Production and perception of a novel, second-language phonetic contrast

James Emil Flege
Department of Biocommunication, University of Alabama at Birmingham, Birmingham, Alabama 35294-0019

(Received 7 September 1991; revised 3 September 1992; accepted 27 November 1992)

Four experiments, all of which focused on vowel duration, assessed Chinese subjects’ production and perception of the contrast between /t/ and /d/ in the final position of English words. Vowel duration was measured in minimal pairs in the first experiment. The stimuli in natural-edited beat-bead and bat-bad continua in which vowel duration varied in 20-ms steps were then presented to native English and Chinese subjects in a forced-choice test, in an experiment using the method of adjustment, and in an imitation task. The non-natives who learned English in childhood closely resembled native speakers in all four experiments. Three groups of non-natives who had learned English as a second language in adulthood, on the other hand, differed from the native speakers. The late learners produced significantly longer vowels in words ending in /d/ than /t/. However, the late learners’ vowel duration differences were much smaller than the native speakers’, and were correlated significantly with degree of foreign accent in English. The late learners differed from the native speakers in several ways in the two perception experiments, and also in the imitation task. The pattern of significant and nonsignificant between-group differences, but not data for individual subjects, was consistent with the hypothesis that L2 (second language) production accuracy is limited by the adequacy of perceptual representations for sounds in the L2.

PACS numbers: 43.70.Kv, 43.71.Hw

INTRODUCTION

This study assessed production and perception of the word-final English /t/-/d/ contrast by individuals who had learned English as a second language either in childhood (“child learners”) or in adulthood (“late learners”). Flege (1992a,b) hypothesized that late learners may eventually add phonetic categories to their existing phonetic inventory in order to accommodate second language (henceforth L2) sounds that differ substantially from any sound in the native language (L1). He claimed that this is much less likely to occur, on the other hand, for L2 sounds that are related to a corresponding sound in the L1 yet differ acoustically from the L1 counterpart. For such “similar” L2 sounds, phonetic category formation may be blocked by the perceptual mechanism of equivalence classification. The hypothesized difference in how new and similar sounds are treated perceptually leads to the prediction that new but not similar sounds in an L2 may be mastered eventually by adult L2 learners. The prediction concerning similar L2 consonants has been confirmed in a number of previous studies (e.g., Flege, 1991). The primary aim of the present study was to test the prediction concerning L2 consonants that might potentially be treated as new. The L1 of some late learners tested in the present study has neither /t/ nor /d/ in word-final position; the L1 of other non-natives lacks a word-final /d/. If the late learners treated word-final English /d/’s and/or /t/’s as new then, by hypothesis, they should master the word-final English /t/-/d/ contrast.

Of the numerous acoustic differences between English voiced and voiceless stops in word-final position, vowel duration has probably received the most attention, although some researchers regard closure voicing as more important (Kluender et al., 1988). The great interest in vowel duration arises from its apparently unique status in English. In pre-pausal position, the effect of stop voicing on preceding vowel duration is larger in English than in other languages with final stops (e.g., Chen, 1970). As one might expect from this, vowel duration influences English speakers’ perception of the voicing feature in word-final final stops and fricatives (e.g., Raphael, 1972). Of course, so do other dimensions such as F1 offset frequency, stop closure duration, the presence/absence of closure voicing, and the duration, intensity, and spectrum of the final release burst (e.g., Hogan and Rozsypal, 1980; Fischer and Ohde, 1990). The present study nevertheless focused on vowel duration because vowel duration is measured easily, and an incorrect specification of vowel duration contributes to foreign accent (Elsendoorn, 1984). Also, production and perception of vowel duration can be compared readily. Given the diversity of cues to the voicing feature in final stops, the production of nativelike vowel durations would not, in itself, demonstrate mastery of the English /t/-/d/ contrast. However, the absence of nativelike vowel durations would certainly demonstrate a lack of mastery, especially given that children acquiring English as their L1 begin producing large vowel duration differences in /t/ versus /d/-final words at an early age (e.g., Naeser, 1970).

Previous research suggests that most adult L2 learners produce smaller voicing effects than native speakers, at least in list-reading experiments (see Flege, 1988a). It also appears that the number and nature of stop consonants in non-natives’ L1 may influence the magnitude of voicing effects.
Evidence of cross-language phonetic interference can be seen clearly in French-accented English, for example. Differences in the duration of vowels in words ending in /p t k/ vs /b d g/ is smaller in French than English, and French speakers of English produce smaller voicing effects in English words than do native speakers (Mack, 1982; Flege and Hillenbrand, 1986). The implication is that French speakers of English identify /b d g/ and /p t k/ tokens in English words in terms of the corresponding French stop categories, and produce them accordingly.

Cross-language phonetic interference is also evident in Arabic-accented English. Native speakers of Arabic produce smaller voicing effects in English words than native speakers, apparently because there is very little if any effect of stop voicing on preceding vowel duration in Arabic (e.g., Flege and Port, 1981; Mitleb, 1981; Munro, 1993). Here, however, a clear case cannot be made that an incorrect specification of vowel duration in English words has a perceptual basis. Flege (1984) examined the identification of final fricates in a synthetic peace–peas continuum in which vowel duration varied. Inexperienced Arabic speakers of English showed as large an effect of vowel duration on /s/ vs /z/ judgments as did native speakers, even though vowels are about equally long before /s/ and /z/ in Arabic (Mitleb, 1981). One might interpret these results to mean that the production of vowel duration is unrelated to the perceptual awareness, or use of, vowel duration cues (see Obler, 1982).

This could account for why some Japanese subjects, for example, may produce English /t/ and /l/ better than they identify these non-Japanese sounds (Goto, 1971; Sheldon and Strange, 1982). However, other L2 research suggests that L2 productive abilities lag behind the development of perceptual abilities (Snow and Hoefnagel-Höhle, 1979; Elsendoorn, 1984). Perhaps Arabic subjects in the Flege (1984) study did not make phonetic use of vowel duration as a cue to the /s/-/z/ contrast. They may have "reinterpreted" vowel duration differences in the stimuli as a phonemic vowel length contrast (see Flege, 1988a, p. 286), or else arbitrarily assigned relatively long-duration stimuli to one button, and short-duration stimuli to the other button, without actually perceiving different final consonants or vowels (Crowther and Mann, 1992).

Elsendoorn (1984) examined Dutch subjects' production and perception of vowel duration in English words ending in /p t k/ and /b d g/. The effect of the voicing feature in word-medial (intervocalic) stops on preceding vowel duration in Dutch words is nearly as large as the effect seen in English (see Staatsen and Leijten, 1976). The effect in Dutch is much smaller for vowels preceding word-final stops, however, because there the distinction between voiced and voiceless stops is neutralized by a phonological rule. In one production experiment, Elsendoorn noted very small voicing effects in English words that had been read from a list by Dutch high school students. He nevertheless observed English-like voicing effects in a sentence repetition task. In an identification experiment, vowel duration had a larger effect on experienced Dutch subjects' judgments of the final stop in CVC English words than it did for less experienced Dutch subjects. In still another production experiment, Elsendoorn had Dutch subjects read English words in isolation, then adjust the duration of vowels in English words such as lit and lid until the words sounded "correct." Dutch university students majoring in English produced even larger voicing effects than native speakers. The least experienced of the four Dutch groups who were tested produced virtually no voicing effect. They did, however, adjust vowels to be longer in words ending in /d/ than /t/. Significant correlations between produced and perceived (i.e., adjusted) vowel durations were noted, even for inexperienced Dutch subjects. This led Elsendoorn to conclude that Dutch learners of English establish an "internal representation" of the vowel duration correlate of word-final stop voicing, albeit one that may be less "sharply defined" than native speakers' (1984, p. 100).

Although the Dutch results just reported support the hypothesis that perceptual sensitivity to vowel duration precedes its appearance in production, they might not generalize to speakers of other languages. Certain pairs of vowels in Dutch differ primarily according to duration. Perhaps the existence of phonemic vowel length distinctions in the L1 will heighten sensitivity to vowel duration differences in English words. Perhaps, too, the Dutch subjects were able to transfer their use of vowel duration cues to word-medial Dutch stops to word-final English stops. The studies that have most directly addressed the claim that new L2 consonants can be mastered by adult learners are those that have examined speakers of languages that do not have word-final stops. Transcriptions provided by Eckman (1981) suggested that speakers of Mandarin produced voiced and voiceless word-final English stops accurately (see also Hiramatsu, 1990). However, several recent studies have failed to support the prediction that speakers of L1's without final stops will master word-final English stops (Flege, 1988c; Crowther and Mann, 1992; Flege et al., 1992). Most non-native subjects examined in those studies produced smaller vowel duration differences than native speakers.

Four experiments were carried out in the present study. The duration of vowels in English minimal pairs ending in /t/ vs /d/ was measured in experiment 1. In experiment 2, subjects identified, as /t/ or /d/, the final stops in two natural-edited continua in which vowel duration had been varied in 20-ms steps. The same two continua were used in experiment 3, which employed the method of adjustment to assess the subjects' perceptual sensitivity to vowel duration as a cue to the voicing feature in word-final stops. In experiment 4, subjects imitated the duration of vowels in the two continua and in a continuum of isolated vowels. In addition to assessing the relation between the production and perception of vowel duration as a cue to the word-final English /t/-/d/- distinction, the present study tested differences between groups of subjects differing in the age at which English was learned and/or English-language experience.

Performance of subjects in a native English group was compared to that of subjects in four non-native groups. The "childhood" (or "child") L2 learners, most of whom were native speakers of Taiwanese, were first exposed to English as children. Two groups of mostly Taiwanese "late learners," who began learning English as adults, differed primar-
ily according to length of residence in the U.S. The native Mandarin late learners in another group were matched, in terms of length of residence in the U.S., to the less experienced of the two Taiwanese groups. Subjects in the child learner group arrived in the U.S. between the ages of 3 years and 13 years. The latter age is often said to demarcate individuals who will or will not speak an L2 with an accent (e.g., Scovel, 1988), although more recent research suggests that foreign accents first emerge at around that age of 7 years (Flege and Fletcher, 1992). Given that an incorrect specification of vocal duration may contribute to foreign accent (Elsendoorn, 1984), some subjects in the child learner group might be expected to produce smaller voicing effects, or to show a smaller influence of vowel duration on perceptual judgments, than the native speakers. The experienced Taiwanese late learners were compared to the relatively inexperienced Taiwanese late learners. Although results cited earlier suggest that the experienced Taiwanese late learners would not produce nativelike vowel duration differences in words ending in /t/ vs /d/, it was of interest to learn if they would also show smaller perceptual effects of vowel duration than the native speakers.

The inexperienced Mandarin and Taiwanese late learners were also compared. Taiwanese has words ending in /p t k/ but not /b d g/; Mandarin permits no obstructions of any kind at the end of words (Cheng, 1973; Norman, 1988). Results obtained by Flege and Wang (1990) suggest that native speakers of Taiwanese may allocate more attention than native Mandarin speakers to word endings because of differences in the range of word-final consonants in the two languages. If so, Taiwanese speakers might “pick up” perceptual cues to the word-final /t/-/d/ distinction more readily than Mandarin speakers when they learn English. Competing predictions could be made regarding how the Mandarin and Taiwanese subjects would produce word-final English stops. If final L2 consonants are related perceptually only to word-final consonants in the L1 (which is plausible given allophonic differences between consonants in word-initial and word-final position), then the Mandarin subjects might produce larger voicing effects than the Taiwanese subjects.1 The contrastive analysis notion that “different is difficult” (e.g., Moisio and Valento, 1976), on the other hand, implies that Taiwanese speakers’ previous experience producing stops in word-final position might give them an advantage in producing both /t/ and /d/ in the final position of English words.

I. EXPERIMENT 1

The purpose of this experiment was to measure vowel durations in English words ending in /t/ and /d/ that had been spoken by native speakers of English and by four groups of non-native speakers.

A. Method

1. Subjects

Characteristics of the five groups of subjects who participated are summarized in Table I. There were nine subjects in the “child learner” group, but ten subjects in all other groups. All subjects were living in Birmingham, Alabama at the time of testing; and most were affiliated with the University of Alabama at Birmingham. Subjects in the native English group spoke only American English. The subjects in the remaining four Chinese groups are designated TA (inexperienced Taiwanese late learners), TB (experienced Taiwanese late learners), TC (childhood L2 learners), and MA (inexperienced Mandarin late learners) in figure legends. Most subjects in the groups with a “T” designation were native speakers of Taiwanese from Taiwan. The few who were not spoke a native language (Cantonese, Fukienese, Hunanese) which, like Taiwanese, has words ending in /p t k/ but not /b d g/, and had learned Taiwanese as young children. They were expected to perform just like the subjects who had learned Taiwanese as an L1. The principal difference between the inexperienced native Mandarin subjects and the inexperienced native Taiwanese subjects was native language. Most inexperienced Mandarin subjects were from the Beijing region. The principal difference between the experienced and inexperienced Taiwanese late learners was length of residence in the U.S. (1.2 vs 5.1 years). The experienced Taiwanese late learners differed from the child learners according to length of residence in the U.S. but—more importantly—these groups differed according to age of L2 learning. The experienced Taiwanese late learners arrived in the U.S. at an average age of 25 years, the child learners at an average age of just eight years (range = 3 to 13 years).

2. Procedures

The production experiment was carried out first to avoid possible effects of exposure to the perceptual stimuli on production. The vowels /ɪ/, /ʌ/, /æ/, /ɒ/, /eɪ/, /e/, and /u/ were inserted into /b t/ and /b d/ frames, yielding seven minimal pairs. Subjects read seven randomized tokens of each word from a list at the end of a carrier phrase (I will say __). The speech samples were recorded in a sound booth using a cassette tape recorder (Sony model TCD5M). The subjects were told to produce the utterances conversationally, maintaining a constant rate and loudness level. Produc-

---

**TABLE I. Characteristics of the subjects in five groups who participated in experiments 1 to 4. Standard deviations are in parentheses.**

<table>
<thead>
<tr>
<th>Group</th>
<th>Age</th>
<th>Gender</th>
<th>LOR</th>
<th>AOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA: inexperienced</td>
<td>35.0</td>
<td>5 males</td>
<td>1.1</td>
<td>34.0</td>
</tr>
<tr>
<td>Mandarin late learners</td>
<td>35.2</td>
<td>5 females</td>
<td>0.9</td>
<td>6.1</td>
</tr>
<tr>
<td>TA: inexperienced</td>
<td>28.4</td>
<td>9 males</td>
<td>1.2</td>
<td>27.6</td>
</tr>
<tr>
<td>Taiwanese late learners</td>
<td>1.6</td>
<td>1 female</td>
<td>0.7</td>
<td>2.0</td>
</tr>
<tr>
<td>TB: experienced</td>
<td>30.5</td>
<td>7 males</td>
<td>5.1</td>
<td>25.4</td>
</tr>
<tr>
<td>Taiwanese late learners</td>
<td>2.4</td>
<td>3 females</td>
<td>2.7</td>
<td>2.3</td>
</tr>
<tr>
<td>TC: childhood</td>
<td>21.2</td>
<td>5 males</td>
<td>12.7</td>
<td>8.3</td>
</tr>
<tr>
<td>L2 learners</td>
<td>4.7</td>
<td>4 females</td>
<td>3.8</td>
<td>3.3</td>
</tr>
<tr>
<td>NE: native English speakers</td>
<td>25.6</td>
<td>7 males</td>
<td>5.1</td>
<td>25.4</td>
</tr>
<tr>
<td>TA: inexperienced</td>
<td>28.4</td>
<td>9 males</td>
<td>1.2</td>
<td>27.6</td>
</tr>
<tr>
<td>Mandarin late learners</td>
<td>5.2</td>
<td>5 females</td>
<td>0.9</td>
<td>6.1</td>
</tr>
</tbody>
</table>

---

1. Age, chronological age, in years.
2. LOR, length of residence in the U.S., in years.
3. AOA, age of arrival in the U.S., in years.
tion of the utterances was modeled by the experimenter, who said non-test CVCs at the end of the carrier phrase. She spoke the model utterances at a moderate rate, placing main sentence stress on the variable CVCs. The experimenter instructed subjects to begin again if she perceived them to have begun speaking rapidly or slowly.

3. Measurement

The middle five repetitions of each test word were measured acoustically. Vowel duration was measured to the nearest 4 ms using a PM pitch analyzer (Voice Identification, Inc.), from the first sampling point in the intensity trace that exceeded 66% of the eventual peak syllable intensity to the last above-zero data point in the intensity trace. The segmentation criteria excluded the release burst of /b/ (when one was present) and, at the amplitude setting employed, the closure voicing in /d/ (when present). A single research assistant made all measurements. To assess measurement reliability and validity, 30 randomly selected words (14 ending in /d/, 16 in /t/) spoken by a female native speaker were examined. These words were each measured twice by the assistant using the PM pitch analyzer, a Kay (model 7800) sound spectrograph, and a digital waveform editing program. The author also measured the 30 words using the PM pitch analyzer. As summarized in Table II, the mean values obtained using the PM device were very similar to those obtained using the other two measurement techniques in terms of mean values and variability. The two sets of PM measurements made by the assistant differed by only 0.2 ms on average, and were highly correlated ($r = 0.999$). The differences between replicate pairs of PM measurements ranged from $-4$ to 8 ms. Measurement reliability appeared to be slightly better for the PM device than for the other two techniques. ANOVAs indicated that the three sets of PM measurements (two from the assistant and one from the author) did not differ significantly [$F(2,58) = 2.18, p > 0.10$].

B. Results

Figure 1 shows that the subjects in all five groups made vowels longer before /d/ than /t/. Voicing-conditioned differences in preceding vowel duration averaged 137 ms for the native speakers, 104 ms for child learners, 63 ms for experienced Taiwanese late learners, 39 ms for inexperienced Taiwanese late learners, and 40 ms for the inexperienced Mandarin late learners. The mean duration of vowels in the 35 /t/-final words spoken by each subject was subtracted from the mean duration of vowels in 35 /d/-final words. The average vowel duration differences produced by native English speakers ranged from a low of 84 to a high of 172 ms. Seven (78%) child learners, three (30%) experienced Taiwanese late learners, and just one (5%) of the inexperienced L2 learners produced vowel duration differences that fell within the native English range. The mean vowel duration differences produced by the 49 subjects were examined in a one-way ANOVA, which yielded a significant effect of group [F(4,44) = 22.3, p < 0.01]. Tukey's HSD procedure was used to test for pairwise differences between groups (a procedure used throughout this article). The post-hoc test revealed that the native speakers and child learners produced significantly larger differences than the experienced Taiwanese late learners, the inexperienced Taiwanese late learners, and the inexperienced Mandarin late learners ($p < 0.05$).

TABLE II. Summary of tests evaluating the validity and reliability of vowel duration measurements. Two sets of measurements were made in milliseconds by an assistant ("1" and "2") using the PM pitch analyzer, a sound spectrograph, and a digital waveform editor. The author made a third set of measurements ("3") using the PM device. Each mean is based on 30 measurements of words produced by a single native speaker of English (see text).

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>s.d.</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrograms-1</td>
<td>274</td>
<td>65</td>
<td>268</td>
<td>176</td>
<td>400</td>
</tr>
<tr>
<td>Spectrograms-2</td>
<td>272</td>
<td>66</td>
<td>264</td>
<td>176</td>
<td>392</td>
</tr>
<tr>
<td>Waveform editor-1</td>
<td>272</td>
<td>69</td>
<td>270</td>
<td>155</td>
<td>406</td>
</tr>
<tr>
<td>Waveform editor-2</td>
<td>273</td>
<td>68</td>
<td>275</td>
<td>158</td>
<td>401</td>
</tr>
<tr>
<td>PM device-1</td>
<td>268</td>
<td>71</td>
<td>253</td>
<td>168</td>
<td>391</td>
</tr>
<tr>
<td>PM device-2</td>
<td>268</td>
<td>71</td>
<td>253</td>
<td>168</td>
<td>391</td>
</tr>
<tr>
<td>PM device-3</td>
<td>269</td>
<td>72</td>
<td>254</td>
<td>172</td>
<td>394</td>
</tr>
</tbody>
</table>

FIG. 1. Mean duration, in ms, of vowels in English words spoken by inexperienced Mandarin late learners (MA), inexperienced Taiwanese late learners (TA), experienced Taiwanese late learners (TB), Taiwanese speakers who learned English as a second language in childhood (TC), and native English speakers (NE). The brackets enclose ± 1 standard error.
duced by the native speakers and child learners in *bat–bad* (128, 127 ms) did not differ significantly. The native speakers produced significantly larger differences than the experienced and inexperienced Taiwanese late learners (74, 42 ms) and the inexperienced Mandarin late learners (37 ms) ($p < 0.05$). The child learners produced larger differences than the two groups of inexperienced late learners ($p < 0.05$).

The mean duration of vowels produced by subjects in each of the 14 CVC test words were submitted to a (5) group $\times$ (7) vowel $\times$ (2) final stop voicing ANOVA. It yielded a significant vowel $\times$ final stop interaction [$F(6,264) = 7.71$, $p < 0.01$] and a significant three-way interaction [$F(24,264) = 1.74$, $p < 0.05$]. The simple main effect of final stop (/t/ vs /d/) was significant in all seven minimal pairs for the subjects in all but one group. Subjects in the remaining group, the inexperienced Mandarin late learners, made vowels significantly longer before /d/ than /t/ in just six of the seven minimal pairs ($p < 0.05$). The simple main effect of group was nonsignificant for all seven words ending in /t/ ($p > 0.10$). It was significant, however, for all words ending in /d/ ($p < 0.05$).

**C. Discussion**

The late learners, even those who were relatively experienced in English, produced smaller vowel duration differences than the native speakers, which agrees with findings obtained previously (Flege, 1988c; Flege et al., 1992; Crowther and Mann, 1992). This suggests that the lack of a word-final /t/ vs /d/ contrast in an adult learner’s L1 does not guarantee mastery of this contrast in English. The vowel duration differences produced by the experienced late learners were only about half as large as the native speakers’, but they nevertheless made vowels significantly longer before /d/ than /t/ in all seven minimal pairs examined. Thus it is unlikely that the late learners were unaware that /t/ differs phonologically from /d/ in English. As suggested by Flege and Port (1981), their problem seems to have been one of phonetic implementation. The lack of a significant difference between the experienced and inexperienced Taiwanese late learners suggests that an additional 4 years of English-language experience is insufficient to increase vowel duration differences measurably. This may mean that the presence versus absence of word-final stops in the L1 does not influence greatly how well late learners will produce the English word-final /t/ vs /d/ contrast. Perhaps a difference between groups would have been noted had more experienced Mandarin and Taiwanese subjects been compared.

As hypothesized in the Introduction, the child learners produced significantly smaller differences than the native speakers in one minimal pair (viz., *beat–beat*). They did not differ significantly from the native speakers for *bat–bad*, however; nor did they differ from the native speakers when the vowels in all seven /t/- and /d/-final words were averaged. Greater differences between the native speakers and the child learners might have been observed had more of the child learners first been exposed to English after 7 years of age. Three subjects who arrived in the U.S. between 3 and 6 years of age produced a difference that averaged 125 ms; child learners whose first exposure to English occurred between the ages of 9 and 13 years produced differences averaging just 92 ms.

**II. EXPERIMENT 2**

The 49 subjects next identified the final stops in CVC stimuli as /t/ or /d/. Vowel duration varied in 17 steps in natural-edited continua ranging from *beat* to *bead* and from *bat* to *bad*. Using similar stimuli, Elsendoorn (1984) found that the effect of vowel duration was larger for native speakers of English and experienced Dutch L2 learners than for inexperienced Dutch L2 learners. The experienced Taiwanese subjects in the present experiment might therefore be expected to show a larger effect of vowel duration than the two inexperienced late learner groups, even though they did not produce larger voicing effects than the inexperienced subjects in experiment 1. If so, this would support the hypothesis that perceptual sensitivity to vowel duration precedes the production of large vowel duration differences.

**A. Method**

The results of previous research suggested that, for the purposes of the present study, it would be more appropriate to use natural-edited than synthetic stimuli. Results obtained using synthetic speech might not be replicable with natural stimuli (e.g., Burton et al., 1989). The use of synthetic speech might obscure individual listener differences (Hazan and Rosen, 1991). Response variability might be greater with synthetic than natural stimuli (see Clark et al., 1985). Finally, non-native subjects may perform more poorly with synthetic than natural realizations of novel L2 phonetic contrasts (Mochizuki, 1981).

Two continua were created by cross-splicing, then editing portions of naturally spoken CVC words. Pilot testing with native English listeners revealed that neither the deletion of glottal periods from vowels preceding /d/, nor the repetition of glottal periods in vowels preceding /t/, resulted in a complete shift in stop voicing judgments (see also Hogan and Rozsypal, 1980; Elsendoorn, 1984). The barrier seemed to be the presence of spectral cues in transitions into word-final stops in the natural tokens (see, e.g., Fischer and Ohde, 1990). To augment the effect of vowel duration manipulations, therefore, a portion of /t/ transitions in *beat* and *bat* tokens were substituted for comparable signal portions in *bead* and *bad* tokens. The final stop closures of the *bad* and *bead* tokens were 75 and 102 ms in duration, respectively, with 52 and 68 ms of glottal pulsing. This meant that the final release bursts of the two /d/ tokens were acoustically voiceless. F1 offset frequency was substantially lower in the *bad* than *bat* token (699 vs 1066 Hz), but there was little F1 difference between the *bead* and the *beat* tokens (284 vs 270 Hz).

The last 30.0 ms of final-stop transitions in the *beat* token replaced a 29.2-ms portion in the *bead* token; and a 32.8-ms portion from *bat* replaced a 32.5-ms portion from *bad*. The resulting hybrid stimuli had conflicting cues to the stop voicing feature. The relatively high F1 offset frequency in the /t/ portions (at least in the hybrid stimulus with /x/), and the voiceless release bursts present in the original /d/’s,
might be expected to cue a voiceless stop. The presence of closure voicing and the relatively short duration of the stop closure intervals themselves, on the other hand, might be expected to cue a final voiced stop.

Vowel duration was decreased 16 times in each hybrid stimulus, yielding two 17-member continua. First, a single glottal pulse two-thirds of the way into the vocalic portion of the beat–bead hybrid was repeated until it reached 461 ms. A glottal pulse in the bat–bad hybrid was repeated until it reached 451 ms in duration. The two hybrid stimuli were normalized for overall rms intensity, then shortened by deleting glottal pulses from the vowel "steady states" (i.e., the portions between initial and final transitions). Starting near the beginning of the vowel, three or four glottal periods at a time were deleted. To minimize F0 perturbations, only every other glottal pulse was deleted. Given the extensive shortening that was intended, a second pass that again started near the beginning of the vowel was needed in order to create the three shortest stimuli. The periodic portion of members of the beat–bead continuum that were created in this way ranged from 140–461 ms. The bat–bad stimuli ranged in duration from 133–451 ms. The step size between successive stimuli in the two continua averaged 20 ms (range: 17 to 23 ms).

The two CVC continua were presented in counterbalanced order using a PDP 11/73 computer. The stimuli were low-pass filtered at 4.0 kHz and presented binaurally via headphones at 75 dB SPL(A). The subjects were instructed to push a button marked "t" or "d," and to guess if unsure. The interval between each response and the next stimulus was 1.0 s. The first of the nine random presentations of each stimulus was not analyzed. The dependent variable was the percentage of times each stimulus was identified as ending in /t/.

B. Results

1. Vowel duration

The mean percentages of /t/ responses given to stimuli in the beat–bead and bat–bad continua are shown in Fig. 2. The identification functions in Fig. 2 are similar to those obtained by Crowther and Mann (1992), especially the bat–bad functions. The identification functions of the native English subjects and child learners were similar. For both

FIG. 2. (a) The mean percentage of /t/ responses by native English (NE) speakers and child learners (TC) in response to stimuli in a beat–bead continuum; (b) responses to a bat–bad continuum by the subjects in groups NE and TC; (c) responses by inexperienced Mandarin late learners (MA), inexperienced Taiwanese late learners (TA), and experienced Taiwanese late learners (TB) to the beat–bead continuum; (d) responses to the bat–bad continuum by the subjects in groups MA, TA, and TB.
groups, slopes were somewhat steeper for bat–bad than beat–bead. The functions obtained for the three groups of late learners were more gradual, which agrees with the results obtained by Elsendoorn (1984). Some late learners showed a strong bias toward one response category. Others, as in the Elsendoorn (1984) study, had nonmonotonic functions. As expected (e.g., Crowther and Mann, 1992), still others reversed the labels used by native speakers (see below). Therefore, the effect of vowel duration on the perception of word-final stops was assessed by subtracting the percentage of /t/ responses given to the three stimuli with the longest vowel durations (viz., stimuli 15 to 17) from the percentage of /t/ responses given to the three stimuli with the shortest vowel durations (viz., stimuli 1 to 3).

As shown in Fig. 3, the native speakers’ difference scores were slightly larger for the beat–bead than bat–bad continuum, whereas the reverse held true for all four non-native groups. A two-way ANOVA examining the difference scores yielded a significant main effect of the group factor \( F(4,44) = 3.00, p < 0.05 \) and the continuum factor \( F(1,44) = 15.5, p < 0.05 \), but the two-way interaction missed reaching significance \( F(4,44) = 2.13, p = 0.093 \). The simple effect of group was nevertheless tested because the effect of vowel duration was expected to differ for the two continua owing to F1 offset frequency differences (see Sec. II A). The group factor was significant for beat–bead \( F(4,44) = 3.49, p < 0.05 \). Tukey’s HSD test revealed that the native speakers’ effect was significantly greater than that obtained for the inexperienced Mandarin late learners (83% vs 23%). The native speakers did not differ significantly, however, from the child learners (72%), the experienced Taiwanese late learners (72%), or the inexperienced Taiwanese late learners (38%). For the bat–bad continuum, the difference scores averaged 81% for the native speakers and 83% for the child learners. Differences scores for the experienced and inexperienced Taiwanese late learners (66%, 53%), and for inexperienced Mandarin late learners (58%), were somewhat smaller, but the effect of group did not reach significance \( F(4,44) = 1.77, p > 0.10 \).

All native speakers and child learners identified stimuli with long vowels as ending in /d/, and stimuli with short vowels as ending in /t/. Of the 60 data sets obtained for late learners, this pattern of labeling was reversed six times for beat–bead and twice for bat–bad. One possible explanation for such reversals is that certain late learners associated long vowels with /t/, and short vowels with /d/. However, if this were so, one would not have expected five of the six late learners who reversed labels for beat–bead to show the expected pattern for bat–bad. Perhaps some late learners partitioned the CVC continua by making auditory rather than phonetic use of vowel duration. If so, then the decision to assign long vowels to one button and short vowels to the other button would have been arbitrary, and reversals would be expected. In the absence of compelling evidence in favor of either explanation, a second ANOVA in which the reversals were “corrected” was carried out. The data for native speakers and child learners were, of course, unchanged. However, as shown in Fig. 3, the late learners’ difference scores were somewhat higher than they were before the reversals were corrected. Also, the late learners showed a somewhat smaller difference between the two continua than before the corrections. The simple effect of group was significant for beat–bead \( F(4,44) = 3.23, p < 0.05 \) but not for bat–bad \( F(4,44) = 1.79, p > 0.10 \). Post-hoc tests revealed that the native speakers’ difference scores for beat–bead (viz., 83%) were significantly greater than the inexperienced Mandarin and Taiwanese late learners’ scores (47%, 45%). The native speakers’ scores did not differ significantly from the child learners’ or the experienced Taiwanese late learners’ (72%, 61%).

![FIG. 3. The mean effect of the vowel duration manipulation on stop voicing judgments, estimated as the difference in the average percentage of /t/ responses given to the three stimuli with the shortest vowel durations (stimuli 1–3) and the three stimuli with the longest vowel durations (stimuli 15–17). Differences scores for CVC continua are shown for inexperienced Mandarin late learners (MA), inexperienced Taiwanese late learners (TA), experienced Taiwanese late learners (TB), childhood L2 learners (TC), and native English speakers (NE). Labeling reversals by a few late learners have been corrected in (b) but not in (a), as discussed in the text. The brackets enclose ± 1 standard error.](image-url)
2. Boundary analyses

Curve-fitting techniques were used to determine which of the 98 identification functions were suitable for phoneme boundary analysis. The beat–bead functions of four inexperienced Taiwanese late learners and one experienced Taiwanese late learner did not show a systematic effect of vowel duration (i.e., the slopes obtained did not differ significantly from zero). The same was true for four non-native speakers' bat–bad functions. Of the remaining data sets, 23 were best fit by a linear function, two by an exponential function, and 64 by a logistic function (30 for beat–bead, 34 for bat–bad). The data sets best fit by a logistic function generally showed two plateaus near the ends of the continuum separated by a region of rapid change. Phoneme boundaries and slopes were calculated for these data sets using probit analysis.

The boundaries in beat–bead occurred at somewhat greater stimulus values for the native speakers and child learners than for the three groups of late learners, whose mean boundaries ranged from 7.2 to 8.3. Despite this, the group effect was nonsignificant in a one-way ANOVA \( F(4,25) = 1.59, p > 0.10 \). The mean boundaries obtained for the native speakers and child learners for bat–bad (9.0, 8.6) also occurred at somewhat greater values than did the boundaries obtained for the three groups of late learners (6.7 to 8.3). Once again, the effect of group was nonsignificant \( F(4,30) = 2.52, p = 0.057 \). The nonsignificance of the group factor might be attributed to the small number of values. When the data for the native speakers and child learners were pooled in one group, and those of all late learners were pooled in another group, the effect of age of learning (childhood versus adulthood) was significant for both beat–bead \( F(1,28) = 5.84, p < 0.05 \) and for bat–bad \( F(1,33) = 6.35, p < 0.05 \). In both instances, phoneme boundaries occurred at greater stimulus values for the native speakers and child learners than for the late learners (9.3 vs 7.5 for beat–bead, 8.8 vs 7.5 for bat–bad).

3. Slope and consistency analyses

Slopes of identification functions are used in developmental speech perception research as an index of how well mental speech perception research as an index of how well expected phoneme boundaries and slopes were calculated for these data sets using probit analysis.

Slope values were available for all 49 subjects in the second slope analysis. These values were obtained by fitting a straight line to the responses given to stimulus 6 to stimulus 12. The slope of the best fitting lines were examined in a (5) group \( \times (2) \) continuum ANOVA. The slopes for beat–bead were significantly shallower than those for bat–bad (0.468 vs 0.828) \( F(1,44) = 9.08, p < 0.01 \), suggesting that the manipulation of vowel duration was more effective when \( F1 \) offset frequency provided a clear cue to voicelessness. The two-way interaction missed reaching significance \( F(4,44) = 2.21, p = 0.080 \). However, given an a priori expectation that the effect of vowel duration would differ for the two continua, the simple main effect of group was tested for both continuum. The group factor proved to be nonsignificant for beat–bead \( F(4,44) = 1.25, p > 0.10 \) but significant for bat–bad \( F(4,44) = 4.18, p < 0.01 \). This was opposite to what was found in analyses examining the difference in percentage of /t/ responses given to stimuli with the longest and shortest vowel durations. Tukey's HSD test revealed that the native speakers' and child learners' slopes (0.861, 0.885) were significantly steeper than those of the experienced and inexperienced Taiwanese late learners (0.532, 0.579) and the inexperienced Mandarin late learners (0.307) \( p < 0.05 \).

The consistency with which the subjects identified members of the two continua was examined using a measure based on relative entropy (Atteenne, 1959). Consistency scores were obtained only for the nine odd-numbered stimuli, which differed by 40-ms steps. (Scores were computed for only half of the stimuli because stimulus was to be included as a factor in the ANOVA examining the consistency scores.) The consistency scores ranged from 0.0, in instances where a stimulus was labeled /t/ and /d/ four times each, to 1.0, for stimuli labeled /t/ or /d/ in all eight instances. Averaged over groups and stimuli, the scores were higher for bat–bad than beat–bead (0.468, 0.412). The continuum factor reached significance in a (5) group \( \times (2) \) continuum \( \times (9) \) stimulus ANOVA \( F(1,44) = 4.70, p < 0.05 \). The ANOVA yielded a significant group \( \times \) stimulus interaction \( F(32,352) = 1.72, p < 0.05 \), but not a significant three-way interaction \( F(32,352) = 0.88, p > 0.10 \). The two-way interaction was explored by examining bat–bad and beat–bead scores in one-way ANOVAs. As seen in Fig. 4, the consistency scores were higher for the short-duration than for the long-duration stimuli for all groups except the inexperienced Mandarin late learner group. This may have been due to the incompatibility of long vowels and low \( F1 \) offset frequency in the long-duration stimuli (see Sec. II A). The five groups differed more for short- than long-duration stimuli. The simple main effect of group proved to be significant for stimuli 1, 3, and 5 \( F(4,44) = 3.59 \) to 5.45, \( p < 0.05 \). The group factors for stimuli 7 to 17 were nonsignificant \( p > 0.10 \). Tukey's HSD tests revealed that scores obtained for the native English subjects were higher than those obtained for the inexperienced Mandarin late learners for stimuli 1, 3, and 5. The child learners' scores were higher than the inexperienced Mandarin subjects' for stimuli 1 and 5. The native English subjects' scores for stimulus 3 were significantly higher than those obtained for the experienced and inexperienced Taiwanese late learners \( p < 0.05 \).

C. Discussion

As in experiment 1, the childhood L2 learners closely resembled the native speakers. However, the present results for the late learners differed from those obtained in experiment 1. In experiment 1, all three groups of late learners
produced significantly smaller voicing effects than the native speakers. In the present experiment, shortening the duration of vowels in natural-edited bat-bad stimuli decreased the percentage of /t/ judgments given by the subjects in all groups. For beat-bead, the effect size was significantly greater for the native English subjects than for the inexperienced Mandarin subjects, and also for the inexperienced Taiwanese late learners once labeling reversals had been corrected. There was not a significant difference between the native speakers and the experienced Taiwanese late learners, suggesting that the experienced late learners had acquired a nativelike sensitivity to vowel duration. If so, then the present results, when taken together with the production data from experiment 1, could be viewed as providing support for the hypothesis that the perception of vowel duration cues precedes its production.

This conclusion can be questioned on several grounds, however. The lack of between-group differences in experiment 2, upon which the conclusion stands, might be attributed to a lack of statistical power arising from a small number of subjects per group, or to intersubject variability. Also, several findings of experiment 2 suggested that the late learners did not use vowel duration in a truly "nativelike" manner. For bat-bad, slopes were steeper for the native speakers and child learners than for the inexperienced Mandarin late learners. The Mandarin subjects were significantly less consistent in labeling short-duration stimuli than the native speakers and child learners. Experienced and inexperienced Taiwanese late learners were less consistent than the native speakers in identifying one short-duration stimulus. Also, some late learners reversed the labels used by all native speakers. Finally, for identification functions that could be subjected to probit analysis, the late learners' phoneme boundaries occurred at shorter vowel durations than those obtained for the native speakers and child learners.

III. EXPERIMENT 3

In experiment 2, only inexperienced late learners showed a smaller effect of vowel duration in identifying word-final stops than did the native speakers (and then only for one of the two CVC continua). However, the experienced late learners did differ from the native speakers in response consistency, which suggests that they did not use vowel duration in a truly nativelike fashion. Experiment 3

![Graph](image1)

**FIG. 4.** The average consistency with which odd-numbered members of two vowel duration continua (beat to bead, bat to bad) were labeled by native speakers of English and childhood L2 learners (left) and three groups of late learners (right).

![Graph](image2)

**FIG. 5.** The mean vowel difference in stimuli chosen as the best example of words ending in /t/ (beat, bat) and /d/ (bead, bad) for inexperienced Mandarin late learners (MA), inexperienced Taiwanese late learners (TA), experienced Taiwanese late learners (TB), childhood L2 learners (TC), and native English speakers (NE). The brackets enclose ± 1 standard error.
assessed further the non-natives’ perceptual use of vowel duration as a cue to the word-final English /t/-/d/ contrast. It made use of the method of adjustment [see Nooteboom (1973)]. Elsendoorn (1984) used this technique to examine the range of acceptable vowel durations preceding /t/ and /d/ in English words for native speakers of English and also Dutch speakers who had learned English as an L2. The size of voicing effects produced by the non-natives was found to be correlated with the size of vowel duration differences between the subjects’ preferred /t/ and /d/-final words. The correlations between produced and perceived (i.e., adjusted) vowel durations were greater for relatively experienced than inexperienced Dutch subjects, suggesting that, as in L1 acquisition, alignment of production and perception may occur gradually in adult L2 learning (see Flege, 1988a).

In experiment 1 of the present study, late learners produced significantly smaller voicing effects than did native speakers. Given the correlations obtained by Elsendoorn, the possibility existed that the method of adjustment would reveal significant perceptual differences between the native speakers and the experienced late learners that were not evident in experiment 2. Such a finding, if obtained, would tend to undermine the conclusion that the perception of vowel duration precedes the production of an English-like voicing effect.

A. Method

The 49 subjects participated in experiment 3 of this study at the end of the second day of testing. They were instructed to choose one of the 17 beat-bead stimuli as the “best example” of the word beat. In a separate session, they were to choose the best example of bead from the same 17-member array. The same procedure was followed for the bat-bad stimuli. The orders of the target words and continua were counterbalanced across subjects. At the beginning of each session, a card bearing the target word was placed on the response box. The stimuli were played out in order from shortest to longest before each session. The stimuli were presented at the end of the same carrier phrase used in earlier experiments (I will say _). The subjects were told to push button #1 if they wanted to hear a word with a shorter vowel. They pushed button #2 to hear a longer vowel, button #3 to have the last stimulus repeated, and button #4 to choose the last stimulus presented as the “best example” of the target word being examined in the session. This process was repeated 17 times in each session, with each member of the continuum serving (randomly) as the starting point for one trial. The first two of 17 trials in each session were not analyzed. The four mean values obtained for each subject (one each for the best examples of bat, bad, beat and bead) were thus based on 15 choices.

B. Results

Frequency histograms were plotted to show how often each stimulus was chosen as the best example of a word ending in /t/ or /d/. The native speakers and child learners showed two distinct distributions for words judged to have the best sounding /t/’s and /d/’s. The distributions of the inexperienced Mandarin and Taiwanese late learners’ preferred /t/’s and /d/’s overlapped considerably. Those of experienced late learners overlapped somewhat less. To test for between-group differences, two values were derived for each subject: the average difference between choices for beat and bead (expressed in number of stimuli), and the average difference between preferred bat and bad stimuli. The mean difference scores are shown in Fig. 5. Scores were larger for the native English subjects and the child L2 learners than for the experienced Taiwanese late learners, whose scores were larger than those of the inexperienced Mandarin and Taiwanese late learners. Recall that, in experiment 2, all four non-native groups showed slightly larger effects of vowel duration on the identification of stops in the bat–bad than beat–bead continuum. In the present experiment, the child learners and experienced Taiwanese late learners resembled the native speakers in showing a somewhat larger effect of vowel duration for beat–bead than bat–bad. The inexperienced Mandarin and Taiwanese late learners, on the other hand, showed the opposite pattern, and thus maintained the pattern seen in experiment 2.

Despite these differences, the group x continuum interaction did not reach significance in a two-way ANOVA [F(4,44) = 0.54, p > 0.10]. The main effect of continuum was nonsignificant [F(1,44) = 0.02, p > 0.10], but the group main effect was significant [F(4,44) = 8.42, p < 0.01]. A post-hoc test revealed that average difference scores were significantly greater for the native speakers and child learners (5.4, 5.7) than for inexperienced Mandarin and Taiwanese late learners (1.7, 0.8) (p < 0.05). The native speakers and child learners did not differ significantly, however, from the experienced Taiwanese subjects (3.51). The differences between the experienced and inexperienced Taiwanese late learners, and between the inexperienced Taiwanese and Mandarin late learners, were also nonsignificant (p > 0.05).

C. Discussion

The native speakers of English and the child learners specified greater vowel duration differences between words ending in /t/ vs /d/ than did the inexperienced Taiwanese and Mandarin late learners, but they did not differ significantly from the experienced Taiwanese late learners. Recall that the experienced and inexperienced Taiwanese late learners, and the Mandarin late learners, produced significantly smaller voicing effects than the native speakers in experiment 1. The data obtained here for the inexperienced Taiwanese and Mandarin late learners therefore neither support nor disconfirm the hypothesis that perception precedes production. The results for the experienced Taiwanese late learners, on the other hand, support this hypothesis when taken together with the experiment 1 results. Of course, the lack of a significant perceptual differences between the native speakers and experienced late learners could be attributed to a lack of statistical power. An imitation experiment was therefore carried out. It sought to resolve whether the experienced late learners did or did not make the same perceptual use of vowel duration as the native speakers.
IV. EXPERIMENT 4

It is generally accepted that aspects of segmental production and perception are linked. For example, language-specific voice onset time (VOT) values in production and perception coincide in mature speakers (Flege and Eefting, 1986b). The linkage of production and perception values seems to take place over several years in early childhood (Bailey and Haggard, 1980; Zlatin and Koenigsknecht, 1975, 1976). It is an essential characteristic of phonetic representations, for it is through the linkage that sensory properties of target-language sounds are related to the learner’s vocal output (Fry, 1966). A speech imitation task affords a unique opportunity to assess the relation between sensory and motor processes. The structural properties of stimuli being imitated are perceived, coded, and stored in memory, then retrieved and regenerated in a form suitable to guide skilled movements.

Categorization may take place during the coding storage stage of imitation. Chistovich et al. (1966) suggested that vowel stimuli may not be categorized, but there is evidence that categorization occurs in the imitation of consonants. Flege and Eefting (1987, 1988) had native speakers of English and Spanish identify, then later imitate the members of a/da/to/ta/continuum. VOT values produced by the subjects were correlated with VOT values in the stimuli. However, the difference in VOT values produced in the imitation of pairs of stimuli straddling the /t/-/d/-phoneme boundary obtained in the identification experiment was greater than for stimulus pairs located on either side of the boundary. Stimulus-response “discontinuities” occurred at greater VOT values for the native English than Spanish subjects, suggesting that the stimuli were being categorized before or during the imitations. Similar results were obtained by Lehiste and Shockey (1980), whose native English subjects identified, then imitated the members of a synthetic beat-bead continuum. It appeared that discontinuities in imitated vowel durations resulted from the subjects’ having covertly categorized the final stop as /t/ or /d/ based on variations in preceding vowel duration.

The present experiment assumed that non-native speakers would show discontinuities in imitated vowel durations only if they covertly categorized word-final stops in the CVC stimuli as /t/ or /d/. The subjects were asked to imitate the duration of vowels in members of the beat-bead and bat-bad continua as accurately as possible. They also imitated members of an isolated vowel continuum that was derived from the beat-bead continuum. If late learners who reversed labels in the identification experiment did so because they used vowel duration acoustically (or auditorially) rather than phonetically, one would not expect discontinuities near their experiment 2 “phoneme” boundaries. This was because the short interval available for imitation was likely to discourage an auditory use of vowel duration. If native but not non-native subjects categorized final stops in beat-bead and bat-bad stimuli, one might expect the native speakers to be less accurate in imitating vowel duration than the non-native speakers because of categorization induced inaccuracy. Vowel duration is apparently not used as a cue to segmental phonetic distinctions in Mandarin or Taiwanese (Leather, 1988; Lin and Repp, 1989), so the native speakers might be more accurate than the non-natives, on the other hand, in imitating isolated vowels because of their experience using vowel duration as a cue to segmental phonetic distinctions.

A. Method

The native and non-native subjects imitated the isolated vowels on the first day of testing, the two CVC continua (in counterbalanced order) on the second day of testing day. The isolated vowel stimuli were derived from members of the beat-bead continuum by editing out the release bursts and transitions of the initial and final consonants, leaving just the “steady-state” vocalic portion. Neither an initial nor a final stop was perceptible after the editing. The isolated vowel stimuli averaged 59.2 ms (s.d. = 0.5) shorter than the periodic portion of the beat-bead stimuli, ranging in duration from 81 to 403 ms with an average step size of 20 ms (s.d. = 3).

The stimuli to be imitated were presented as described earlier. Each stimulus was randomly presented six times following a 0.6-s carrier phrase (/will say_). The subjects were told to repeat the whole utterance, including the carrier phrase, taking care to imitate the duration of the vowel in the variable V or CVC stimuli “as accurately as possible.” The interval between the end of each isolated vowel stimulus and onset of the carrier phrase was fixed at 3.0 s. The intertrial interval was a constant 2.0 s for the CVC conditions. This left 1.6 to 1.9 s for imitating the utterances made up of the carrier phrase + isolated vowel stimuli, and 1.8 to 2.2 s for utterances with CVC stimuli. The duration of vowels in the final five of six repetitions of each V and CVC stimulus was measured using procedures described in experiment 1. This yielded a total of 12 495 measurements (49 subjects × 3 continua ×17 stimuli ×5 repetitions). As expected (e.g., Lehiste and Shockey, 1980), there was a great deal of inter- and intrasubject variability, so the medians of the five imitations by each subject of each stimulus were used in all analyses.

B. Results

The mean durations of the subjects’ imitations of isolated vowels are shown in Fig. 6. Means plotted above the diagonal indicate “overshoot,” those below the diagonal indicate “undershoot” of stimulus durations. The results for native speakers and child learners are directly compared in Fig. 6(a). There was little systematic difference between the subjects in these two groups. Both the native speakers and child learners produced vowels that were roughly 100 ms longer than the stimulus vowels. The two groups compared in the remaining panels were: native English versus experienced Taiwanese late learners [Fig. 6(b)], experienced versus inexperienced Taiwanese late learners [Fig. 6(c)], inexperienced Taiwanese vs Mandarin late learners [Fig. 6(d)]. The experienced Taiwanese late learners and the inexperienced Mandarin late learners tended to overshoot stimulus durations to a greater extent for short than long-duration stimuli, as if they had divided the continuum of isolated vowels into two portions.

Figure 7(a)-(d) shows the mean durations of imitations of the beat-bead stimuli. The native speakers and child
Imitation of Isolated Vowels

FIG. 6. Mean duration (in ms) of vowels produced in imitation of a 17-member isolated vowel continuum (L = late; LS = learners). The four panels show the imitations of: (a) native English (NE) speakers and childhood L2 learners (TC); (b) NE speakers and experienced Taiwanese late learners (TB); (c) the subjects in TB and inexperienced Taiwanese late learners (TA); the subjects in TA and inexperienced Mandarin late learners (MA). The means for all groups except TC are based on 50 imitations.

Learners showed discontinuities not evident in their imitations of isolated vowels. The discontinuity in the stimulus-response function of the native speakers occurred between stimulus 7 and stimulus 8 [Fig. 7(a)]. That is, their imitations of stimuli 7 and 8 differed more than the average of the differences between all other pairs of adjacent stimuli (58 vs 14 ms). The child learners' discontinuity occurred at slightly longer stimulus durations, that is, between stimulus 8 and stimulus 9 [see Fig. 7(b)]. The differing locations of these discontinuities correspond to differences in identifications of the beat-bead stimuli by the native speakers and child learners in experiment 2. This strongly suggests, then, that the discontinuities arose from covert categorization of final stops in the beat-bead stimuli. A smaller discontinuity was evident for the experienced Taiwanese late learners [Fig. 7(b)]. Their discontinuity occurred at the same location as the one observed for the child learners. There was no evidence of a discontinuity for the two inexperienced late learner groups [Fig. 7(c) and (d)].

Stimulus-response functions for bat-bad stimuli are shown in Fig. 8(a)–(d). The native speakers and child learners had discontinuities at the same location, viz., between stimulus 8 and stimulus 9. This was expected because the two groups' identification functions in experiment 2 were very similar for bat-bad. A small discontinuity was evident for the experienced Taiwanese late learners [Fig. 8(b)], but it spanned two stimuli (viz., 7 to 9) rather than just one stimulus. Once again, there was no evidence of a discontinuity for the inexperienced Taiwanese or Mandarin late learners [Fig. 8(c) and (d)].

The imitation data were analyzed in two ways. Pearson product-moment correlation coefficients between stimulus and response vowel durations were calculated for each subject's imitations of the isolated vowels, beat-bead, and bat-bad stimuli. The r values were examined in three separate one-way ANOVAs. The presence of a categorization-induced discontinuity in the CVC stimuli might decrease r by causing imitation durations to differ from stimulus values near the phoneme boundary. If so, one might expect lower r values for the native speakers than for the inexperienced Taiwanese and Mandarin late learners (and, possibly, the experienced Taiwanese late learners). For isolated vowels, on the other hand, larger r's might be expected for the native speakers than for the late learners, at least if the native speakers' previous experience controlling vowel duration as a cue to segmental phonetic distinctions in their L1 would prove helpful in the imitation task.

ANOVAs examining the r values failed to confirm these
FIG. 7. Mean duration (in ms) of vowels produced in imitation of a 17-member beat-bead continuum (see the Fig. 6 legend).

expectations. The effect of group was nonsignificant for the isolated vowels \( F(4,44) = 1.36, p > 0.10 \), for beat-bead \( F(4,44) = 0.65, p > 0.10 \), and for bat-bad \( F(4,44) = 2.46, p = 0.060 \). The lack of between-group differences may have been a ceiling effect, for all stimulus-response correlations were high. The mean \( r \) values for the isolated vowels ranged from 0.931–0.971; for beat-bead they ranged from 0.930–0.951; and for bat-bad they ranged from 0.918–0.962. Significant between-group differences also failed to emerge when correlation coefficients were calculated for several increasingly small ranges of stimuli centered at the middle of the continua.

Other ANOVAs examined "discontinuity" scores based on the imitations of stimuli 4 to 14. The largest of the differences between all pairs of stimuli in this range that differed by 80 ms (e.g., stimulus 4 vs 8, stimulus 9 vs 13) were determined for each data set. (Eighty milliseconds was determined to be the smallest step size that was sufficiently large to straddle all subjects' phoneme boundaries.) As expected, all of the mean discontinuity scores in Fig. 9 are larger than 80 ms (the value that would have been obtained had all stimuli been imitated with perfect accuracy). The child learners closely resembled the native speakers of English. For isolated vowels, smaller discontinuity scores were obtained for the native speakers (130 ms) than for the three groups of late learners (101 to 112 ms). For bat-bad, larger scores were also obtained for the native speakers (135 ms) than for the late learners (110 to 131 ms).

A significant interaction \( F(8,88) = 3.51, p < 0.01 \) was obtained in a (5) group \( \times (3) \) continuum ANOVA that examined the discontinuity scores. The simple main effect of continuum was significant for all three groups of late learners \( F \) values ranging from 5.56–19.69, \( p < 0.01 \). Post-hoc tests revealed that scores obtained for the inexperienced Mandarin and Taiwanese late learners for beat-bead and bat-bad were significantly smaller than their isolated vowel continuum scores. For the experienced Taiwanese late learners, beat-bead but not bat-bad scores were significantly smaller than scores for the isolated vowel continuum. The between-continuum differences obtained for the experienced Taiwanese late learners and the inexperienced Mandarin late learners might be attributed to two factors. One is the tendency, noted earlier, for them to produce isolated vowels with relatively short or long durations. The other is the small size (or complete absence) of discontinuities in their imitations of the CVC stimuli. For the inexperienced Taiwanese late learners, the difference between continua can be attributed primarily to the absence of discontinuities in CVC stimuli. The native speakers of English showed nonsignificant differences in a direction that was opposite to the one observed for the late learners \( F(2,88) = 0.62, \)
The aim of the imitation experiment was to determine if the three groups of late L2 learners would show the same evidence of sensory-motor linkages as the native speakers of English. The subjects in all five groups showed high stimulus-response correlations. However, only the native speakers and the childhood L2 learners showed clear discontinuities in their imitations of CVC stimuli that could be attributed to categorization of final stops as /t/ or /d/. The results obtained for the experienced Taiwanese late learners resembled the results obtained for the inexperienced Taiwanese and Mandarin late learners rather than those of the native speakers and child learners in that they were more

\[ p > 0.10 \], as did the child learners \[ F(2,88) = 0.42, p > 0.10 \].

The simple main effect of group was tested for all three continua. It was significant for the isolated vowel continuum \[ F(4,123) = 5.41, p < 0.05 \] but nonsignificant for \textit{beat–bead} \[ F(4,123) = 5.41, p < 0.05 \] and \textit{bat–bad} \[ F(4,123) = 5.41, p < 0.05 \]. \textit{Post-hoc} tests revealed that, for the isolated vowels, discontinuity scores obtained for the native speakers and child learners were significantly smaller than those obtained for the inexperienced Mandarin late learners (MA) \( p < 0.05 \).

C. Discussion

The aim of the imitation experiment was to determine if the three groups of late L2 learners would show the same evidence of sensory-motor linkages as the native speakers of English. The subjects in all five groups showed high stimulus–response correlations. However, only the native speakers and the childhood L2 learners showed clear discontinuities in their imitations of CVC stimuli that could be attributed to categorization of final stops as /t/ or /d/. The results obtained for the experienced Taiwanese late learners resembled the results obtained for the inexperienced Taiwanese and Mandarin late learners rather than those of the native speakers and child learners in that they were more
TABLE III. Mean effect sizes obtained in four experiments from inexperienced Mandarin late learners (MA), inexperienced Taiwanese late learners (TA), experienced Taiwanese late learners (TB), Taiwanese speakers who learned English as children (TC), and native English speakers (NE).

<table>
<thead>
<tr>
<th></th>
<th>NE</th>
<th>TC</th>
<th>TB</th>
<th>MA</th>
<th>TA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>beat vs bead continuum</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp. 1: production (ms)</td>
<td>145</td>
<td>85</td>
<td>49</td>
<td>43</td>
<td>31</td>
</tr>
<tr>
<td>Exp. 2: % change in /t/</td>
<td>83</td>
<td>72</td>
<td>43</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Exp. 3: difference in number of stimuli preferred difference in ms</td>
<td>5.54</td>
<td>6.03</td>
<td>3.67</td>
<td>1.44</td>
<td>0.47</td>
</tr>
<tr>
<td>Exp. 4: discontinuity scores (ms)</td>
<td>111</td>
<td>121</td>
<td>73</td>
<td>29</td>
<td>9</td>
</tr>
<tr>
<td><strong>bat vs bad continuum</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp. 1: production (ms)</td>
<td>128</td>
<td>127</td>
<td>74</td>
<td>37</td>
<td>42</td>
</tr>
<tr>
<td>Exp. 2: % change in /t/</td>
<td>81</td>
<td>83</td>
<td>66</td>
<td>57</td>
<td>53</td>
</tr>
<tr>
<td>Exp. 3: difference in number of stimuli preferred difference in ms</td>
<td>5.17</td>
<td>5.43</td>
<td>3.34</td>
<td>1.87</td>
<td>1.12</td>
</tr>
<tr>
<td>Exp. 4: discontinuity scores (ms)</td>
<td>103</td>
<td>109</td>
<td>67</td>
<td>37</td>
<td>22</td>
</tr>
</tbody>
</table>

accurate in imitating CVC than isolated V stimuli. This suggests that the experienced late learners probably did not establish nativelike phonetic representations for the contrast between word-final English /t/ and /d/.

V. RELATION BETWEEN PRODUCTION AND PERCEPTION

As summarized in Table III, the magnitude of effects obtained in experiments 1 to 4 generally increased in the following order: native speakers of English > child learners > experienced Taiwanese late learners > inexperienced Mandarin late learners > inexperienced Taiwanese late learners. (The few exceptions to this ordering involved mostly small differences in mean effect sizes.) The experienced Taiwanese late learners showed a larger, and thus more English-like, effect than the inexperienced Taiwanese late learners in every instance. Thus, although the experienced and inexperienced Taiwanese late learners never differed statistically, the consistent difference in their rankings suggests that additional L2 experience nevertheless did enable the relatively experienced late learners to approximate English phonetic norms to a greater extent than the relatively inexperienced late learners. The data in Table III provide no support, however, for two other hypotheses. There was no systematic difference between the native speakers of English and the child learners; nor were there systematic differences between the inexperienced Taiwanese and Mandarin late learners.

Correlation analyses were carried out to assess the relation between the production and perception data obtained for the 39 non-native subjects, and between data obtained here for the non-native subjects and foreign accent scores obtained previously for all but one child learner by Flege (1988b). The Pearson product-moment correlation coefficients obtained in these analyses are presented in Table IV.

TABLE IV. Correlation between variables obtained in experiments 1 to 4 and foreign accent scores from Flege (1988b). Values in boldface are significant at the 0.01 level.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>foreign accent</td>
<td>-0.044</td>
<td>0.451</td>
<td>-0.039</td>
<td>0.555</td>
<td>0.714</td>
<td>0.714</td>
<td>0.305</td>
<td>0.344</td>
<td>0.509</td>
<td>0.424</td>
<td>-0.491</td>
<td>0.136</td>
</tr>
<tr>
<td>beat duration</td>
<td>0.734</td>
<td>0.412</td>
<td>0.353</td>
<td>-0.107</td>
<td>0.131</td>
<td>0.242</td>
<td>0.156</td>
<td>0.069</td>
<td>0.067</td>
<td>0.001</td>
<td>0.216</td>
<td>0.331</td>
</tr>
<tr>
<td>bead duration</td>
<td>0.296</td>
<td>0.539</td>
<td>0.506</td>
<td>0.059</td>
<td>0.455</td>
<td>-0.014</td>
<td>0.011</td>
<td>0.239</td>
<td>0.199</td>
<td>-0.231</td>
<td>0.016</td>
<td>0.311</td>
</tr>
<tr>
<td>bad duration</td>
<td>0.611</td>
<td>0.053</td>
<td>0.017</td>
<td>0.124</td>
<td>0.197</td>
<td>0.124</td>
<td>0.239</td>
<td>0.179</td>
<td>0.074</td>
<td>-0.179</td>
<td>0.041</td>
<td>0.156</td>
</tr>
<tr>
<td>beat vs bead duration</td>
<td>0.372</td>
<td>0.795</td>
<td>0.073</td>
<td>0.288</td>
<td>0.459</td>
<td>0.358</td>
<td>-0.256</td>
<td>0.229</td>
<td>0.301</td>
<td>0.229</td>
<td>0.301</td>
<td>0.301</td>
</tr>
<tr>
<td>bat vs bad duration</td>
<td>0.511</td>
<td>0.227</td>
<td>0.200</td>
<td>0.268</td>
<td>0.094</td>
<td>-0.304</td>
<td>0.001</td>
<td>0.184</td>
<td>0.184</td>
<td>0.184</td>
<td>0.184</td>
<td>0.184</td>
</tr>
<tr>
<td>bat vs head identification*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.200</td>
<td>0.269</td>
<td>0.444</td>
<td>0.445</td>
<td>0.355</td>
<td>0.323</td>
<td>0.206</td>
<td></td>
</tr>
<tr>
<td>bat vs head identification*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.627</td>
<td>0.529</td>
<td>0.346</td>
<td>0.119</td>
<td>0.225</td>
<td>0.444</td>
<td>0.444</td>
<td></td>
</tr>
<tr>
<td>bat vs head adjustment*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.496</td>
<td>0.494</td>
<td>0.004</td>
<td>0.190</td>
<td>0.321</td>
<td>0.251</td>
<td>0.251</td>
<td></td>
</tr>
<tr>
<td>bat vs bad adjustment*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.704</td>
<td>-0.102</td>
<td>0.115</td>
<td>0.269</td>
<td>0.184</td>
<td>0.184</td>
<td>0.184</td>
<td></td>
</tr>
<tr>
<td>isolated vow. imitation*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.179</td>
<td>0.246</td>
<td>0.003</td>
<td>0.258</td>
<td>0.503</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>beat vs head imitation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.003</td>
<td>0.258</td>
<td>0.503</td>
<td>0.503</td>
<td>0.503</td>
<td>0.503</td>
<td>0.503</td>
<td></td>
</tr>
</tbody>
</table>

*a Vowel duration effect from the identification experiment (exp. 2).
* Differences in preferred /t/ vs /d/ final stimuli obtained using the method of adjustment (exp. 3).
* Discontinuity scores from the imitation experiment (exp. 3).
Two interesting correlations were those between the magnitude of voicing effects produced in beat-bead \( r = 0.714, p < 0.01 \) and bat-bad \( r = 0.714, p < 0.01 \) and the foreign accent scores. The larger—and thus more nativelike—were the voicing effects, the less accented were the non-natives' productions of three English sentences. This agrees with several other studies showing a relation between global foreign accent and segmental articulation (Ryan et al., 1975; Flege and Eefting, 1986a; Major, 1987). The correlations between degree of foreign accent and the duration of vowels in words ending in /d/ were stronger than those between accent and vowel durations in /t/-final words. This was expected because the non-natives' vowels differed from the native speakers' in the context of /d/ but not /t/.

A significant correlation existed between discontinuity scores obtained in the imitation experiment and degree of foreign accent. The larger the discontinuity scores for isolated vowels (see above), the more accented the non-natives' sentences were judged to be \( r = -0.491, p < 0.01 \). It was suggested earlier that non-native subjects might be less accurate in imitating isolated vowels than the native English subjects because of a lack of experience using vowel duration as a segmental phonetic cue. Inspection of Fig. 6 suggested that the inexperienced Mandarin late learners and the experienced Taiwanese late learners partitioned the isolated vowel continuum into "short" and "long" portions. It is known that listeners can detect very small divergences from the temporal norms for the vowels of their native language (e.g., Nooteboom, 1973), and that an incorrect specification of vowel duration may contribute to the perception of foreign accent (e.g., Elsendoorn, 1984). The correlations just reported may therefore mean that degree of accent in an L2 is related to motoric differences in ability to specify vowel duration in the L2.

The foreign accent scores were also correlated with perceptual data obtained using the method of adjustment. The larger the difference in vowel duration for preferred beat and bead stimuli \( r = 0.509, p < 0.01 \), and the larger the difference between bat and bad stimuli \( r = 0.424, p < 0.01 \), the better the non-natives' pronunciation of English sentences. This establishes directly the link between production and perception discussed above. Other correlations did so as well. The larger the vowel duration difference between bat and bad produced by the non-natives, the larger was their preferred difference in vowel durations for bat and bad in experiment 3 \( r = 0.445, p < 0.01 \). A similar correlation was obtained for production of beat-bead \( r = 0.444, p < 0.01 \).

Figure 10 shows vowel duration differences produced in experiment 1 (x axes) and the perceptual effects of vowel duration obtained in experiment 3 using the method of adjustment (y axes). The perceptual difference scores, which were originally expressed in terms of the number of stimuli between preferred /t/ and /d/-final words, have been multiplied by the average step size between stimuli (20 ms) to make these data compatible with the production data. All but three non-natives produced longer vowels in bead than in beat. All but three made vowels longer in bad than bat. Less consistency was evident in the perceptual data, however. Eight non-natives chose stimuli with longer vowels when selecting their preferred beat stimulus than when selecting their preferred bead stimulus. Nine subjects showed such a reversal for bat-bad.

The diagonal line represents equal vowel duration effects in production and perception. The perception effect was larger than the corresponding production effect for points above the diagonal, whereas the opposite held true for points below the diagonal. Nearly the same number of points fell above and below the diagonal (34 vs 39), suggesting a lack of directionality in the relation between production and perception.
perception. A consideration of the number of individuals who showed disproportionate production and perception effects leads to the same conclusion. The number of subjects who showed a 50-ms or greater effect in the perception experiment, but who produced a vowel duration difference of 20 ms or less, was tabulated. These two conditions were met in 11 instances. Also tabulated was the number of subjects who showed a disparity in the opposite direction (i.e., subjects who produced vowel duration differences of at least 50 ms, but showed perception effects of 20 ms or less). These conditions were also met in 11 instances.

VI. GENERAL DISCUSSION
A. Production-perception relation

This study assessed the extent to which 39 non-native subjects from mainland China and Taiwan made vowels longer before /d/ than /t/ in a word-list reading experiment (experiment 1), and used vowel duration as a perceptual cue to the / d/ vs / t/ distinction in perception (experiments 2 and 3). Correlation analyses pointed to a link between production and perception. The non-natives' degree of foreign accent in English sentences was found to be correlated with the method of adjustment results obtained in experiment 3. The larger the difference in vowel duration between preferred examples of beat and bead, and bat and bad, the less accented the non-natives' production of English sentences were judged to be. Also, significant correlations were found to exist between the size of voicing effects observed in the production of English minimal pairs and the size of the difference in vowel durations for stimuli that were chosen as the preferred instances of beat versus bead, and of bat versus bad.

The pattern of significant and nonsignificant between-group differences obtained here is consistent with the hypothesis that non-natives will resemble native speakers more closely in perceiving than producing vowel duration differences. The native speakers produced significantly larger voicing effects in the minimal pair beat–bead than the child learners and the inexperienced Taiwanese late learners. The native speakers also produced a significantly larger voicing effect in bat–bad than the experienced Taiwanese late learners. However, the child learners and experienced Taiwanese late learners did not differ significantly from the native speakers in perceiving vowel duration cues to the / t/–/ d/ distinction in these two minimal pairs.

Data for individual subjects, on the other hand, were not consistent with the "perception before production" hypothesis. As many late learners showed large production effects in the absence of large perception effects as showed large perceptual effects of vowel duration without producing large vowel duration differences. The individual data obtained in the present study therefore agrees with the results of a recent tone-training experiment. Leather (1991) provided either tone production training or tone perception training to Dutch subjects unfamiliar with Mandarin. The ability of subjects trained to identify Mandarin tones was correlated with their ability to produce Mandarin tones in the absence of prior production training. The spontaneous tone produc-

tion ability of subjects trained to vocally match computer-displayed tones was correlated with their ability to identify tones in the absence of prior perceptual training. Parallel production and perception errors were evident for the subjects in both training groups. Interestingly, the production-perception correlations were about equally strong for both groups. From this, Leather concluded that there is "no simple correspondence" between production and perception, although the two may manifest an "independent association" even in very early stages of L2 learning (1991, p. 140).

Individual data obtained in the present study might be interpreted to mean that success in producing a novel L2 contrast is not, as some claim (e.g., Flege, 1988a, 1992a,b), limited by the extent to which perceptual representations have developed. One might hypothesize, alternatively, that some of the non-native subjects had difficulty with "the timing of ... articulatory gestures" (see Crowther and Mann, 1992, p. 721) and that this—not inadequate perceptual representations—was responsible for the small vowel duration differences they produced. This alternative hypothesis received indirect support from the inverse correlation found to exist between the non-natives’ accuracy in imitating isolated vowels and their degree of foreign accent in English. Subjects who had large discontinuity scores in the vowel imitation task (indicating relatively large divergences from stimulus values) had stronger accents in English than those with relatively small discontinuity scores. Self-hearing might also have played a role. The production of inaccurate vowel durations might have reduced the non-native subjects’ sensitivity to vowel duration (see Ohde and Sharf, 1988, p. 566).

There are, however, reasons for caution in abandoning the claim that L2 production accuracy may be constrained by the development of nativelike perceptual representations for sounds in an L2. The grouped data were in accord with the "perception before production" hypothesis. Gass (1984) noted an inherent difficulty in comparing perception and production data. Perceptual processing is more "automatic" than speech production, making it less subject to variation in attention and socially conditioned stylistic factors. There is no guarantee that the production results obtained here would generalize to other speaking situations, or even to other phonetic contexts. In a study by Elsendoorn (1984), the strength of correlation between production and perception was greater for experienced than inexperienced L2 learners (see the Introduction). Perhaps, as in L1 acquisition (Zlatin and Koenigsknecht, 1975, 1976; Bailey and Haggard, 1980; Flege and Eefting, 1986b), it takes a number of years for L2 perception and production to align optimally. Of the 30 late learners examined in the present study, 16 showed an effect of vowel duration in perception that was opposite in direction to the effects observed for all native speakers and child learners in at least one instance. Perhaps the study would have supported the "perception before production" hypothesis had more experienced late learners been examined.

B. Between-group differences

Several questions were raised in the Introduction concerning differences between groups of non-native speakers.

One concerned whether individuals who began learning English between the ages of 3 and 13 years would differ significantly from native speakers of English in producing and perceiving the word-final English /t/-/d/ contrast. The "child learners" did not differ significantly from the native speakers in perceiving or imitating the /t/-/d/ contrast, yet they produced significantly smaller voicing effects than the native speakers in one of the minimal pairs examined (beat– bead). Moreover, those who learned English after the age of 9 years produced somewhat smaller overall voicing effects than those who began learning English before the age of 9 years. This is consistent with the hypothesis that foreign accents first emerge at an age of L2 learning (AOL) of roughly 7 years (Flege and Fletcher, 1992). It will require further study to pinpoint the AOL at which voicing effects first begin to diverge systematically from those produced by native speakers of English.

A second question was whether groups of Taiwanese late learners who had lived in the U.S. for averages of 5.1 and 1.2 years would differ. The two groups never differed significantly yet, in all measures of production and perception, the experienced group always resembled the native speakers to a greater extent than the inexperienced group. Thus, as in previous research (e.g., Flege et al., 1992), the effect of additional L2 experience for late learners seems to have been a modest movement in the direction of L2 phonetic norms. Another question was whether the inexperienced Taiwanese late learners would differ from inexperienced native Mandarin late learners. Only small, nonsignificant differences between these two groups were observed. It appears that the presence of /p t k/ in the final position of L1 words (Taiwanese), as against the complete absence of word-final obstruents (Mandarin), has little effect on the production and perception of the word-final English /t/-/d/ contrast. It is conceivable, however, that more experienced groups of Mandarin and Taiwanese subjects would be found to differ.

A final question was whether the relatively experienced Taiwanese late learners would differ from the native speakers. This question was of interest in connection with the claim that certain "new" L2 consonants can be mastered by adult learners of an L2 (e.g., Flege, 1992a,b). The experienced Taiwanese late learners produced substantially smaller voicing effects than the native speakers (63 vs 137 ms). Only three (30%) of the experienced late learners produced voicing effects that fell within the native English range. This finding agrees with the results obtained in previous acoustic (see Flege et al., 1992; Crowther and Mann, 1992), physiological (Flege et al., 1987; Flege, 1988c), and auditory perceptual (e.g., Flege and Davidian, 1984; Heyer, 1986; Flege et al., 1992) experiments examining Chinese subjects' production of word-final English stops. Taken together, the data now available indicate that Chinese late learners usually do not produce nativelike vowel duration differences preceding word-final /t/ vs /d/ in English words.

The question that must now be answered is "Why is acquisition of the contrast between voiced and voiceless word-final English stops so difficult for Chinese late learners?" Flege et al. (1992) identified five potential explanations: (1) The /t/-/d/ contrast may be more difficult to learn than /p/-/b/, /k/-/g/ contrasts because of the variability with which /t/ and /d/ are produced in English; (2) late learners may produce stops more accurately when speaking spontaneously than when reading lists of words; (3) the late learners examined in previous research may not have received sufficient native-speaker phonetic input; (4) they may have identified word-final English stops in terms of word-initial L1 stop categories; or (5) they may have passed a "critical period" that rendered mastery of L2 sounds difficult or impossible. To this list might be added a sixth possibility. When late learners establish a new category for sounds in an L2, they might make use only of features that are phonetically important in the L1. This might explain why, for example, Spanish and Finnish learners of English produce seemingly nativelike contrasts in closure voicing between /p t k/ and /b d g/ yet, at the same time, produce smaller vowel duration differences in words ending in /b d g/ and /p t k/ than native speakers (Flege et al., 1992; Suomi, 1976; and Moisio and Valento, 1976).

Data obtained in the present study bear on the third potential explanation. In the identification experiment, experienced late learners did not differ significantly from the native speakers. This might be taken to mean that they had received sufficient L2 input for, if they had not, how could they have shown "nativelike" perceptual sensitivity to vowel duration? In point of fact, however, the experienced late learners' perceptual sensitivity to vowel duration probably differed from that of the native speakers. Some of them reversed the labels used by the native speakers in the identification experiment. As a group, the experienced late learners showed less consistency than the native speakers; and their phoneme boundaries occurred at a different location than those of the native speakers. Finally, the experienced late learners showed little evidence of a categorization effect on their imitations of CVC stimuli. When data obtained for the imitation of isolated vowels and CVC stimuli were compared, the experienced late learners closely resembled the inexperienced late learners rather than the native speakers. This suggested that the experienced late learners had not integrated vowel duration into what Volyatis and Miller (1992) refer to as "the internal perceptual structure of voicing categories."

In summary, Chinese subjects who learned English in childhood seem to have established phonetic representations for /t/ and /d/ in the final position of English words. Like native speakers, the childhood L2 learners made vowels considerably longer before /d/ than /t/. They showed perceptual effects of vowel duration that closely resembled those obtained for the native speakers in an identification experiment, and in an experiment using the method of adjustment. Also, the childhood learners resembled the native speakers in showing a discontinuity in the vicinity of their /t/ vs /d/ phoneme boundaries in an imitation task. Inexperienced late learners, on the other hand, produced smaller voicing effects than the native speakers. They also differed from the native speakers in terms of consistency and slopes in the identification experiment, and showed significantly smaller effects of vowel duration in the method of adjustment experiment. Ex-
perceived late learners showed production and perception effects that were intermediate in magnitude to those observed for the native speakers and inexperienced late learners. The grouped data, but not the data for individual subjects, were consistent with the hypothesis that production accuracy is limited by the adequacy of adult learners' perceptual representations for L2 sounds and contrasts.

ACKNOWLEDGMENTS

This research was supported by NIH Grant DC00257. The author thanks E. James for her important contributions to the study. Thanks are also extended to K. Johnson, M. Munro, A. Schmidt, and A. Walley for comments on previous drafts of the article.


