logan, and Pisoni, 1993; Lively *et al.*, 1994; Bradlow *et al.*, 1997). These studies have shown that the identification of non-native speech contrasts improved with training, and the improvement was extended to novel phonetic contexts and was retained long after training.

While these studies show the effect of training on segmental learning, Wang *et al.* (1999) extended the training procedure to the suprasegmental level by training American listeners to identify Mandarin Chinese tones. Mandarin phonemically distinguishes four tones, with tone 1 having high-level pitch, tone 2 high-rising pitch, tone 3 low-dipping pitch, and tone 4 high-falling pitch (Chao, 1948). Studies of the acoustic characteristics of Mandarin tones found that the differences in tones are manifested physically by different fundamental frequency ($F0$) values (Liu, 1924), with $F0$ height and $F0$ contour as the primary acoustic parameters characterizing Mandarin tones (Howie, 1976). For learners whose native language is nontonal, tone has presented great difficulty, since the functional association between these $F0$ characteristics and the segmental structure is unfamiliar to them (e.g., Kiriloff, 1969; Bluhme and Burr, 1971; Shen, 1989). Mandarin tone thus provides an ideal case for the study of suprasegmental training.

In Wang *et al.* (1999), eight American learners of Mandarin were trained in eight sessions during the course of 2 weeks to identify the four tones in natural words produced by native Mandarin talkers. Results show that, consistent with the previous segmental training studies, the perception of Mandarin tone improved significantly after training (21% improvement). Moreover, this improvement generalized to new stimuli (18% improvement) and new voices (25% improvement), and was retained when probed 6 months after training (21% improvement). These results are consistent with the previous findings at the segmental level, suggesting that training produces highly generalized perceptual learning that yields long-term modifications of the learners' perceptual system.

One subsequent question is whether perceptual training can affect production, so that training efforts could result in positive transfer from one modality to the other. The transfer of learning from perception to production has been reported in studies training learners to perceive non-native segmental contrasts. Rochet (1995) examined the transfer effect of perceptual training on production of French voice onset time (VOT) categories by native speakers of Mandarin Chinese.

[a)]Electronic mail: yuewang@buffalo.edu

Using an imitation task, productions of voiced and voiceless stops (labials, dentals, and velars) in a variety of vowel contexts were elicited both before and after perceptual training. Perceptual training involved synthetic stimuli consisting of a labial (/b/ or /p/) stop followed by a single vowel (/u/) context. An assessment of production gains following perceptual training was carried out by measuring VOT values in pretest and post-test productions. Mean VOT durations for initial stops did show improvement towards more French-like VOT values. Production accuracy was also assessed by native speaker judgments, with voiceless stops exhibiting more misidentifications in pretest productions compared to post-test productions. While perceptual training did transfer to production of voiceless stops in initial position, the improvement in production was not significant for voiced stops, and did not generalize to the production of stops in intervocalic position. Although Rochet suggests that lack of generalization may be the result of phonetic cues being actualized differently for stop consonants in initial versus intervocalic position, the training procedures may also be responsible. Using different methodological manipulations, Bradlow *et al.* (1997, 1999) also investigated the effects of perceptual training on production, by examining the production of the English /r–l/ contrast by Japanese learners. Similar to Rochet, perception and production data were gathered both at pretest and post-test, and native speakers were used to assess production gains. In the Bradlow *et al.* studies, however, perceptual training involved identification of naturally produced English /r/ and /l/ minimal word pairs using a high variability perceptual training procedure (Logan *et al.*, 1991). The results are consistent with Rochet, showing that native speakers identified the post-test productions more accurately than the pretest productions. Moreover, these production improvements were generalized to novel stimuli, and were retained 3 months after the perceptual training. However, the studies found no correlation between degree of learning in perception and production. That is, it is not the case that improvement in perception and production proceeded in parallel within individual learners (Bradlow *et al.*, 1997). Overall, these studies have shown that perceptual training has a facilitatory effect on the production domain, but the nature of the relationship between perception and production is still not clear.

Although nearly all training studies have focused on perception, Leather (1990) reports an initial attempt to examine the effect of production training on perception. This study examined a group of Dutch speakers who were trained to produce four Mandarin words (with the same syllable “*yu*”) differing in tone. Leather found that these Dutch speakers were able to perceive the differences in tone without perceptual training. The author concluded that training in one modality tended to be sufficient to enable learners to perform in the other. However, since only one syllable was used in training as well as testing, the generalizability of the learning effect was not easily determined. We do not know if this type of training can produce long-term learning that can be extended across stimuli, voice, as well as speech modality.

In the present study, American trainees were recorded, both before and after perceptual training, producing a list of Mandarin words. Since the perceptual training of Mandarin tones has resulted in long-term perceptual improvement across stimuli and voice (Wang *et al.*, 1999), the goal of this study is to examine whether perceptual learning of this suprasegmental property can be transferred to the production domain. This study presents the results of the American learners’ productions before and after training, first with an assessment by native Mandarin listeners in an identification task, and then with an acoustic analysis of pitch track comparisons. This is the first attempt to quantify the training effect with acoustic analysis, to capture the nature of the production improvements following perceptual training.

II. NATIVE LISTENER EVALUATION

A. Method

1. Participants

The participants were the 16 native speakers of American English in the perceptual study, with eight trainees who participated in the 2-week perceptual training program, and eight controls who did not receive training (Wang *et al.*, 1999). The participants were randomly selected from the students at Cornell University who had taken one or two semesters of Mandarin Chinese. None of them had ever lived in a Mandarin-speaking environment, and most of them had no experience with a tone language prior to learning Mandarin (except for four with limited Cantonese). All were paid for their participation. (For details of the characteristics of the participants, see Wang *et al.*, 1999.)

Eighty-two adult native speakers of Mandarin Chinese (18–35 years old) with no reported speech or hearing impairments participated voluntarily as judges. They were all raised and educated in Beijing, and were familiar with the *pinyin* system and the tonal diacritics which were used in this study.

2. Stimuli

The stimuli consisted of 80 real monosyllabic Mandarin words presented in isolation, 20 with each of the four tones. Half of these stimuli were used in the perceptual training (“old” stimuli), while the other half did not appear in training (“new” stimuli), in order to test the generalization of the production gains. The same stimuli were used for both pretest and post-test. Thus, each subject provided four sets of production stimuli: 40 old stimuli from pretest, 40 new stimuli from pretest, 40 old stimuli from post-test, and 40 new stimuli from post-test.

3. Procedure

Before the tone perception pretest, the trainees and the controls were tape recorded in a sound-insulated booth in the Cornell Phonetics Laboratory, using a cardioid microphone (Electrovoice RE20) and cassette recorder (Carver TD-1700). They were asked to read the list of 80 stimuli (blocked for old and new stimuli in the perceptual training) at a normal speaking rate, and were encouraged to repeat or correct whenever they felt necessary. The stimuli were presented on a sheet of paper in *pinyin* using the tonal diacritics familiar to the speakers. The speakers were recorded reading the same stimuli after their perception post-test 2 weeks later.

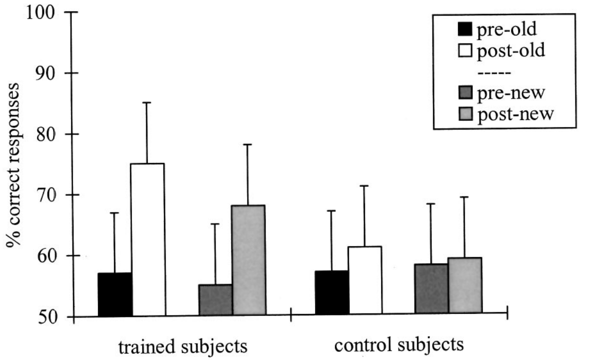

FIG. 1. Mean percent-correct identification of the four tone productions from the American trainees and the control subjects at pretest and post-test as judged by native Mandarin listeners. "Pre-old" and "post-old": pre- and post-test identification of the stimuli included in perceptual training; "pre-new" and "post-new": pre- and post-test identification of the stimuli not used in perceptual training.

The stimuli provided by the eight trainees and eight controls at both pretest and post-test were digitized at 11 kHz and low-pass filtered at 5 kHz using WAVES+/ESPS speech analysis software running on a SUN Sparc Station, after which they were edited and segmented.

These production stimuli were then transferred to a PC for play out in a perception experiment, using BLISS software (Mertus, 1989). The final output included four separate sets of data for each subject, i.e., 40 old and 40 new stimuli in perceptual training at both pretest and post-test, with a 3 s interstimulus interval. These stimuli were recorded to audio tape for evaluation by native Mandarin listeners in Beijing.

Prior to the perceptual judgment task, all potential Chinese judges were asked to identify a list of Mandarin words produced by a native Mandarin speaker. Only those who were able to identify all the tones correctly were included in the present study. Two listeners failed to reach this criterion and were excluded from participation.

The 16 American subjects' productions were evaluated by a total of 80 native Mandarin judges. Five different judges were asked to identify a single subject's productions at both pretest and post-test presented from a portable tape recorder. Answer sheets were provided containing the 40 stimuli in each of the four sets, written in *pinyin* with no tonal diacritics. Next to each stimulus, there were five categories, i.e., the four tonal diacritics, and a "none" category. Judges were to circle a tonal diacritic corresponding to the tone they heard, and to circle "none" if they decided what they heard did not correspond to any of the four tones. The order of presentation of the four stimulus sets for each of the 16 subjects was counterbalanced among judges. Thus, the identification task resulted in a total of 800 observations for each of the 16 speakers ($5\ \text{judges} \times 40\ \text{stimuli} \times 4\ \text{sets}$).

B. Results

1. Overall improvement and generalization

Figure 1 shows the overall results of the production judgments. It displays the percentage of correctly identified productions by the American trainees and the control subjects as evaluated by the Mandarin judges. It should be noted that in the identification task, the judges could also categorize the trainees' productions as being none of the four tones. However, the results reveal that this category constitutes only a small proportion (2.7%) of the pre- and post-test judgments, indicating that for most of the cases, the trainees' productions were judged as one of the four Mandarin tones.

As shown in Fig. 1, the trainees showed an improvement in their production evaluation scores from 57% in the pretest to 75% in the post-test for the old stimuli, and from 55% to 68% for the new stimuli. This indicates the trainees' substantial improvement in production after perceptual training; that is, their improvement not only occurred on the stimuli used in perceptual training (18% increase), but was also generalized to new stimuli that were not included in the perceptual training (13% increase).

In contrast, although the control subjects started at approximately the same level as the trainees in the pretest ("old stimuli": 57%, and "new stimuli": 58%), they exhibited little improvement in the post-test ("old stimuli": 61%, and "new stimuli": 59%). The lack of substantial improvement for the controls occurred both for the stimuli included in the perceptual training (4% change) and also for the new stimuli (1% change).

A three-way repeated measures ANOVA was calculated with test (pretest, post-test) as within-subject factor and group (trained, control), and stimulus (old, new) as between-subject factor. The results revealed a significant main effect of test $[F(1,28)=49.3,\ p<0.0001]$, and a significant test by group interaction $[F(1,28)=26.9,\ p<0.0001]$. The effects of group $([F(1,28)=0.7,\ p>0.412])$ and stimulus $([F(1,28)=0.19,\ p>0.663])$ did not reach significance. There was no interaction of test×stimulus $[F(1,28)=2.0,\ p>0.168]$, group×stimulus $[F(1,28)=0.10,\ p>0.778]$, or test×stimulus×group $[F(1,28)=0.14,\ p>0.712]$. These results suggest that the pretest and post-test performance was different for the trained and control subjects, and this difference occurred across the old and new stimuli in the perceptual training. More specifically, for the trained group, paired samples t tests showed a significant difference between pretest and post-test: $[t(15)=8.90,\ p<0.0001]$. In contrast, for the control group, no significant difference was observed for test $[t(15)=1.30,\ p>0.226]$.

In sum, the above results show a significant improvement in production as the result of perception training for the trainees. Native Mandarin listeners more often perceived the intended tone after training as compared to before training. Moreover, this improvement in production was observed both for stimuli used in training and was extended to novel stimuli not included in perceptual training. However, the perceptual ratings of the native Mandarin listeners suggested no such improvement for the controls, judging controls' post-test productions as accurately as their pretest productions.

2. Individual trainees

Further analyses of these data examined productions for individual participants and individual tones. These subsequent analyses concentrated on the trainees and focused on the stimuli that were used in the perceptual training.

TABLE I. American trainees' percent-correct tone production as judged by Mandarin listeners.

Trainee	Pretest	Post-test	Increase
1	44	47	+3
2	24	55	+31
3	59	74	+15
4	69	88	+19
5	74	92	+18
6	77	90	+13
7	47	67	+20
8	61	83	+22
Mean	**57**	**75**	**+18**

Individual trainees' percent-correct tone production as judged by Mandarin listeners at pre- and post-test is presented in Table I, which shows that each trainee's production accuracy improved after perceptual training. Across all eight trainees, percent-correct tone production improved on average 18% from pretest to post-test. It is also noted that there is a large degree of variability among the eight trainees in terms of initial accuracy levels (24% to 77%), as well as amount of improvement (ranging from 3% to 31%).

3. Individual tones

The trainees' productions for each individual tone are illustrated in Fig. 2. Trainees' performance for each tone improved from the pretest to the post-test. A two-way ANOVA (test×tone) showed a main effect for test $[F(1,30) = 12.25, p<0.002]$, indicating significant improvement from the pre- to the post-test. There was also a main effect for tone $[F(3,28) = 7.45, p<0.001]$. *Post hoc* analyses (Tukey HSD) showed that across pre- and post-test, tone 3 is significantly worse than tones 1, 2, and 4. The interaction of test and tone did not reach significance $[F(3,28) = 0.67, p > 0.577]$, showing the improvement from pretest to post-test was consistent across all four tones.

4. Tone confusion

Table II presents a confusion matrix, for both the pre- and post-test, showing the number of production errors the trainees made for each tone as judged by native listeners of Mandarin. Tone confusions were examined, summing over the errors obtained for each tone pair. For example, the number of errors for tone pair 1 and 2 is the sum of errors of both tone 1 produced as tone 2, and tone 2 produced as tone 1. It can be seen that, at pretest, the most confusing tone pair was tones 2 and 3, followed by tones 2 and 4, tones 1 and 4, tones 1 and 2, tones 3 and 4, and tones 1 and 3. A similar rank order was retained at post-test (Spearman $\rho = 0.81$, $p < 0.05$), except for tones 1 and 4 which became the second most confusing pair.

TABLE II. Confusion matrices for the American trainees' tone productions (as judged by native Mandarin listeners) at (a) pretest and (b) post-test (10 stimuli×8 trainees×5 judges=400 responses for each tone). Correct responses are shown in bold.

	Stimulus			
	Tone 1	Tone 2	Tone 3	Tone 4
Produced as				
(a) pretest				
Tone 1	**300**	23	3	52
Tone 2	49	**255**	231	59
Tone 3	9	31	**101**	21
Tone 4	35	73	39	**262**
None	7	18	26	6
(b) post-test				
Tone 1	**365**	26	11	46
Tone 2	9	**340**	161	36
Tone 3	1	23	**184**	8
Tone 4	21	4	32	**305**
None	4	7	12	5

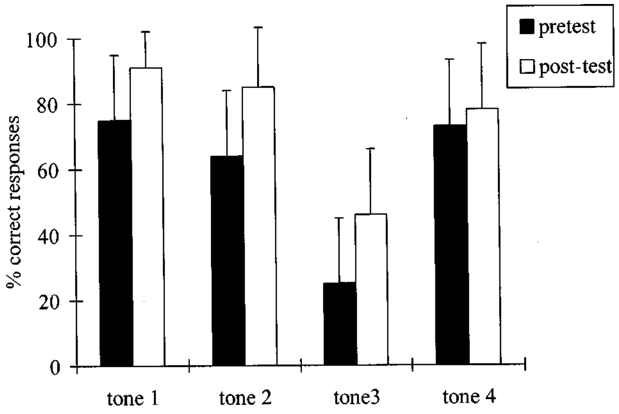

FIG. 2. American trainees' mean percent-correct productions for each tone at pretest and post-test as judged by native Mandarin listeners. The pitch contour shapes of the four tones are: tone 1, high-level pitch; tone 2, high-rising pitch; tone 3, low-dipping pitch; and tone 4, high-falling pitch.

III. ACOUSTIC ANALYSIS

A. Method

To examine the American learners' productions before and after training, an acoustic analysis of the pitch tracks of the pre- and post-training productions was undertaken. Native speaker norms were derived and trainees' productions were compared to these native speaker norms for each of the Mandarin tones.

1. Participants and stimuli

To provide native norms for the four tones, four native speakers of Mandarin Chinese (2 males, 2 females) were asked to produce the same list of words as those used for the trainees. The setting and procedure were identical to those for the trainees.

The 40 stimuli that were included in the perceptual training were used in the acoustic analysis for both the Chinese and American speakers, yielding a total of 160 stimuli (10 syllables×4 tones×4 speakers) for the native speaker productions, and a total of 640 stimuli (10 syllables ×4 tones×8 trainees×2 tests) for the non-native productions.

2. Analysis

A total of 800 pitch contours (160 native productions, 320 non-native pretest productions, 320 non-native post-test productions) was derived using the WAVES+/ESPS software, at a sampling rate of 1 ms.

In order to directly compare the productions across speakers and stimuli, the pitch contours were normalized in two ways: $F0$ normalization, to accommodate the pitch range differences among speakers (especially between males and females), and duration normalization, to adjust for differences in speaking rate and syllable context. Both the native productions and the non-native productions were normalized in this manner.

$F0$ was normalized per speaker across the four tones. That is, the $F0$ values obtained from each speaker were converted to their logarithms, using a formula commonly adopted for such purposes (e.g., Liao, 1983; Ladd *et al.*, 1985; Rose, 1987; Shi, 1986, 1994)

$$T=[(\lg X-\lg \mathrm{L})/\lg \mathrm{H}-\lg \mathrm{L}]\times 5,$$

where H is the highest and L is the lowest $F0$ for a given speaker, and X is any given point of a pitch contour. The output (T) is a value ranging from 1 to 5, corresponding to the 5-point pitch scale for Mandarin tone proposed by Chao (1948).

Duration was normalized per tone across speakers. For each tone, the longest pitch contour was first determined (containing a certain number of $F0$ values depending on the length of the production at the sampling rate of 1 ms). Taking this number of $F0$ values, all other pitch contours for that tone were then time normalized by deriving the same number of $F0$ values, thus interpolating between observed $F0$ values. For example, as the longest pitch contour for tone 1 across speakers and tokens was 571 ms (571 $F0$ points), all tone 1 productions were ‘‘stretched’’ to have 571 $F0$ points at 1-ms intervals. The pitch contours of the other three tones were stretched in the same fashion, resulting in 569 $F0$ points for tone 2, 616 $F0$ points for tone 3, and 520 $F0$ points for tone 4. Thus, the duration of each tone was equalized across all speakers and tokens.

Using the converted $F0$ values (T values), the native norm for each tone was generated by averaging the four native Mandarin speakers’ productions across all words. Likewise, for the non-native productions, two averaged productions were derived, one for the pretest and the other for the post-test by averaging across the eight trainees’ productions.

For each contour, pitch values at 0% (onset), 25%, 50%, 75%, and 100% (offset), as well as the highest (peak) and the lowest (valley) points of the contour were calculated to record overall pitch height and shape. Furthermore, a number of critical attributes were analyzed on the basis of their acoustic characteristics and perceptual salience. First, pitch range (range) was calculated, since this feature has been identified as a perceptual cue distinguishing tones 1 and 2, and tones 1 and 4 (e.g., Leather, 1983; Lin and Wang, 1985; Fox and Qi, 1990). Second, both falling pitch range from onset to valley (falling range) and rising pitch range from valley to offset (rising range) were calculated. The falling pitch range, also termed ‘‘$\Delta F0$,’’ has been found to be critical in the identification of tones 2 and 3 (Shen and Lin, 1991; Moore and Jongman, 1997); whereas the rising pitch was found to cue the identification of tone 2 (e.g., Blicher *et al.*, 1990). Finally, the relative temporal position (position) from the onset to the valley of the pitch contour and from the onset to the peak of the pitch contour was calculated, since the duration from the onset to turning point (corresponding to the valley) has been found to be shorter for tone 2 than for tone 3 (Dreher and Lee, 1966; Moore and Jongman, 1997).

To eliminate occasional spurious values obtained from the pitch-tracking algorithm at the beginning and the end of the stimuli, the first and last 5 points for each pitch contour were excluded from the analysis.

B. Results

Figures 3(a)–(d) illustrate, for each tone, the pitch contours of the pretest and post-test productions averaged across trainees and stimuli, as compared to the native norm. The pitch contours are normalized for $F0$ and duration. Pitch values are represented on a 5-point pitch scale as T values.

As shown in the figure, for each tone, the post-test production resembles the native norm more closely than the pretest production both in terms of pitch height and contour shape, clearly showing an improvement in production resulting from perceptual training. An analysis of the critical points for each tone provides details about the improvement.

Table III displays the averaged T values at 0%, 25%, 50%, 75%, and 100% of the pitch contour, as well as the pitch range (range), for tone 1 at pretest and post-test across trainees, as compared to the native norm.

The native norm shows that, as a high-level tone, the T value remains constant (range: 0) with a relatively high pitch (4.2). The trainees’ pretest production shows that the contour is relatively high and level, although the mean pitch values are a bit lower than the native norm and the pitch contour also decreases to a certain degree (range: 0.3). The pattern is mostly retained at post-test, except that the mean pitch values are even closer to the native norm.

A comparison of the pretest and post-test productions relative to the native norm was conducted using a two-way repeated measures ANOVA. The dependent variable is the ‘‘deviation score’’ (the absolute difference in T values between the native norm and either the pretest or the post-test production at a given point). The within-subject factor is test (pretest, post-test), and the between-subject factor is position (0%, 25%, 50%, 75%, 100%).

A significant effect of test was observed $[F(1,34)=15, p<0.0001]$, with an overall deviation score (i.e., an averaged deviation score across all points in a contour) of 0.45 for the pretest and 0.35 for the post-test. There was no reliable effect of position $[F(4,34)=0.08, p>0.987]$, nor was there a test by position interaction $[F(4,34)=0.3, p>0.904]$. These results show that across all 5 points in a contour, the difference between the post-test T value and the native norm (0.35) is significantly smaller than that between the pretest T value and the native norm (0.45), indicating greater degree of approximation to the native norm at post-test.

Overall, Table III, as well as Fig. 3(a) consistently reveal trainees’ improvement at post-test for tone 1. It should also be noted that both pretest and post-test pitch contours are ‘‘high’’ and ‘‘level,’’ suggesting that the trainees’ tone 1 production was relatively good at pretest, and further improved at post-test. This is consistent with the native speakers’

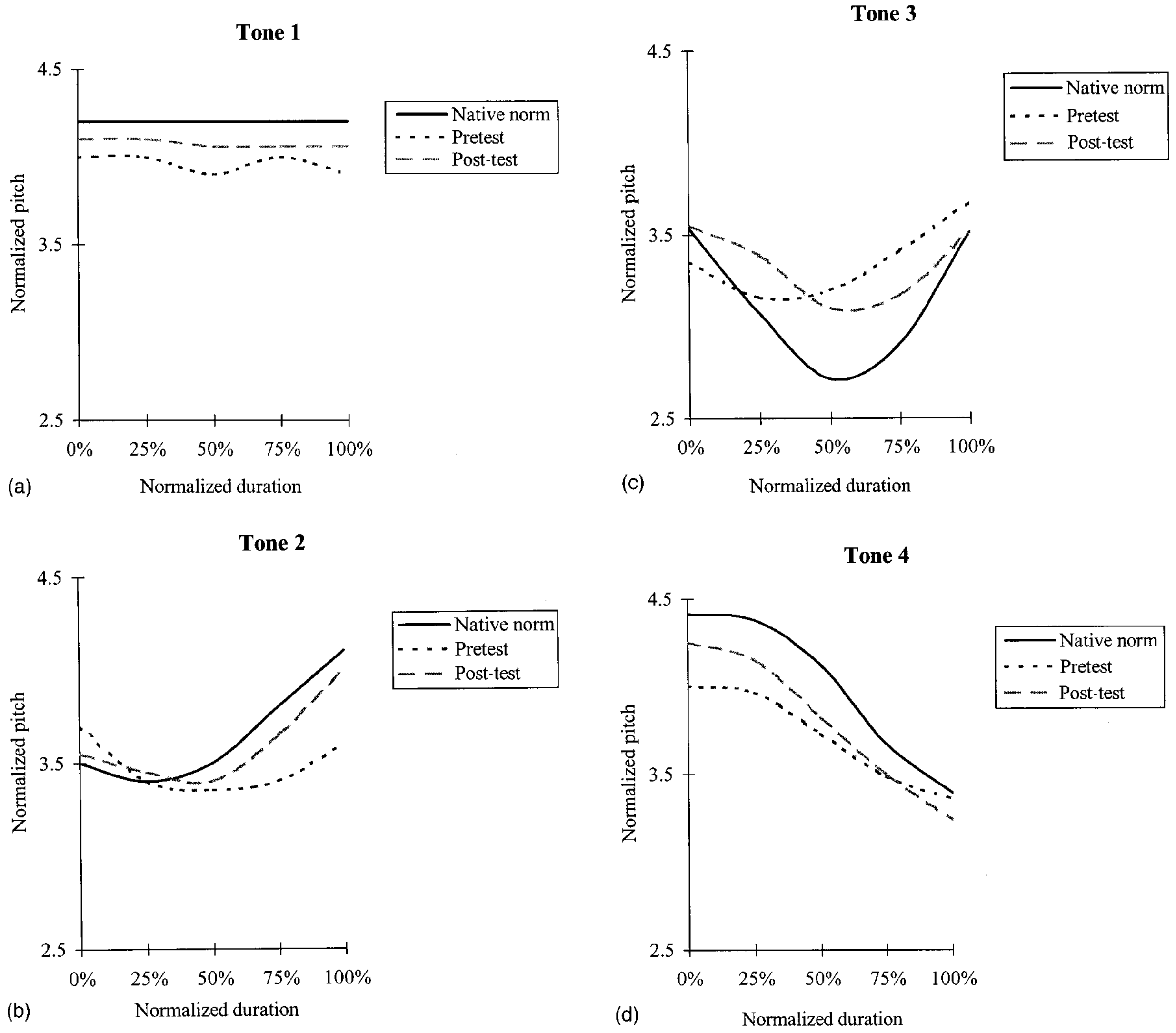

FIG. 3. (a)–(d) Pitch contours on the 5-point pitch scale (T values), comparing the native norm, and the pre- and post-test productions averaged across trainees, for tones 1–4, respectively. The pitch contour shapes of the four tones are: tone 1, high-level pitch; tone 2, high-rising pitch; tone 3, low-dipping pitch; and tone 4, high-falling pitch.

evaluation data, with 75% of the intended tone 1 productions judged by native speakers as correct tone 1 productions at pretest and this increased to 91% at post-test (cf. Fig. 2).

Table IV shows, for tone 2, the averaged T values at 0%, 25%, 50%, 75%, and 100% of the pitch contour, as well as the peak and valley T values and positions, and the pitch range values of the falling and rising portions of the contour, for the native norm and trainees' productions at pretest and post-test.

A two-way repeated measures ANOVA was conducted comparing the pretest and post-test productions relative to the native norm, with the deviation score as dependent variable, test as within-subject factor, and position as between-subject factor. A significant effect was observed for test [$F(1,34)=8.2$, $p<0.007$], but not for position [$F(4,34)=0.2$, $p>0.94$] or test by position interaction [$F(4,34)=1.3$, $p>0.284$]. The overall deviation values show that the difference between the pretest and native norm (deviation

TABLE III. Tone 1: Mean T values at 0%, 25%, 50%, 75%, and 100% (with standard deviations, s.d., in parentheses), as well as the peak (highest point), valley (lowest point), and range (pitch range), of the pitch contour of the native norm, and trainees' pre- and post-test productions.

	0% (s.d.)	25% (s.d.)	50% (s.d.)	75% (s.d.)	100% (s.d.)	Peak	Valley	Range
Native norm	4.2 (0.6)	4.2 (0.6)	4.2 (0.6)	4.2 (0.6)	4.2 (0.6)	4.2	4.2	0
Pretest	4.0 (0.6)	4.0 (0.6)	3.9 (0.6)	4.0 (0.6)	3.9 (0.6)	4.1	3.8	0.3
Post-test	4.1 (0.6)	4.1 (0.5)	4.0 (0.5)	4.0 (0.4)	4.0 (0.5)	4.1	3.9	0.2

TABLE IV. Tone 2: Mean T values at 0%, 25%, 50%, 75%, and 100% (with standard deviations, s.d., in parentheses) of the pitch contour, as well as mean T values at the lowest (valley) and highest (peak) points and their temporal position in % (expressed as a percentage of total time duration) (in brackets), and the falling pitch range (the difference between onset and valley) and the rising pitch range (the difference between valley and offset) of the pitch contour of the native norm, and trainees' pre- and post-test productions.

	0% (s.d.)	25% (s.d.)	50% (s.d.)	75% (s.d.)	100% (s.d.)	Valley [position]	Peak [position]	Falling range	Rising range
Native norm	3.5 (0.6)	3.4 (0.6)	3.5 (0.6)	3.8 (0.6)	4.1 (0.6)	3.4 [25%]	4.1 [100%]	0.1	0.7
Pretest	3.7 (0.5)	3.4 (0.5)	3.4 (0.6)	3.4 (0.6)	3.6 (0.7)	3.2 [65%]	3.7 [0%]	0.5	0.4
Post test	3.6 (0.5)	3.5 (0.4)	3.4 (0.4)	3.6 (0.4)	4.0 (0.4)	3.3 [47%]	4.0 [100%]	0.3	0.7

score: 0.50) is greater than that between the post-test and native norm (deviation score: 0.34), indicating significantly greater approximation to the native norm in the post-test than the pretest productions.

Detailed analysis of the critical points of the tone 2 native norm shows that, as a rising tone, the initial falling portion of the contour is relatively short (25% from onset) with minimal change in pitch range (0.1), while the rising portion is relatively long, reaching its peak (4.1) at the offset of the contour. Comparing the pre- and post-test test productions with the native norm, the post-test reveals greater approximation to the norm in a number of ways. First, the initial falling portion is shorter (47% versus 65%) and less steep (range: 0.3 versus 0.5) for the post-test than for the pretest. Second, similar to the native norm, in the post-test the major rising contour is long and reaches its peak at the offset (4.0). For the pretest, however, the peak (3.7) occurs at the onset, while the offset (3.6) does not rise as high as the native and post-test productions. These differences are clearly shown in Fig. 3(b).

Taken together, these results show for tone 2 a significantly higher degree of resemblance of the post-test production to the native norm both in terms of pitch contour and height, as compared to that of the pretest.

The learners' improvement is also reflected in the native speakers' evaluation described previously, as learners' correct tone 2 productions as judged by Mandarin speakers improved from 64% in the pretest to 85% in the post-test.

The average pitch values (T) for tone 3 at 0%, 25%, 50%, 75%, and 100% of the pitch contour, as well as the peak and valley T values and positions, and the pitch range values of the falling and rising portions of the contour, for the native norm and the trainees' productions are shown in Table V.

A two-way repeated measures ANOVA (test×position) with deviation score as dependent variable was conducted. An analysis of the difference between the pretest and the post-test as compared to the native norm at each of these 5 points shows that the post-test production (deviation score: 0.46) is significantly more similar to the native norm than is the pretest production (deviation score: 0.60) [$F(1,34)=7.6$, $p<0.009$]. Neither position [$F(4,34)=0.3$, $p>0.85$] nor test by position interaction [$F(4,34)=0.6$, $p>0.654$] reached significance.

Detailed analysis of the critical points shows that, for the native norm, the turning point (valley: 2.4) is very low relative to both the peak at the onset (falling range: 1.2) and the offset (rising range: 1.1), and appears relatively late (55% from the onset) in the contour. In comparison, the turning point of the pretest production is not as low (3.2), and occurs in the initial portion (25%) of the contour. For the post-test production, although the turning point pitch value (3.1) is similar to that of the pretest, its position (65%) is much closer to the native norm, appearing relatively late in the contour as compared to that of the pretest. Moreover, the peak of the post-test production occurs in the same position as that of the native norm (0%), different from the pretest production peak position (100%).

Overall, the post-test production of tone 3 is significantly more similar to the native norm than the pretest production. Analyses show that the approximation is more in terms of pitch shape than pitch height [also cf. Fig. 3(c)], indicating that the post-test production has not fully reached the native norms. These results are supported by the native

TABLE V. Tone 3: Mean T values at 0%, 25%, 50%, 75%, and 100% (with standard deviations, s.d., in parentheses) of the pitch contour, as well as mean T values at the lowest (valley) and highest (peak) points and their temporal position in % (expressed as a percentage of total time duration) (in brackets), and the falling pitch range (the difference between onset and valley) and the rising pitch range (the difference between valley and offset) of the pitch contour of the native norm, and trainees' pre- and post-test productions.

	0% (s.d.)	25% (s.d.)	50% (s.d.)	75% (s.d.)	100% (s.d.)	Valley [position]	Peak [position]	Falling range	Rising range
Native norm	3.6 (0.6)	3.1 (0.4)	2.7 (0.4)	2.9 (0.5)	3.5 (0.3)	2.4 [55%]	3.6 [0%]	1.2	1.1
Pretest	3.4 (0.7)	3.2 (0.7)	3.2 (0.7)	3.4 (0.6)	3.7 (0.6)	3.2 [25%]	3.7 [100%]	0.2	0.5
Post-test	3.6 (0.6)	3.4 (0.5)	3.1 (0.6)	3.2 (0.5)	3.5 (0.4)	3.1 [65%]	3.6 [0%]	0.5	0.4

TABLE VI. Tone 4: Mean *T* values at 0%, 25%, 50%, 75%, and 100% (with standard deviations, s.d., in parentheses) of the pitch contour, as well as mean *T* values at the highest (peak) and lowest (valley) points, and their temporal position in % (expressed as a percentage of total time duration) (in brackets), and pitch range of the pitch contour of the native norm, and trainees' pre- and post-test productions.

	0% (s.d.)	25% (s.d.)	50% (s.d.)	75% (s.d.)	100% (s.d.)	Peak [position]	Valley [position]	Falling range
Native norm	4.4 (0.5)	4.3 (0.5)	4.1 (0.5)	3.7 (0.4)	3.4 (0.5)	4.4 [0%]	3.4 [100%]	1.0
Pretest	4.0 (0.6)	4.0 (0.6)	3.7 (0.5)	3.5 (0.5)	3.4 (0.6)	4.0 [0–25%]	3.4 [100%]	0.6
Post test	4.2 (0.6)	4.1 (0.6)	3.8 (0.6)	3.5 (0.4)	3.2 (0.5)	4.2 [0%]	3.2 [100%]	1.0

Chinese listeners' evaluation in that although significantly more post-test productions (46%) were judged to be correct than pretest productions (25%), the overall percent-correct productions is still relatively low in the post-test.

Table VI shows the pitch values for tone 4 at 0%, 25%, 50%, 75%, and 100% of the pitch contour, as well as the highest and lowest pitch values, for the native norm and the trainees' productions.

A two-way repeated measures ANOVA (test×position) with deviation score as dependent variable was conducted. An analysis of the difference between the pretest and the post-test as compared to the native norm at each of these 5 points shows that the post-test production (deviation score: 0.37) is significantly more similar to the native norm than the pretest production (deviation score: 0.47) $[F(1,34)=5.8,\ p<0.022]$. Furthermore, there is no significant effect of position $[F(1.34)=0.1,\ p>0.977]$ nor a position by test interaction $[F(4,34)=0.5,\ p>0.709]$.

As a high-falling tone, the native norm shows a high initial pitch (4.4) followed by a relatively steep fall (range: 1.0). Although both the pretest (4.0) and post-test (4.2) tone 4 productions start at a relatively high pitch, post-test values are higher than pretest values and consequently closer to the native norm. In terms of falling range, the post-test (range: 1.0) resembles the native norm more closely than the pretest (range: 0.6).

Overall, the post-test production of tone 4 is significantly more similar to the native norm than the pretest production. These results are also consistent with the native listener judgment data, showing that the percentage of productions that was identified as correct was greater in the post-test (78%) than in the pretest (73%). Both the acoustic analysis and the native listener evaluation data indicate that the pretest productions improved in the post-test.

The above results show that the native productions are consistent with the patterns described in Chao (1948) and Wu (1986) for the four Mandarin tones, showing a high-level pitch contour for tone 1, a mid-high rising contour for tone 2, a low-dipping contour for tone 3, and a high-falling contour for tone 4. The American learners' results show that, as a consequence of perceptual training, the post-test productions approximate the native productions more closely than do the pretest productions.

In terms of individual tones, the present results show that among the four tones, although the learners' pretest production of tone 1 was similar to the native tone 1 both in terms of pitch height and contour, tone 1 was even more native-like in the post-test. The ease of articulation for tone 1 might be attributed to the fact that, since the contour shape is constant, learners only need to grasp the single dimension of pitch height to correctly produce this tone. That the production of tone 1 is relatively easy as compared to the other three tones for American learners of Mandarin has also been reported in previous studies analyzing learners' productions both in laboratory (Leather, 1983) and classroom (Miracle, 1989) settings. It is also noted that, despite the closer approximation to a level contour, the post-test production still did not achieve full accuracy in terms of pitch height.

Tone 2 and tone 3 have consistently been found to be confusing in both first language acquisition (e.g., Li and Thompson, 1977; Clumeck, 1980) and second language acquisition (e.g., Leather, 1983; Miracle, 1989) studies. The confusion may well be due to the acoustic similarities of these two tones (Leather, 1990), in that they both involve a falling followed by a rising contour. However, differences exist in that the rising contour for tone 2 starts much earlier and ends much higher than that for tone 3. In addition, the valley for tone 2 is not as low as that for tone 3. In the present study, learners' pretest tone 2 rising contour started relatively late and did not reach high levels at the offset, showing more resemblance to a tone 3 pattern. However, in the post-test, both the frequency and temporal position of the valley, as well as the contour offset height closely approximated the native patterns.

Learners' pretest tone 3 productions reveal a rising contour starting relatively early just as a typical tone 2. Although in their post-test production the turning point position shifted later toward the native tone 3 direction, its frequency was still not as low as that of the native turning point. Together, these data show that tone 2 and tone 3 were confusable for the American learners.

The learners' pretest production of tone 4 is different from the native norm in two dimensions: it starts at a lower pitch, and its slope is less steep. In the post-test, the pitch range of the learners' productions was significantly increased, which resembled the native value. Although post-test pitch height was also increased, it was still lower than the native norm.

Taken together, the analysis of the productions of the four tones in pretest and post-test seems to suggest that the two dimensions of pitch height and pitch contour are not

TABLE VII. Tone pair confusion patterns for perception and production at pretest and post-test, in terms of percent errors for each tone pair.

	Pretest errors (%)		Post-test errors (%)	
	Perception	Production	Perception	Production
Tones 2&3	25	33	8.3	23
Tones 2&4	10	17	3.5	5.0
Tones 1&2	8.8	9.0	2.8	4.4
Tones 1&4	7.3	11	5.5	8.4
Tones 3&4	5.0	7.5	1.0	5.0
Tones 1&3	2.5	1.5	0	1.5

mastered in parallel. As compared to pitch contour, pitch height is more resistant to improvement.

IV. RELATION BETWEEN PRODUCTION AND PERCEPTION

Wang *et al.* (1999) showed that, after perceptual training of Mandarin tone, the American trainees' identification greatly improved (21% increase). The present study demonstrated that their production accuracy also increased significantly, indicating a relation between the perception and production of Mandarin tone. Consequently, American trainees' perception and production of Mandarin tone in the pretest and post-test are compared to examine the nature of this relationship.

The tone confusion data of the perception results (Wang *et al.*, 1999) show that, in the pretest, the most confusing tone pair was tones 2 and 3, followed by tones 2 and 4, tones 1 and 2, tones 1 and 4, tones 3 and 4, tones 1 and 3. This rank order was mostly retained in the post-test, except that tones 1 and 4 became the second most confusing pair. Interestingly, the present tone production confusion results reveal strikingly similar patterns in both pretest and post-test.

A comparison of the perception and production confusion patterns in the pretest and post-test is shown in Table VII, in terms of the percent errors for each tone pair. As shown in the table, in the pretest the percent errors for perception and production are highly correlated [$r(5) = 0.98$, $p < 0.0001$]. The rank order in terms of tone pair is also highly correlated for perception and production [$\rho(5) = 0.94$, $p < 0.005$]. Similarly, perception and production are significantly correlated in the post-test, both in terms of errors [$r(5) = 0.9$, $p < 0.01$] and in terms of tone pair rank order [$\rho(5) = 0.9$, $p < 0.015$].

These results show that trainees' tone perception and production are highly related. However, despite this general consistency, differences do exist between perception and production.

It is noted that although tone pair 2 and 3 is the most confusing pair for both perception and production, the direction of confusion is different for these two modalities. That is, tone 2 was incorrectly perceived as tone 3 more frequently than tone 3 was incorrectly perceived as tone 2. In contrast, tone 3 was incorrectly produced as tone 2 more frequently than the reverse. Similar patterns are also found for tones 3 and 4, in that tone 3 was more often incorrectly produced as tone 4, but less often perceived as tone 4. These patterns are illustrated in Table VIII.

TABLE VIII. Confusion patterns for tone pair 2 and 3, and tone pair 3 and 4 at pretest and post-test, in terms of percent errors of the total number of stimuli, showing the difference between perception and production in terms of confusion direction.

	Pretest errors (%)		Post-test errors (%)	
Correct−incorrect	Perception	Production	Perception	Production
Tone 2 as tone 3	16	4	5.6	2
Tone 3 as tone 2	9	29	2.7	21
Tone 3 as tone 4	1.7	4.8	0	4
Tone 4 as tone 3	3.2	2.6	1	1

Comparing the pretest and post-test data, it is also noted that tone pair 2 and 3 errors decreased to a large degree in perception. However, a similar decrease is not as evident in production comparing pretest to post-test. Similarly, tones 3 and 4 did not improve greatly in the production post-test.

These patterns are also reflected in the overall results for individual tones, in that the perception of tone 3 was relatively good to start with and significantly improved after training (see Fig. 2 in Wang *et al.*, 1999), whereas its production was poor in the pretest and remained so in the post-test (see Fig. 2). Taken together, these data show that while tone 3 was relatively easy to identify, it was difficult to produce, and was resistant to improvement.

V. DISCUSSION AND CONCLUSIONS

The present study shows that, after perceptual Mandarin tone training, the American learners' productions of Mandarin tone improved without any production training. The native Chinese listeners' evaluation of the trainees' pretest and post-test productions indicates that, after training, there was an improvement for each of the four tones, and the improvement in production was even extended to novel stimuli which were not used in the perception training. Native Chinese listeners' evaluation of the controls' pretest and post-test productions did not show a similar improvement. Acoustic analysis consistently revealed that the trainees' post-test productions were significantly more similar to the native norm in terms of both pitch height and contour than were pretest productions. These results indicate that the effect of training in perception transferred to the production domain.

The present study is consistent with previous training studies in the segmental domain showing the transfer of perceptual learning to production, such as the production of French VOT categories by native Chinese speakers (Rochet, 1995), and the production of the English /r–l/ contrast by Japanese learners (Bradlow *et al.*, 1997, 1999; Akahane-Yamada, 1999). Together, these studies coupled with the present results show that the facilitatory effect of perception training on production not only occurs for segmental learning, but also extends to suprasegmental learning.

The facilitatory effect of perception training for production learning supports the view in segmental acquisition research that the two speech modalities are related, with perception ''leading'' production (Flege, 1997). Indeed, the current phonetic learning theories are all perception oriented, stating that perceptual experience can guide sensory-motor learning (Kuhl, 2000a, b), and the accuracy with which L2 segments are perceived limits how accurately they can be produced (Flege, 1999). Studies of non-native segmental acquisition have found significant, albeit modest, correlations between learners' perception and production of L2 vowels (e.g., Flege, Bohn, and Jang, 1997) and consonants (e.g., Flege, 1993), showing that production accuracy is constrained by perception accuracy. The results in the present study provide supporting evidence of the nature of this relationship in suprasegmental learning. While the high correlation of the tone pair confusion patterns of the pretest perception and production shows the relationship of these two domains, the high correlation of the post-test perception and production tone pair confusion patterns clearly demonstrates how perceptual learning guided production. For example, the tone pairs that had been greatly improved perceptually, e.g., tones 2 and 3, tones 2 and 4, tones 1 and 2, also showed great improvement in production. In contrast, tone pair 1 and 4, which was most resistant to improvement in perception, had minimally improved in production as well.

Despite the general claim of a positive correlation between perception and production, the learning of these two modalities may not always be in parallel, as not all aspects of perceptual learning can be incorporated in production (Flege, 1999). Flege (1999) further pointed out that not all instances of non-native phonetic production have a perceptual origin. Some segments that are not used in learners' L1 phonetic system may present difficulty for production learning. The present results may also provide some support for this segment-based claim. Although, after training, tone 3 became relatively easy to perceive, it remained difficult to produce. Table VIII further revealed that, contrary to the patterns for perception, more tone 3 stimuli were incorrectly produced as tone 2 or tone 4. On this account, the difficulty in the production of tone 3 might not be due to a failure in perception learning, but rather to the novelty of the sound itself. It might be that the low dipping nature of the tone 3 pitch contour is so unfamiliar to the American learners that it makes articulation difficult. It is not known whether additional perceptual (or production) training may improve this particular tone.

The relationship of perception and production learning suggests an integrated system underlying these two mechanisms. Thus, learning a speech contrast involves mastery of both perception and production. The present finding that the effect of perceptual training not only extended to new speech contexts but was also transferred to the production domain further indicates that perceptual training results in highly generalized learning, suggesting that new tonal categories might have been established as a consequence of training. Taken together, the present results concerning the training of suprasegmental contrasts are consistent with the notion of a malleability of the adult learner's speech learning system across both perception and production.

ACKNOWLEDGMENTS

This research was conducted as part of a doctoral dissertation at Cornell University by the first author under the direction of the second and third authors. Portions of this study were reported at the 137th meeting (1999, Berlin), and the 142nd meeting of the Acoustical Society of America (2001, Fort Lauderdale). We would like to thank Michelle Spence, Brian Kim, Alicia Mackay, Eric Evans, Dorris Lok, and Bertrand Kotewall for assistance and technical support. Part of this research was supported by a grant from the University of Kansas General Research Fund to the third author.

Akahane-Yamada, R. (**1999**). ''Toward further understanding of second language speech learning: An approach utilizing speech technology,'' J. Acoust. Soc. Am. **105**, 1032.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. (**1990**). ''Effects of syllable duration on the perception of the Mandarin tone2/tone3 distinction: Evidence of auditory enhancement,'' J. Phonetics **18**, 37–49.

Bluhme, H., and Burr, R. (**1971**). ''An audio-visual display of pitch for teaching Chinese tone,'' Stud. Linguist. **22**, 51–57.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., and Tohkura, Y. (**1997**). ''Training Japanese listeners to identify English /r/ and /l/. IV. Some effects of perceptual learning on speech production,'' J. Acoust. Soc. Am. **101**, 2299–2310.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. (**1999**). ''Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production,'' Percept. Psychophys. **61**, 977–985.

Chao, Y. R. (**1948**). *Mandarin Primer* (Harvard University Press, Cambridge).

Clumeck, H. (**1980**). ''The acquisition of tone,''in *Child Phonology*, edited by G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson (Academic, New York), Vol. I.

Dreher, J. J., and Lee, P. C. (**1966**). *Instrumental Investigation of Single and Paired Mandarin Tonemes* (Advanced Research Laboratory, Douglas Aircraft Company, Huntington Beach, CA).

Flege, J. E. (**1993**). ''Production and perception of a novel, second-language phonetic contrast,'' J. Acoust. Soc. Am. **93**, 1589–1608.

Flege, J. E. (**1997**). ''The role of phonetic category formation in second-language speech learning,'' in *New Sounds 97. Proceedings of the Third International Symposium on the Acquisition of Second-Language Speech*, edited by J. Leather and A. James, pp. 79–88.

Flege, J. E. (**1999**). ''The relation between L2 production and perception,'' in *Proceedings of the XIVth International Congress of Phonetics Sciences*, edited by J. Ohala, Y. Hasegawa, M. Ohala, D. Granveille, and A. Bailey (Department of Linguistics, University of California at Berkeley, Berkeley, CA), pp. 1273–1276.

Flege, J. E., Bohn, O-S., and Jang, S. (**1997**). ''The production and perception of English vowels by native speakers of German, Korean, Mandarin, and Spanish,'' J. Phonetics **25**, 422–470.

Fox, R., and Qi, Y. Y. (**1990**). ''Context effects in the perception of lexical tone,'' J. Chin. Ling. **18**, 261–283.

Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge, MA).

Jamieson, D. G., and Morosan, D. E. (**1986**). ''Training non-native speech contrasts in adults: Acquisition of the English /θ/–/ð/ contrast by francophones,'' Percept. Psychophys. **40**, 205–215.

Jamieson, D. G., and Morosan, D. E. (**1989**). ''Training new, non-native speech contrasts: A comparison of the prototype and perceptual fading techniques,'' Can. J. Psychol. **43**, 88–96.

Kiriloff, C. (**1969**). ''On the auditory discrimination of tones in Mandarin,'' Phonetica **20**, 63–67.

Kuhl, P. K. (**2000a**). ''Language, Mind, and Brain: Experience alters perception,'' in *The New Cognitive Neurosciences*, 2nd ed., edited by M. S. Gazzaniga (The MIT Press, Cambridge, MA), pp. 99–115.

Kuhl, P. K. (**2000b**). ''A new view of language acquisition,'' Proc. Natl. Acad. Sci. U.S.A. **97**, 11850–11857.

Ladd, D. R., Silverman, K., Tolkmitt, F., Bergmann, G., and Scherer, K. (**1985**). ''Evidence for the independent function of intonation contour type, voice quality, and *F*0 range in signaling speaker affect,'' J. Acoust. Soc. Am. **78**, 435–444.

Leather, J. (**1983**). ''Speaker normalization in perception of lexical tone,'' J. Phonetics **11**, 373–382.

Leather, J. (**1990**). ''Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers,'' in *New Sounds 90: Proceedings of the Amsterdam Symposium on the Acquisition of Second Language*

Speech, edited by J. Leather and A. James (University of Amsterdam, Amsterdam).

Li, C. N., and Thompson, S. (**1977**). ''The acquisition of tone in Mandarin-speaking children,'' J. Child Lang **4**, 185–199.

Liao, R. (**1983**). ''Suzhouhua danzidiao shuangzidiao de shiyan yanjiu,'' Yuyan Yanjiu **5**, 24–50.

Lin, T., and Wang, W. Y.-S. (**1985**). ''Shengdiao ganzhi wenti,'' Zhongguo Yuyan Xuebao **2**, 59–69.

Liu, F. (**1924**), *Szu Sheng Shih Yen Lu* (Ch'un Yi, Shanghai).

Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). ''Training Japanese listeners to identify English /r/ and /l/. II. The role of phonetic environment and talker variability in learning new perceptual categories,'' J. Acoust. Soc. Am. **94**, 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., and Yamada, T. (**1994**). ''Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories,'' J. Acoust. Soc. Am. **96**, 2076–2087.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). ''Training Japanese listeners to identify English /r/ and /l/: A first report,'' J. Acoust. Soc. Am. **89**, 874–886.

McClaskey, C. L., Pisoni, D. B., and Carrell, T. D. (**1983**). ''Transfer of training of a new linguistic contrast in voicing,'' Percept. Psychophys. **34**, 323–330.

Mertus, J. (**1989**). BLISS *Manual* (Brown University, Providence).

Miracle, W. C. (**1989**). ''Tone production of American students of Chinese: A preliminary acoustic study,'' J. Chin. Lang. Teach. Assoc. **24**, 49–65.

Moore, C. B., and Jongman, A. (**1997**). ''Speaker normalization in the perception of Mandarin Chinese tones,'' J. Acoust. Soc. Am. **102**, 1864–1877.

Pisoni, D. B., Aslin, R. N., Perey, A. J., and Hennessy, B. L. (**1982**). ''Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants,'' J. Exp. Psychol. Hum. Percept. Perform. **8**, 297–314.

Rochet, B. L. (**1995**). ''Perception and production of second-language speech sounds by adults,'' in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York, Baltimore), pp. 379–410.

Rose, P. (**1987**). ''Considerations in the normalization of the fundamental frequency of linguistic tone,'' Speech Commun. **6**, 343–351.

Shen, X. S. (**1989**). ''Toward a register approach in teaching Mandarin tones,'' J. Chin. Lang. Teach. Assoc. **24**, 27–47.

Shen, X. S., and Lin, M. C. (**1991**). ''A perceptual study of Mandarin tones 2 and 3,'' Lang Speech **34**, 145–156.

Shi, F. (**1986**). ''Tianjin fangyan shuangzizu shengdiao fenxi,'' Yuyan Yanjiu **10**.

Shi, F. (**1994**). ''Beijinghua de shengdiao geju,'' in *Yuyin Conggao*, edited by F. Shi and R. Liao (Beijing Foreign Language Institute Press, Beijing), pp. 10–19.

Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (**1999**). ''Training American listeners to perceive Mandarin tones,'' J. Acoust. Soc. Am. **106**, 3649–3658.

Wu, Z. J. (**1986**). *The Spectrographic Album of Mono-syllables of Standard Chinese* (Social Science, Beijing).