Category restructuring during second-language speech acquisition

Ian R. A. MacKay
Department of Linguistics, University of Ottawa, Ottawa K1N 6N6, Canada

James Emil Flege[a)]
Department of Rehabilitation Sciences, University of Alabama at Birmingham, Birmingham, Alabama 35294

Thorsten Piske
English Department, Kiel University, Kiel 24098, Germany

Carlo Schirru
Department of Linguistics, University of Padua, Padua 35137, Italy

(Received 27 January 2000; revised 5 April 2001; accepted 10 April 2001)

This study examined the production of English /b/ and the perception of short-lag English /b d g/ tokens by four groups of bilinguals who differed according to their age of arrival (AOA) in Canada from Italy and amount of self-reported native language (L1) use. A clear difference emerged between early bilinguals (mean AOA = 8 years) and late bilinguals (mean AOA = 20 years). The late bilinguals showed a stronger L1 influence than the early bilinguals did on both the production and perception of English stops. In experiment 2, the late bilinguals produced a larger percentage of prevoiced English /b/ tokens than early bilinguals and native English (NE) speakers did. In experiment 3, the late bilinguals misidentified short-lag English /b d g/ tokens as /p t k/ more often than the early bilinguals and NE speakers did. Experiment 4 revealed that the frequencies with which the bilinguals prevoiced /b d g/ in Italian and English were correlated. The observed differences between the early and late bilinguals were attributed to differences in the quantity and quality of English phonetic input they had received, not to a greater likelihood by the early than late bilinguals to establish new phonetic categories for English /b d g/. © *2001 Acoustical Society of America.*
[DOI: 10.1121/1.1377287]

PACS numbers: 43.71.Ft [KRK]

I. INTRODUCTION

Research examining second language (L2) speech acquisition has shown that ''early'' bilinguals who were first exposed to their L2 as children generally produce and perceive L2 phonetic segments more like monolingual speakers of the target L2 than do ''late'' bilinguals who were first exposed to the L2 in adolescence or adulthood. For example, Flege *et al.* (1999a) found that early Italian–English bilinguals produced and perceived English vowels more accurately than late bilinguals did (see also Flege *et al.*, 1995, for consonants). Flege *et al.* (1999b) attributed the observed age effect to the greater likelihood of phonetic category formation for L2 vowels by the early than the late bilinguals. It appears that in some instances, however, L2 vowels and consonants (or ''speech sounds,'' for short) are not identical to native language (L1) speech sounds but nevertheless do not differ sufficiently from the closest L1 sound(s) for category formation to occur, even for early bilinguals. The question addressed by this study was whether phonetic learning takes place for such L2 speech sounds in the absence of category formation.

The Speech Learning Model, or SLM (Flege, 1995), proposes that category formation for an L2 speech sound will be blocked if it is perceptually ''equated'' with an L1 speech sound. Equivalence classification is said to occur when the perceived instances of an L2 speech sound continue to be assimilated by the closest L1 speech sound even after many years of L2 use (e.g., Flege and Hillenbrand, 1984; Flege, 1995). According to the SLM, equivalence classification does not prevent phonetic learning from occurring. By hypothesis, a merged (or ''composite'') category that subsumes the equated L1 and L2 speech sounds will develop over time. It will be used to process the equated L1 and L2 speech sounds, and will reflect the properties of all L1 and L2 speech sounds that have been perceived to be instances of either the L1 category, the L2 category, or the merged category that eventually replaces the original L1 category that has undergone restructuring during L2 acquisition due to its perceptual linkage to an L2 sound.

The development of a merged category during L2 acquisition is predicted to yield two interrelated effects, even after many years of L2 use. Bilinguals' production and perception of the L2 speech sound will continue to differ from L2 monolinguals' production and perception because the bilinguals' production and perception will partially resemble patterns that are typical for the corresponding L1 speech sound. At the same time, bilinguals' production and perception of the L1 speech sound will gradually change so as to partially

[a)]Author to whom correspondence should be addressed. Electronic mail: jeflege@uab.edu

resemble the corresponding L2 speech sound (see e.g., Flege, 1987; Flege *et al.*, 1995, 1999a).

The SLM's prediction of phonetic learning in the absence of category formation appears to differ from the views offered by other theories. According to the Native Language Magnet model (Kuhl, 2000, p. 106), listeners remain sensitive to subcategorical phonetic differences across languages, although they may no longer ''attend'' to such differences. One might hypothesize that L2 speech learning is governed by a kind of cross-language ''categorical perception.'' Such a view is related to the notion that the L1 phonological system acts as a kind of ''sieve'' that filters out the acoustic properties of L1 sounds that are needed to distinguish sounds in the L1 but not in the L2 (Polivanov, 1931; Trubetzkoy, 1939/1969; Hallé *et al.*, 1999). However, if L2 speech learning was governed by a kind of cross-language categorical perception, one would not expect phonetic learning to take place in the absence of category formation because the sensory input needed to guide learning for an L2 speech sound would be unavailable.

There are several reasons to think that, as proposed by the SLM (Flege, 1995), phonetic learning does take place for an L2 speech sound in the absence of category formation. Subcategorical phonetic differences across languages (or language varieties) appear to remain auditorily accessible to language learners. Flege and Hammond (1982) found that native English (NE) adults reproduced the voice onset time (VOT) values often heard in Spanish-accented English (i.e., values midway between those typical for short-lag and long-lag stops) when asked to mimic a Spanish accent in English sentences. Whalen *et al.* (1997) found that NE adults had difficulty discriminating unaspirated and aspirated allophones of English /p/ ([p] and [p^h]), but they could generally reproduce the VOT difference between the allophones in an imitation task. More globally, Munro *et al.* (2000) observed a measurable shift in the native language (L1) pronunciation of monolingual adults who were exposed to a nonprestige dialect of their L1 that differed from their native L1 dialect primarily in terms of subcategorical phonetic differences.

The results obtained in discrimination studies also suggest that subcategorical phonetic differences across languages remain auditorily accessible to language learners. Werker and Logan (1985) found that adult listeners showed sensitivity to certain cross-language phonetic differences under some task conditions (e.g., short ISIs in an AX task) but not others. An analysis of cortical evoked potentials led Sharma and Dorman (2000, p. 2702) to conclude that phonetic segments are processed at a sensory level that is ''not modified by exposure to the phonetic categories of a language'' and also at a level where ''language specific categories play a role.'' [1]

In this study, we examined Italian–English bilinguals' production of English /b/ and their perception of short-lag /b d g/ tokens in order to test for phonetic learning in the absence of category formation. We had several reasons to think that native Italian learners of English will not establish categories for English /b d g/. Italian /b d g/ and /p t k/ are realized with lead and short-lag VOT values, respectively (Magno-Caldognetto *et al.*, 1971, 1979). English /b d g/ are realized with short-lag VOT values or, less often, with lead VOT values (Lisker and Abramson, 1964; Flege and Eefting, 1986).[2] If English /b d g/ were realized with the lead VOT values that are typical for Italian (see experiment 1), such realizations would be acceptable in English; and so there would be no communicative pressure for Italian–English bilinguals to establish new categories for English /b d g/ (Port and Mitleb, 1983, p. 223).

Another reason to think that Italian–English bilinguals will not establish categories for English /b d g/ derives from universal constraints on phonetic systems. Stop consonants in the world's languages are realized with one of three modal VOT categories: lead (prevoiced), short-lag, and long-lag (Cho and Ladefoged, 1999). Keating (1984, p. 224) proposed that there may only be as many phonetic categories in languages as there are ''contrasting phonetic types.'' The same appears to hold true for individual bilinguals (see Flege and Eefting, 1988). Also, bilinguals usually identify short-lag stops in much the same way in their L1 and L2 (e.g., Elman *et al.*, 1977; Bohn and Flege, 1993). The establishment of short-lag categories for English /b d g/ is therefore likely to be preempted by existing Italian categories (viz., those for /p t k/).

The French–English and English–French bilinguals examined by Flege (1987) appear to have created merged categories for /t/. These bilinguals tended to produce both French /t/ and English /t/ with values that were intermediate to those observed for French and English monolinguals, respectively. No previous study has investigated the predicted effects of L1–L2 category merger for /b d g/, although L1 effects on the production of /b d g/ in an L2 have been observed for native Spanish and French learners of English (Caramazza *et al.*, 1973; Nathan, 1987; Williams, 1977b, 1979; Flege and Eefting, 1987). Based on this and the evidence reviewed earlier, we hypothesized that the Italian–English bilinguals examined here would detect differences between short-lag tokens of English /b d g/ and prevoiced tokens of Italian /b d g/ even if they did not establish categories for short-lag English /b d g/ tokens. If so, then the SLM (Flege, 1995) would predict the development of merged categories embracing the properties of corresponding L1 and L2 stops (e.g., English /b/ and Italian /b/). This led us to expect that the Italian–English bilinguals would differ from Italian monolinguals and also from English monolinguals. Specifically, the bilinguals should rely less on prevoicing as a perceptual cue to the identification of English stops as /b d g/. They should also prevoice English /b/ less often than Italian /b/ is typically prevoiced, but more often than is typical for English /b/.

If phonetic learning occurs in the absence of category formation, both predicted effects might be less evident for early than late Italian–English bilinguals. The early bilinguals examined in this study were probably exposed to more short-lag realizations of English /b d g/ in their lifetimes than the late bilinguals were. As in previous studies (e.g., Yeni-Komshian *et al.*, 2000), the early bilinguals had used their L2 longer than the late bilinguals had, and so were likely to have received more input from native speakers of the L2 (Jia

and Aaronson, 1999; Stevens, 1999). The predicted effects might also be greater for the bilinguals who continued to use their L1 (Italian) often than for the bilinguals who used their L1 relatively seldom. Hazan and Boulakia (1993) found that language dominance, which depends importantly on language use patterns, exerted a strong influence on the frequency of prevoicing in /b/'s spoken by French–English bilinguals (see also MacKay *et al.*, 2001; Meador *et al.*, 2000).

The present study was organized as follows. Experiment 1 provided an acoustic analysis of /b/ tokens that had been produced by English and Italian monolinguals. Experiment 2 examined the production of English /b/ by NE monolinguals and four groups of Italian–English bilinguals. Experiment 3 examined the same participants' identification of naturally produced /b d g/ and /p t k/ tokens. This experiment focused on the identification of word-initial /b d g/ tokens that had been realized with short-lag VOT values. Finally, experiment 4 examined the frequency with which Italian–English bilinguals prevoiced Italian /b d g/. Its aim was to test the prediction that the bilinguals whose productions of English /b d g/ most resembled NE speakers' productions would show the greatest influence of English on their productions of Italian /b d g/.

II. EXPERIMENT 1

Previous research (Magno-Caldognetto *et al.*, 1971, 1979) has shown that, as in other Romance languages, /b d g/ are produced with lead VOT values (i.e., are prevoiced) in Italian. The purpose of this experiment was to provide a direct comparison of /b d g/ production by monolingual native speakers of Italian and English.

A. Method

The participants were monolingual speakers of Italian (ten males, ten females) with a mean age of 25 years (range=19–33 years) and monolingual speakers of English (four males, eight females) with a mean age of 27 years (range=20–44 years). The Italian monolinguals were recorded in Padua, Italy; the English monolinguals were recorded in Birmingham, AL and Columbus, OH. The participants produced /bVdo/ nonwords (where V=/i e ɛ a ɔ o u/ for the Italian monolinguals, /i ɪ eɪ ɛ æ ɑ ʌ ɚ o u ʊ/ for the English monolinguals) after hearing four real words containing each target vowel of interest (e.g., *rido, fido, lido, nido* for Italian /i/). Twenty-one /b/ tokens produced by each Italian monolingual (7 vowel contexts × 3 repetitions) and 11 /b/ tokens produced by each English monolingual (one for each vowel context) were digitized at 22.05 kHz using a waveform editor (Cool Edit '96, Syntrillium Corp.).

Acoustic measurements were made from time domain waveforms displayed on the screen of a PC, supplemented by reference to digital spectrograms as needed. The duration of voicing lead (prevoicing) was measured from the onset of low-frequency periodicity to the onset of the /b/ release burst. In a subset of the tokens produced with lead VOT, the prevoicing ceased prior to the release of /b/. In these tokens, we also measured the duration of the silent gap from the cessation of prevoicing to the onset of the release burst. The remainder of tokens were produced without any prevoicing at all (i.e., were short-lag stops). VOT was measured in these tokens from the beginning of the release burst to the first upward-going zero crossing in the periodic portion (''vowel'') of the signal.

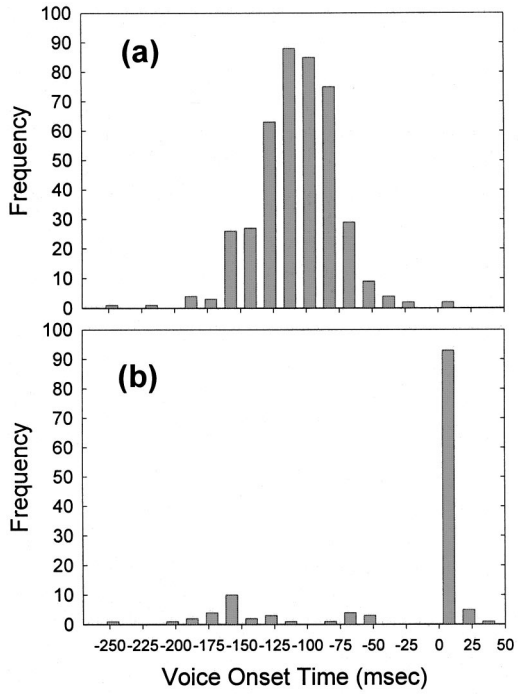

FIG. 1. The mean frequency of voice onset time (VOT) values for productions of /b/ in /bVdo/ nonwords by Italian monolinguals (a) and English monolinguals (b).

B. Results and discussion

Figures 1(a) and (b) show the frequency of VOT values in /b/ tokens produced in the same context by the Italian and English monolinguals, respectively. Of the 420 Italian /b/ tokens, 418 (99.5%) were prevoiced whereas just 32 (24%) of the 131 English /b/ tokens were prevoiced. Prevoicing in the Italian /b/ tokens averaged 106 ms in duration (s.d. =31, range=17–248), as compared to 136 ms (s.d.=9, range=51–251) for the prevoiced English /b/ tokens. The majority (99 or 76%) of the English /b/ tokens were realized as short-lag stops having a mean VOT of 11 ms (s.d.=3, range=1–31). Five of the 12 English monolinguals did not prevoice any /b/ tokens.

One unexpected, but nonetheless interesting, finding involved a qualitative difference in the prevoicing produced by the English and Italian monolinguals. Prevoicing ceased before the /b/ release in 23 (72%) of the prevoiced English /b/ tokens. The duration of the silent gap in these tokens averaged 47 ms (s.d.=24, range=13–95). Prevoicing did not cease before release in any of the prevoiced Italian /b/ tokens, however. The production of voicing through the entire closure of a labial stop requires an active volumetric expansion of the oral cavity, which must be learned (Flege *et al.*, 1987). It is uncertain if the presence of a silent gap before the release of a prevoiced stop will influence perception in Italian. However, the acoustic measurements and perceptual data obtained by Williams (1977a; see also Hazan and Bou-

TABLE I. Characteristics of the five groups of participants in experiment 2.

	AGE	GENDER	AOA	LOR	L1 USE	EDUC	B2FA
Native	50	9m				18	11.6
English	(5)	9f	⋯	⋯	⋯	(2)	(11.7)
Early-	50	8m	7	42	7%	14	12.1
low	(4)	10f	(3)	(4)	(4)	(3)	(8.9)
Early-	49	8m	8	40	43%	11	11.6
high	(6)	10f	(4)	(4)	(15)	(6)	(8.8)
Late-	51	10m	20	31	10%	2	15.0
low	(7)	8f	(3)	(8)	(5)	(2)	(12.0)
Late-	49	8m	20	29	53%	2	12.5
high	(8)	10f	(3)	(9)	(13)	(2)	(9.4)

Note: AGE=chronological age, in years; AOA and LOR=age of arrival and length of residence in Canada, in years; L1 USE=self-reported overall percentage use of Italian; EDUC=years of formal eduction in Canada; B2FA=best two-frequency average obtained in the heating screening, in dB. Standard deviations are in parentheses.

lakia, 1993, for French) suggested to her that such a silent gap may encourage Spanish speakers to hear /b d g/ tokens as /p t k/. Further research will be needed to determine if native Italian speakers learn to avoid producing /b/ with a silent gap in prevoicing in order to prevent perceptual confusions.

The results obtained here confirmed (e.g., Magno-Caldognetto *et al.*, 1971; Lisker and Abramson, 1964) that /b/ is prevoiced more frequently in Italian than English. These results may not necessarily generalize to the native dialect (or variety) of all of the Italian–English bilinguals examined subsequently in this study. However, it seems reasonable to think that when native Italian speakers first begin to speak English, they will prevoice English /b d g/ more frequently than NE speakers do as the result of cross-language differences in the phonetic implementation of /b d g/.

III. EXPERIMENT 2

Research with French–English and Spanish–English bilinguals has shown that they prevoice English /b d g/ more often than NE speakers do because /b d g/ is always, or nearly always, prevoiced in their L1s (Caramazza *et al.*, 1973; Hazan and Boulakia, 1993; Williams, 1977a, b, 1979; Nathan, 1987). The question addressed here was whether the same would hold true for Italian–English bilinguals and, if so, whether the size of the native versus non-native difference in prevoicing would vary as a function of AOA and/or L1 use.

A. Method

1. Participants

Eighteen participants were NE speakers who were not proficient in another language, and 72 were native speakers of Italian who had emigrated from Italy to Canada. As summarized in Table I, the bilinguals were assigned to one of four groups of 18 participants each (roughly half female).[3] Each of the 90 participants passed a pure-tone hearing screening that established thresholds for both ears at 500, 1000, 2000, 4000, and 8000 Hz. Preliminary analyses revealed that the combination of thresholds that was most strongly correlated with the perception data presented in experiment 3 was the best two-frequency average threshold (or ''B2FA,'' for short). The B2FA of the five groups did not differ significantly [$F(4,85)=0.3$, $p>0.10$].

The bilinguals were selected based on their age of arrival (AOA) in Canada and amount of continued L1 (Italian) use. Thirty-six ''early'' bilinguals arrived in Canada between the ages of 2–13 years (mean=8 years, s.d.=4), whereas 36 ''late'' bilinguals arrived between the ages of 15–26 years (mean=20 years, s.d.=3). The early and late bilinguals were then subdivided according to amount of continued L1 use, 1%–15% for the ''low-use'' bilinguals (mean =8%, s.d.=4) vs 25%–80% for the ''high-use'' bilinguals (mean=49%, s.d.=15).

A (2) AOA×(2) L1 use ANOVA revealed that the AOA difference between the low-use and high-use participants (13.2 vs 13.9 years) was nonsignificant [$F(1,68)=0.7$, $p>0.10$]. The lack of an AOA×L1 use interaction [$F(1,68)=0.9$, $p>0.10$] indicated that the AOA differences between the two groups of high-use participants (early-high, late-high), and between the two groups of low-use participants (early-low, late-low), were comparable. The early and late bilinguals differed significantly (25% vs 31%) according to amount of L1 use [$F(1,68)=7.3$, $p<0.01$]. However, the AOA×L1 use interaction in the analysis of self-reported percentage L1 use was nonsignificant [$F(1,68)=2.45$, $p>0.10$].

All but 3 of the 72 bilinguals had lived in Canada for at least 20 years. The early and late bilinguals differed significantly (means=41 vs 30 years) in length of residence (LOR) in Canada [$F(1,68)=47.5$, $p<0.01$], whereas the LOR difference between the low-use and high-use bilinguals (means=37 vs 35 years) was nonsignificant [$F(1,68)=1.4$, $p>0.10$]. Years of education in schools where the L2 is used as the language of instruction is known to affect certain aspects of L2 acquisition (e.g., Flege *et al.*, 1999b). The difference in number of years of education that the early and late bilinguals had obtained in English-speaking schools in Canada (means=13 vs 2 years) differed significantly [$F(1,68)=168.9$, $p<0.01$]. However, the education difference between the high-use and low-use bilinguals (means =7 vs 8 years) was nonsignificant [$F(1,68)=2.2$, $p>0.10$].

2. Speech materials

A delayed repetition procedure was used to elicit the production of word-initial /b/ tokens. A male and a female native speaker of Canadian English produced 15 test words including one token each of *bade*, *booed*, and *bed* and two tokens of *bad*. Their productions were digitized[4] and then presented via loudspeakers at the beginning of carrier phrases in two conditions. In the ''one-word'' condition, each test word to be repeated was followed by ''*...is the next word to say*.'' The participants repeated the target word after hearing the entire carrier phrase. In the ''three-word'' condition, the test words of interest occurred as the second member of three-word series (e.g., *Hid...bad...heed*) followed by ''*...are the next words to say*.'' The participants repeated all three words after hearing the entire carrier phrase. These

TABLE II. The mean percentage of word-initial English /b/ tokens that were prevoiced, and the percentage of stops that were produced with prevoicing that ceased before the stop release.

	Isolated word		Middle word in series	
	Prevoiced	Ceased	Prevoiced	Ceased
Native English	29% (33)	2% (5)	34% (33)	4% (13)
Early-low	61% (32)	17% (21)	53% (32)	18% (19)
Early-high	69% (35)	13% (18)	61% (36)	13% (17)
Late-low	79% (27)	5% (10)	80% (25)	8% (12)
Late-high	86% (17)	4% (9)	73% (31)	3% (6)

Note: Standard deviations are in parentheses.

procedures yielded 1800 /bVd/ words (90 participants $\times 2$ talkers $\times 5$ words $\times 2$ elicitation conditions) for analysis. Of these, seven words in the one-word condition and 115 words in the three-word condition were declared missing because of noise or because they were not repeated. In 21 other instances (of which 18 were in the three-word condition) the participants said a word that resembled the target word to be repeated (e.g., *bid* instead of *bad*). The /b/'s in these ''substitute'' words, along with the other /bVC/ target words, were digitized and measured as described in experiment 1.

B. Results

The total number of words that each participant repeated correctly in each condition (maximum=10) was tabulated. A (5) group$\times$(2) condition ANOVA examining these scores yielded a significant two-way interaction [$F(4,85)=5.3$, $p<0.01$]. Simple effects tests revealed that the effect of condition (isolated=9.9, series=9.8) was nonsignificant for the NE speakers [$F(1,85)=0.1$, $p>0.10$], but significant for all four bilingual groups (early-low: 9.9 vs 8.9; early-high: 9.9 vs 8.3; late-low: 9.9 vs 8.5; late-high: 9.9 vs 7.6) (F-values ranging from 7.9 to 43.3, $p<0.01$). This suggested that the bilinguals experienced greater difficulty than the NE speakers did in retaining three English words in working memory prior to repeating them. However, inasmuch as this finding does not bear directly on how /b/ was produced, it will not be discussed further.

Of the 1042 /b/ tokens that were measured, 646 (62%) were prevoiced. As shown in Table II, all four groups of bilinguals prevoiced /b/ more often than the NE speakers did in both conditions. The ''percent prevoiced'' scores were submitted to a mixed-design (5) group$\times$(2) condition ANOVA to determine if any group of bilinguals prevoiced /b/ more frequently than the NE speakers did. This analysis yielded a significant main effect of group [$F(4,85)=8.98$, $p<0.01$], a nonsignificant effect of condition [$F(4,85)=3.20$, $p>0.05$], and a nonsignificant two-way interaction [$F(4,85)=1.72$, $p>0.10$]. To test for native versus nonnative differences, the average of scores obtained in the two conditions by the four bilingual groups was compared to the average scores obtained for the NE speakers in a series of t-tests. These tests revealed that all four bilingual groups prevoiced /b/ more often (early-low: 57%, early-high: 65%, late-low: 79%, late-high: 79%) than the NE speakers did (mean=31%) (Bonferroni $p<0.05$). A supplementary Tukey's test did not reveal any significant differences between the four groups of bilinguals ($p>0.10$).

Experiment 1 revealed that prevoicing often ceased before the release burst in stops produced by English monolinguals, but never in stops produced by Italian monolinguals. In this experiment, prevoicing ceased before release in 149 of the English /b/ tokens that were examined. Table II shows the mean percentage of prevoiced /b/ tokens produced in which prevoicing ceased before the release. The duration of these silent gaps averaged 35 ms. More early than late bilinguals produced one or more /b/ tokens in which prevoicing ceased before the release (early-low: 16 participants, early-high: 13, late-low: 9, late-high: 5). The number of NE speakers who did so was small ($n=4$), apparently because they prevoiced so few stops.

The percentages of prevoiced stops in which voicing ceased before release were examined in a (5) group$\times$(2) condition ANOVA. It yielded a significant main effect of group [$F(4,85)=5.2$, $p<0.01$], a nonsignificant effect of condition [$F(4,85)=0.5$, $p>0.10$], and a nonsignificant two-way interaction [$F(4,85)=0.1$, $p>0.10$]. The average of scores obtained in the two conditions for the four bilingual groups were compared to the average scores obtained for the NE speakers in a series of t-tests. The two groups of early bilinguals were found to have produced more prevoiced stops in which the prevoicing ceased before the release than the NE speakers did (early-low: 18%, early-high: 13%, NE: 3%), whereas neither group of late bilinguals (late-low: 6%, late-high: 3%) differed from the NE speakers (Bonferroni $p<0.05$). A supplementary Tukey's test that tested for all possible between-group differences revealed that the Early-low group produced more such stops than the NE and the Late-low groups did ($p<0.01$).

The analysis just presented is potentially misleading inasmuch as the NE speakers produced so few prevoiced /b/'s. We therefore computed the percentage of the 20 /b/ tokens produced by each participant that were ''fully'' prevoiced, that is, had prevoicing that continued without interruption until the release burst. (We pooled the data obtained in the two conditions because the earlier analyses indicated that it would be appropriate to do so.) Figure 2 shows that the NE speakers produced fewer fully prevoiced stops than did any of the four bilingual groups. The one-way ANOVA examining these scores was significant [$F(4,85)=9.8$, $p<0.01$]. A series of t-tests revealed that the participants in the late-high, late-low, and early-high groups produced more fully prevoiced /b/'s (means=52%, 73%, 76%) than the NE speakers did (mean=28%; Bonferroni $p<0.05$), whereas the early-low participants (mean=39%) did not differ significantly from the NE speakers (Bonferroni $p>0.10$). A supplementary Tukey's test testing all possible between-group differences revealed that the late-high and late-low groups fully prevoiced /b/ more often than the NE and the early-low groups did ($p<0.01$).

The scores obtained for the four groups of bilinguals were examined in a series of (2) AOA$\times$(2) L1 use

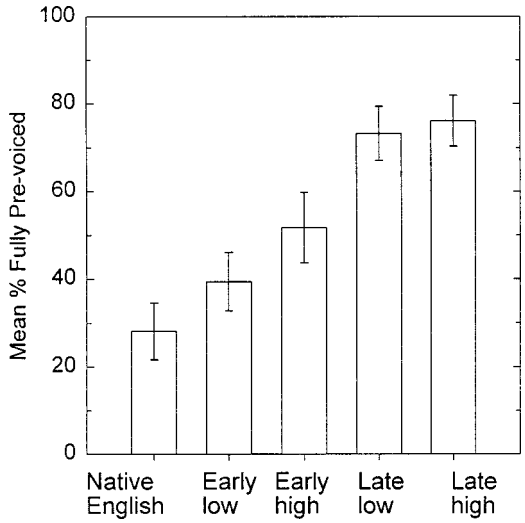

FIG. 2. The mean percentage of fully prevoiced stops produced by the participants in five groups. The brackets enclose ±1.0 SE.

ANOVAs. The variables examined in these analyses were the average percentage of /b/ tokens that were prevoiced in the two conditions (analysis 1); the average percentage of prevoiced /b/ tokens in which the pre-voicing ceased prior to the /b/ release (analysis 2); and the average percentage of /b/ tokens that were produced with full prevoicing (analysis 3). All three analyses revealed a significant main effect of AOA [analysis 1: $F(1,68)=8.1$; analysis 2: $F(1,68)=12.8$; analysis 3: $F(1,68)=18.6$, $p<0.01$]. The early bilinguals produced a smaller percentage of /b/ tokens with prevoicing than the late bilinguals did (61% vs 79%). They produced a larger percentage of prevoiced /b/ tokens in which voicing ceased before the release than the late bilinguals did (15% vs 5%). And they produced a smaller percentage of fully prevoiced /b/ tokens than the late bilinguals did (46% vs 75%). However, in all three analyses, the effect of L1 use was nonsignificant [analysis 1: $F(1,68)=0.3$; analysis 2: $F(1,68)=1.7$; analysis 3: $F(1,68)=1.3$, $p>0.10$], and the two-way interaction was nonsignificant [analysis 1: $F(1,68)=0.4$; analysis 2: $F(1,68)=0.1$; analysis 3: $F(1,68)=0.5$, $p>0.10$]

C. Discussion

The question addressed here was whether bilinguals who were experienced in English would show evidence of phonetic learning for English /b/. It appears that phonetic learning did take place. The percentage of English stops that were prevoiced by the four groups of Italian–English bilinguals ranged from an average of 57% for the early-low group to 79% for the two late bilingual groups. The percentage of stops that were fully prevoiced ranged from 39% for the early-low group to 76% for the early-high group. These percentages are in every case lower than the percentage observed for the production of Italian /b/ by Italian monolinguals in experiment 1 (99.5%).

The present results agree with previous studies examining French–English and Spanish–English bilinguals (Nathan, 1987; Williams, 1977b; Williams, 1979; Hazan and Boulakia, 1993) in showing that the Italian-English bilinguals prevoiced more often than NE speakers did. All four bilingual groups in this study prevoiced English /b/ more frequently than the NE speakers did. When the percentages of /b/ tokens that were fully prevoiced were examined, the participants in all of the bilingual groups except the early-low group differed significantly from the NE speakers. Other analyses revealed that the late bilinguals produced a significantly larger percentage of prevoiced /b/ tokens, and also a larger percentage of fully prevoiced /b/'s than the early bilinguals did. However, amount of continued L1 (Italian) use was not found to influence English /b/ production significantly.

A closer approximation to English phonetic norms by the early than late bilinguals might be attributed to the passing of a critical period (e.g., Scovel, 1988). However, the difference might have arisen from the quantity and quality of L2 input. Recall that the early bilinguals had lived longer in Canada than the late bilinguals had (41 vs 30 years); received more education in English-speaking Canadian schools (13 vs 2 years); and reported using Italian less overall (25% vs 31%). The early bilinguals may, therefore, have used English more often with NE speakers than the late bilinguals had, and may have been exposed to Italian-accented English less often than the late bilinguals had been.

IV. EXPERIMENT 3

Experiment 1 revealed that phonologically voiced stops are prevoiced more often by Italian than English monolinguals. Here we sought to determine if Italian–English bilinguals would misidentify short-lag tokens of /b d g/ as /p t k/ more often than NE speakers due to the lack of prevoicing.

A. Method

The participants from experiment 2 were tested in a quiet room in a single session using a notebook computer after having produced the speech materials examined earlier. The perceptual stimuli used here were non-words of the form /ˈCɑmɑ/, /ˈmɑCɑ/, and /ɑˈmɑC/ (where ''C'' indicates a token of /b d g p t k/). The stimuli were spoken by two NE males, then digitized at 22.05 kHz.[5] The /b d g/ tokens in the /ˈCɑmɑ/ stimuli were realized as short-lag stops having an average VOT of 15 ms. The /b d g/ tokens in the /ˈmɑCɑ/ stimuli, on the other hand, were produced with voicing through most (67 ms or 93%) of the closure interval. The same held true for the /b d g/ tokens in the /ɑˈmɑC/ stimuli (voicing in 80 ms or 94% of the closure intervals).

The stimuli just described were mixed with varying levels of noise to provide stimuli in which ceiling effects would not be evident. After the 36 stimuli (2 talkers×18) were normalized to 50% of full scale, three copies were made of each. The copies were digitally added to three 1000-ms pink noise segments. This yielded 36 stimuli each having S/N ratios of 16, 10, and 4 dB[6] in addition to the original 36 no-noise stimuli.

The 144 stimuli were presented via headphones (Sennheiser Model HD535) at a comfortable level that was determined individually for each participant before the experiment began. Test stimuli similar to the experimental stimuli

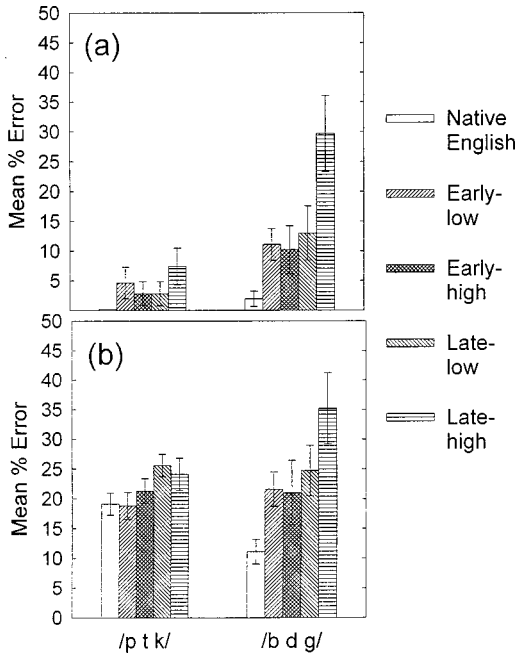

FIG. 3. The mean percentage of errors that the participants in five groups made in the identification of word-initial /p t k/ and /b d g/ tokens that were presented in the quiet (a) or in noise (b). The brackets enclose ±1.0 SE.

were played out. The volume was adjusted repeatedly until each participant confirmed hearing the test stimuli ''clearly'' in both ears.

The participants were told to identify the stop consonant in each stimulus as one of six stop consonants (''p, t, k, b, d, g'') by clicking one of six buttons shown on the computer screen.[7] Stimuli produced by the two talkers were presented in separate, counterbalanced blocks. The 36 no-noise stimuli were randomly presented a single time. The three sets of with-noise stimuli were then presented in a fixed order such that the stimuli became progressively noisier over the final three blocks (i.e., S/N levels of 16, 10, and 4 dB). A fixed order was used to help avoid ceiling effects that might arise as the participants gained familiarity with the small set of stimuli. The participants were given practice with feedback using stimuli produced by another talker before the experiment began. However, they received no feedback during the experiment. The participants were required to label the stop in each stimulus, and were told to guess if uncertain. The interval between each response and the next stimulus was 1.0 s.

B. Results

The percentages of errors that each participant made identifying /b d g/ and /p t k/ in the no-noise condition were calculated. The identification errors arose from confusions of the voicing feature (e.g., /p/ tokens heard as /b/), place of articulation (e.g., /d/ tokens heard as /g/), or both (e.g., /t/s heard as /g/). As shown in Fig. 3(a), the four groups of bilinguals erred more in identifying the /b d g/ than the NE speakers did, but did not differ much from the NE speakers for /p t k/.

The percent error scores were examined in a (5) group ×(2) phonological voicing ANOVA with repeated measures on the voicing factor. It yielded significant main effects of group [$F(4,85)=6.22$, $p<0.01$] and voicing [$F(1,85)=24.9$; $p<0.01$], as well as a significant two-way interaction [$F(4,85)=3.2$, $p<0.05$]. The interaction arose, in part, because the simple effect of voicing was nonsignificant for the NE group [$F(1,17)=2.12$, $p>0.05$], the early-low group [$F(1,17)=3.7$, $p>0.05$] and the early-high group [$F(1,17)=4.2$, $p>0.05$], whereas participants in the late-low [$F(1,17)=5.6$, $p<0.05$] and the late-high group [$F(1,17)=10.1$, $p<0.05$] made significantly more errors identifying /b d g/ than /p t k/. Also, a significant effect of group was obtained for /b d g/ [$F(4,85)=5.9$; $p<0.01$] but not /p t k/ [$F(4,85)=1.5$; $p>0.10$]. A series of *t*-tests was carried out to determine which bilingual group(s) differed from the NE speakers for /b d g/ in the no-noise condition. The late-high participants were found to have made more errors for /b d g/ than the NE speakers did (30% vs 2%; Bonferroni $p<0.05$) whereas those in the remaining three bilingual groups (early-low: 11%, early-high: 10%, late-low: 13%) did not differ from the NE speakers (Bonferroni $p>0.10$.)

As expected, a preliminary analysis revealed that the frequency of errors increased systematically as the stimuli became progressively more noisy (means=10% at 16 dB, 17% at 10 dB level, and 47% at the 4 dB S/N level). However, as in a study by MacKay *et al.* (2001), adding noise appeared to exert a comparable effect on the responses given by all five groups,[8] so we calculated an average percent error score for /b d g/ and /p t k/ in the three with-noise conditions. Each of these scores was based on 18 judgments (2 talkers ×3 stops×3 S/N levels). Figure 3(b) shows the average percent error scores obtained for /b d g/ and /p t k/ in the with-noise conditions.

A (5) group×(2) phonological voicing ANOVA examining the average with-noise scores yielded a significant main effect of group [$F(4,85)=4.2$, $p<0.01$]. The main effect of phonological voicing was nonsignificant [$F(1,85)=0.2$, $p>0.10$] but entered into a significant interaction with group [$F(4,85)=2.5$, $p=0.05$]. The interaction arose, in part, because the NE speakers made fewer errors for /b d g/ than /p t k/ [$F(1,17)=13.7$, $p<0.01$] whereas the simple effect of voicing was nonsignificant for all four bilingual groups [*F*-values ranging from 0.0 to 3.6; $p>0.05$]. Also, the simple effect of group was significant for /b d g/ [$F(4,85)=3.9$, $p<0.01$] but not /p t k/ [$F(4,85)=1.9$, $p>0.10$]. A series of four *t*-tests revealed that the late-high participants made more errors for /b d g/ than the NE speakers did (35% vs 11%; Bonferroni $p<0.05$), whereas the other three bilingual groups (early-low: 22%, early-high: 21%, late-low: 25%) did not differ from the NE speakers (Bonferroni $p>0.10$).

To summarize so far, the same results were obtained for stops presented in the no-noise and with-noise conditions. Only the late-high participants made more errors identifying short-lag tokens of English /b d g/ than the NE speakers did. The late-high participants may have misidentified the English /b d g/ tokens often because they were produced without the prevoicing that is typical for Italian /b d g/ (see experiment 1). We cannot be certain of this, however, because the scores we examined included place of place of articula-

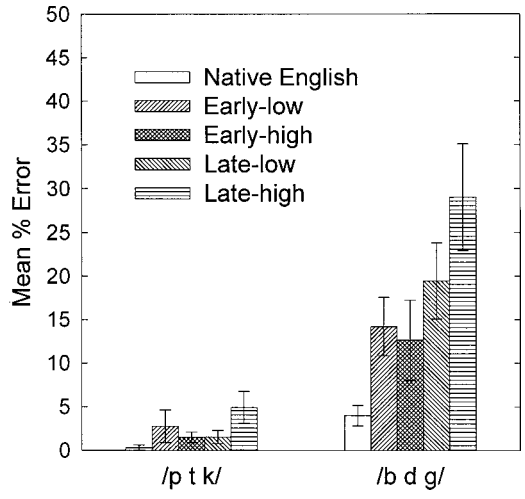

FIG. 4. The mean percentage of errors made by the participants in five groups that were due to misidentification of the voicing feature in word-initial tokens of /p t k/ and /b d g/ presented in noise. The brackets enclose ±1.0 SE.

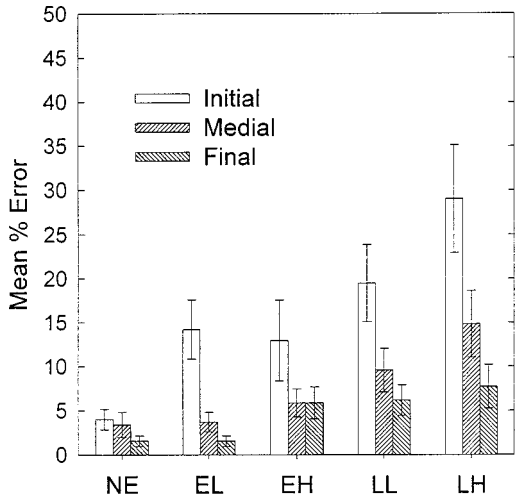

FIG. 5. The mean percentage of times that word-initial, word-medial, and word-final tokens of /b d g/ presented in noise were identified as /p t k/ by five groups of participants. The brackets enclose ±1.0 SE.

tion errors (e.g., the labeling of /t/ as /k/) as well as voicing errors.

We next calculated the percentage of errors involving just the voicing feature (e.g., the labeling of /b/ tokens as /p/) in the with-noise conditions.[9] As shown in Fig. 4, the four bilingual groups made more errors identifying the voicing feature in /b d g/ than the NE speakers did, but did not differ much from the NE speakers for /p t k/. A (5) group×(2) phonological voicing ANOVA examining the voicing error scores yielded significant main effects of group [$F(4,85) = 6.1$, $p < 0.01$] and voicing [$F(1,85) = 46.3$, $p < 0.01$] and a significant interaction [$F(4,85) = 3.0$, $p < 0.05$]. The interaction arose because the simple effect of group was significant for /b d g/ [$F(4,85) = 4.7$, $p < 0.01$] but not /p t k/ [$F(4,85) = 1,9$, $p > 0.10$]. A series of *t*-tests revealed that the late-low and late-high participants misidentified the /b d g/ tokens as /p t k/ more often (means = 19% and 29%) than the NE speakers did (mean = 4%) (Bonferroni $p < 0.05$), whereas neither group of early bilinguals (early-low: 14%, early-high: 13%) differed significantly from the NE speakers (Bonferroni $p > 0.10$).

The percentages of voicing errors made for /b d g/ by the four bilingual groups were examined separately in a (2) AOA×(2) L1 use ANOVA. It revealed that the late bilinguals made more voicing errors than the early bilinguals did (means = 24% vs 13%) [$F(1,68) = 5.3$, $p < 0.05$]. However, the difference between the high-use and low-use bilinguals (means = 21% vs 17%) was nonsignificant [$F(1,68) = 0.7$, $p > 0.10$], as was the two-way interaction [$F(1,68) = 1.4$, $p > 0.10$]. One possible explanation for a difference between the early and late bilinguals is a difference in the amount of English-language input. In support of this, the early–late difference for /b d g/ became nonsignificant [$F(1,67) = 3.0$, $p = 0.09$] when LOR was used as a covariate in a (2) AOA ×(2) L1 use ANOVA.

Unlike the word-initial /b d g/ tokens, those occuring in the word-medial and word-final /b d g/ positions were produced with closure voicing. We compared the frequency of voicing errors in word-initial, word-medial, and word-final /b d g/ tokens to evaluate how the absence of prevoicing in word-initial stops affected the bilinguals' identification judgments. As shown in Fig. 5, the participants in all five groups made more voicing errors for word-initial stops than for word-medial or word-final stops. However, the effect of position was greater for the four bilingual groups than for the NE speakers.

The voicing error scores shown in Fig. 5 were submitted to a (5) group×(3) word position ANOVA. It yielded significant main effects of group [$F(4,85) = 6.5$, $p < 0.01$] and position [$F(2,170) = 28.1$, $p < 0.01$] and also a significant two-way interaction [$F(8,170) = 2.3$, $p < 0.05$]. The interaction arose because word position affected the NE speakers and Italian-English bilinguals differently. The simple effect of position was nonsignificant for the NE speakers [$F(2,34) = 1.3$, $p > 0.10$] but it was significant (or marginally significant) for all four bilingual groups [late-low: $F(2,34) = 6.7$, $p < 0.01$; late-high: $F(2,34) = 9.6$, $p < 0.01$; early-low: $F(2,34) = 10.2$, $p < 0.01$; early-high: $F(2,34) = 2.8$; $0.05 < p < 0.10$]. Tukey's tests revealed that the participants in the early-low, late-low, and late-high groups made more errors for the word-initial /b d g/ tokens than for the word-medial or the word-final /b d g/ tokens ($p < 0.05$).

C. Discussion

This experiment showed that the late but not the early Italian-English bilinguals misidentified word-initial English /b d g/ tokens as /p t k/ more frequently than the NE speakers did. It appears that most of the late bilinguals' errors were due to the absence of prevoicing, for the word-initial /b d g/ stimuli were realized as short-lag stops rather than as lead (prevoiced) stops, as in Italian (see experiment 1). In a follow-up analysis, we compared the frequency of voicing errors for the word-initial, word-medial, and word-final /b d g/ tokens. The effect of position was nonsignificant for the NE speakers, but it was significant (or marginally significant) for all four bilingual groups, whose error rates were highest for the word-initial stops. We attributed the relatively

high rate of errors for the initial stops to the fact that they, but not the medial or final stops, lacked closure voicing. We acknowledge, however, that the medial and final tokens of /b d g/ differed from the word-initial /b d g/ in more than just closure voicing (see, e.g., Flege *et al.*, 1992).

Important individual differences existed in the frequency of voicing identification errors. The following number of 18 participants in the five groups made no errors identifying the voicing feature in the short-lag /b d g/ tokens: Native English: 9, early-low: 4, early-high: 6, late-low: 4, late-high:2. The following number of participants made more than 30% voicing errors for the word-initial /b d g/ tokens: Native English: 0, early-low: 2, early-high: 2, late-low: 6, late-high: 7. The basis of intersubject differences among the Italian–English bilinguals is uncertain. They might have arisen from differences in speech-learning ability, from differences in phonological short-term memory (MacKay *et al.*, 2001), from differences in the quantity and quality of phonetic input that had been received from native speakers of English, from degree of motivation to sound like a native speaker, or from some combination of factors.

V. EXPERIMENT 4

The final experiment examined Italian–English bilinguals' production of Italian /b d g/. Its purpose was to determine if learning English would cause Italian–English bilinguals to prevoice Italian /b d g/ less than Italian monolinguals. We examined the percentage of times that Italian /b d g/ were fully prevoiced because, as shown in experiment 1, /b/ is fully prevoiced less often by English than Italian monolinguals.

Four groups of bilingual participants differing in AOA and L1 use (early-low, early-high, late-low, late-high) participated. Experiment 2 showed that participants in an early-low group produced a smaller percentage of English /b/ tokens with full prevoicing than the participants in two late bilingual groups (late-low, late-high) did. A ''phonetic category merger'' hypothesis (see the Introduction and Flege, 1987, 1991, 1995) would therefore lead one to expect a greater influence of English on the production of Italian /b d g/ by participants in the early-low group than by late Italian–English bilinguals.

Extrapolating from the experiment 2 results for English, one might predict that only the participants in an early-low group would produce fewer fully prevoiced /b d g/ tokens than Italian monolinguals. It would have been ideal to obtain data from Italian monolinguals for this experiment, but we were unable to do so. We therefore evaluated the merger hypothesis by testing the prediction that participants in an early-low group would produce fewer fully prevoiced Italian /b d g/ tokens than participants in a late-low group would, whereas early and late bilinguals who often spoke Italian (early-high, late-high) would not differ significantly.

A. Method

Fifty bilinguals from experiments 2 and 3 (14 early-low, 13 early-high, 9 late-low, and 13 late-high participants) returned for this experiment a year later. Fourteen new participants were recruited in Ottawa to provide four groups of 16 participants each who differed in AOA and L1 use (see Table III). The 64 bilinguals had a mean age of 50 years (s.d.=6, range=30–63), and had been living in Canada for an average of 36 years (s.d.=8, range: 9–51 years). None reported an auditory disorder; and all passed a pure-tone hearing screening at octave frequencies between 500 and 4000 Hz (*re*: 35 dB HL). The 32 early bilinguals had an average AOA of 8 years (s.d.=3, range: 3–13) whereas the 32 late bilinguals had an average AOA of 19 years (s.d.=3, range: 15–28). The 32 low-use bilinguals reported using Italian 8% of the time on the average (s.d.=4, range: 2%–15%) whereas the 32 high-use bilinguals reported using Italian 46% of the time (s.d.=15, range: 29%–75%).

TABLE III. Characteristics of the four groups of native Italian participants in experiment 4.

	AGE	GENDER	AOA	LOR	L1 USE	EDUC	AGE
Early-low	49 (4)	7m 9f	7 (3)	42 (4)	6% (3)	14 (3)	49 (4)
Early-high	49 (6)	8m 8f	8 (4)	41 (6)	40% (13)	11 (5)	49 (6)
Late-low	51 (6)	7m 9f	18 (3)	33 (5)	10% (5)	2 (2)	51 (6)
Late-high	49 (8)	7m 9f	20 (4)	29 (9)	52% (15)	1 (2)	49 (8)
M	50 (6)	···	13 (7)	36 (8)	27% (22)	7 (6)	50 (6)

Note: AGE=chronological age, in years. Standard deviations are in parentheses; AOA and LOR=age of arrival and length of residence in Canada, in years; L1 USE=self-reported percentage use of Italian; EDUC=years of eduction in Canada, in years.

An adult female native speaker of Italian produced a list of 24 Italian words including nine words that began with prevoiced /b d g/ tokens (*babbo, bada, batto, dado, danno, data, gamma, gatta, gatto*). Tokens of these words that had been digitized at 11.025 kHz (with 16-bit resolution) were randomly presented to the participants via headphones at a comfortable level. The participants were told to repeat each word a single time after hearing it presented twice in a row, then to choose the correct English definition for the word from among the three definitions shown on the computer screen (e.g., ''a female cat'' for *gatta*). This last procedure ensured that the bilinguals' L2 (English) system was activated as they repeated the Italian words (see Grosjean, 1999).

The participants' productions of the 24 Italian words were recorded using a portable DAT tape recorder (Sony TCD8). The words beginning in /b d g/ were digitized (11.025 kHz) and then measured as in experiments 1 and 2. A total of 26 (4.5%) of the 576 test words were declared missing because of noise or failure to repeat.

B. Results

Of the 550 nonmissing /b d g/ tokens, 456 (83%) were prevoiced and 94 (17%) were realized as short-lag stops. Voicing ceased before the stop release in 59 (13%) of the prevoiced tokens. The duration of the silent gap in these tokens averaged 26 ms in duration. The percentage of /b d g/ tokens in which voicing continued without interruption until

release was calculated for each participant (maximum=9). All four bilingual groups produced a smaller percentage of fully prevoiced stops (early-low: 55%, early-high: 76%, late-low: 85%, late-high: 73%) than was observed for /b/'s spoken by the Italian monolinguals in experiment 1 (99.5%). The same held true if we consider only the percentage of fully prevoiced Italian /b/ tokens (early-low: 65%, early-high: 81%, late-low: 90%, late-high: 85%).

A two-way ANOVA examining the percentage of fully prevoiced Italian /b d g/ tokens indicated that the effect of AOA (early=65%, late=79%) was significant [$F(1,60)=2.9$, $p<0.01$] whereas the effect of L1 use (low use =70%, high use=74%) was not [$F(1,60)=0.5$, $p>0.10$]. Simple effects tests indicated that a significant two-way interaction was obtained [$F(1,60)=7.2$, $p>0.01$] because the effect of AOA was significant only for the low-use bilinguals. As predicted, the participants in the early-low group produced fewer fully prevoiced /b d g/ tokens than did those in the late-low group [$F(1,30)=13.0$, $p<0.01$], whereas the difference between late-high and early-high groups was nonsignificant [$F(1,30)=0.1$, $p>0.10$].

The low-use bilinguals tended to produce fewer fully prevoiced /b d g/ tokens than the high-use bilinguals did. However, the differences between the early-low and the early-high groups [$F(1,30)=4.0$], and those between the late-low and late-high groups [$F(1,30)=3.4$], were only marginally significant ($0.05<p<0.10$).

C. Discussion

The Italian–English bilinguals fully prevoiced Italian /b/ tokens less often (early-low: 65%, early-high: 81%, late-low: 90%, late-high: 85%) than Italian monolinguals did in experiment 1 (viz., 99.5%). There were, of course, important methodological differences between this experiment and experiment 1. Here we examined the immediate repetition of real Italian words beginning with fully prevoiced /b d g/ tokens whereas the participants in experiment 1 produced nonwords without a direct model. Still, these results suggest that learning English affected how all four groups of Italian–English bilinguals produced /b d g/ in their L1, Italian.

The SLM's (Flege, 1995) prediction that the participants in the early-low group would fully prevoice Italian /b d g/ less often than those in late-low group was confirmed. That is, the bilinguals whose productions of English /b/ was most English-like showed the greatest influence of English on their production of Italian /b d g/. In fact, for the 50 Italian–English bilinguals who participated in this experiment as well as in experiment 2, a significant correlation existed between the percentage of fully prevoiced English and Italian stops that were produced [$r(48)=0.47$, $p<0.01$]. That is, the less often the Italian–English bilinguals fully prevoiced English /b/, the less often they fully prevoiced Italian /b d g/. Significant partial correlations were also obtained when variations in self-rated ability to speak and understand Italian were partialled out [$r(46)=0.47$, $p<0.01$], when variations in self-rated ability to speak and understand English were partialled out [$r(46)=0.42$, $p<0.01$], and when four additional variables (age, LOR, percent use of Italian, and AOA) were partialled out [$r(44)=0.43$, $p<0.01$].

These findings supported our working assumption (see the Introduction) that even those bilinguals who learned English as children and seldom use Italian (early-low) had not developed new phonetic categories for English /b d g/. We suspect that most if not all of the bilinguals continued to identify English /b d g/ tokens as instances of their Italian /b d g/ categories. As the result of using a single category to process the many instances of /b d g/ they encountered in Italian and English words, the bilinguals' representations for /b d g/ may have gradually evolved to reflect all of the phonetic input they received (Flege, 1991, 1995). That is, their learning of English may have resulted in merged categories for /b d g/ that reflected a two-language source of phonetic input.

VI. GENERAL DISCUSSION

This study examined Italian–English bilinguals' production of English /b/ and their perception of short-lag tokens of English /b d g/ to determine if phonetic learning takes place in the absence of category formation. Experiment 1 showed that /b/ is fully prevoiced far more often in Italian than in English. Experiment 2 examined the production of /b/ by Italian–English bilinguals. The early bilinguals were found to prevoice English /b/ significantly less often than the late bilinguals did, and so resembled the NE speakers to a greater extent than the late bilinguals. However, the early bilinguals nevertheless prevoiced /b/ more often than the NE speakers did. Importantly, both the early and the late bilinguals fully prevoiced English /b/ less often than the Italian monolinguals in experiment 1 fully prevoiced Italian /b/. In experiment 3, the late but not the early Italian bilinguals misidentified /b d g/ tokens as /p t k/ significantly more often than NE speakers did, probably because the short-lag English /b d g/ tokens lacked the pre-voicing that is typically found in Italian /b d g/ (see experiment 1).

The difference between the early and late bilinguals might be attributed to the passing of a critical period (e.g., Bever, 1981; Scovel, 1988). However, in our opinion, the early and the late bilinguals differed primarily as the result of differences in the phonetic input they had received. As in previous studies examining immigrants to North America (e.g., Flege *et al.*, 1999b), the early bilinguals had lived for a longer time in an English-speaking environment, had received more education in English-speaking schools, and were likely to have used English more than the late bilinguals did. As a result, the early bilinguals may well have received more phonetic input from NE speakers over the course of their lives than the late bilinguals had (see Jia and Aaronson, 1999; Stevens, 1999). The early Italian–English bilinguals may have resembled English monolinguals more than the late bilinguals did because they had heard /b/ realized as a short-lag stop (or without full pre-voicing) more often than the late bilinguals had. In support of this, the effect of AOA on the percentage of voicing identification errors for the short-lag English /b d g/ tokens became nonsignificant when length of residence in Canada was used as a covariate.

As discussed in the Introduction, our working assumption was that the Italian–English bilinguals generally did not establish new phonetic categories for English /b d g/ in prestressed, word-initial position. According to the SLM (Flege, 1995), segmental production of an L2 speech sound may change in the absence of category formation. By hypothesis, it will do so through the merger of the phonetic properties of corresponding L1 and L2 sounds.

An analysis of L2 production from the perspective of dynamic systems theory provides another potential account of phonetic change in an L2 in the absence of category formation. Sancier and Fowler (1997) measured a Portuguese–English bilingual's production of Portuguese /p t/ and English /p t/ at several times. This bilingual produced longer VOT values in both English and Portuguese stops while living in the United States than in Brazil. As a result of a ''gestural drift'' towards ambient-language VOT norms, the bilingual's L1 (Portuguese) stops became somewhat less authentic in the United States, whereas her L2 (English) stops became less authentic in Brazil (see also Major, 1992). Sancier and Fowler (1997, p. 433) interpreted this to mean that concomitant changes in L1 and L2 stops arose through an ongoing change in a potential function as phonetic input (especially recent input) was received rather than through the establishment of new attractors.[10]

Experiment 4 provided an indirect test of our working assumption that the English stops /b d g/ are too similar to Italian /b d g/ for category formation to occur. It tested the SLM's prediction that, in the absence of category formation for L2 stops, L1 stops will begin to resemble L2 stops. Experiment 4 revealed that the bilinguals whose productions of English /b/ were most English-like (viz., the early-low participants) also showed the greatest influence of English on their production of Italian /b d g/. A positive correlation was found to exist between the production of stops in English and Italian. The less frequently the participants produced /b d g/ with full prevoicing in English, the less frequently they did so in Italian. This finding, which is analogous to the results of Flege (1987) for voiceless stops,[11] is consistent with the view that the Italian–English bilinguals had not established separate phonetic categories for English /b d g/. If they had done so, there would be no reason to expect their production of L1 stops to change so as to resemble L2 stops. We acknowledge, however, that additional research is needed to further probe for category formation for /b d g/ in both early and late bilinguals.

Additional work will also be needed to provide a better understanding of how L1 categories evolve to accommodate the properties of L2 sounds when category formation does not occur (Flege, 1995). We propose that the internal category structure of the bilingual participants' existing (Italian) /b/ category evolved to encompass the phonetic properties of both Italian and English /b/ tokens in proportion to the input they received (see related discussions by Kluender *et al.*, 1998 and Sancier and Fowler, 1997). More specifically, we propose that progressively less weight (or prominence) was accorded prevoicing in the Italian–English bilinguals' perceptual representations for word-initial tokens of /b d g/ as they gained experience with English. This is because prevoicing does not provide a reliable cue to the identity of phonologically voiced stops in English as it apparently does in Italian (see also Williams, 1977a, for Spanish). If so, one would expect the perceptual weight accorded to low-frequency periodicity just prior to the release burst to decrease as English input was received (see Williams, 1977a). This proposal is consistent with the conclusion drawn by Hazan and Boulakia (1993) regarding the perception of stops by French–English and English–French bilinguals. The bilinguals tended to weight spectral and temporal cues to the voicing feature in stops in a way that was not ''language specific'' (i.e., not just like those of English monolinguals or French monolinguals).

A final comment is in order regarding the role of category formation in L2 speech acquisition. Some investigators (e.g., Kluender *et al.*, 1998) have questioned the need for the construct ''phonetic category'' in speech acquisition research. Researchers in Barcelona have suggested that distinct long-term memory representations may not be established for the sounds encountered in an L2, even under seemingly ideal learning conditions (Sebastián-Gallés and Soto-Faraco, 1999; Bosch *et al.*, 2000). According to the SLM (Flege, 1995), on the other hand, the capacity to form new long-term memory representations (phonetic categories) for L2 speech sounds remains intact across the life span. However, the SLM proposes that the likelihood of category formation will vary according to the state of development of L1 categories at the time of first exposure to the L2, and the degree of perceived dissimilarity of an L2 speech sound from the closest L1 speech sound(s). By hypothesis, whether or not a new category is established for an L2 speech sound will affect how accurately the L2 sound will ultimately be produced and perceived.

The present study focused on L2 speech sounds for which category formation was unlikely, even by early bilinguals (see the Introduction). The findings of this study suggested that phonetic learning did take place for these speech sounds. In our view, the limits on learning observed in this study arose from the influence of previous phonetic learning and the distribution of L1 and L2 phonetic input that was received, not maturational constraints due to normal neurological development (e.g., Scovel, 1988). According to the SLM (Flege, 1995), L1 sounds will exert less influence on the perception and production of an L2 speech sound for which an independent category has been established. As a result, L2 sounds for which a category has been formed will be perceived and produced in a more nativelike fashion than L2 sounds that are processed using a merged category (see, e.g., Flege *et al.*, 1996a, b, 1999a). It is important to note, however, that the present study did not provide direct evidence that the Italian–English bilinguals did not establish new categories for English /b d g/. Additional work will be needed to better define the conditions under which categories are or are not established for L2 speech sounds, as well as the effects of category formation on L2 segmental production and perception.

In summary, this study suggested that phonetic learning for L2 stops takes place in the absence of category formation. Early bilinguals perceived English /b d g/ and produced

English /b/ more accurately than late bilinguals did. Some Italian bilinguals (mostly late bilinguals) continued to misidentify short-lag English /b d g/ tokens as /p t k/ and to fully prevoice /b/ more often than NE speakers did despite having spoken English for several decades. The bilinguals' divergences from English phonetic norms can be attributed to the fact that /b d g/ are fully prevoiced far more often in Italian than English. An analysis of Italian stop production suggested that both the early and late bilinguals' /b d g/ categories reflected experience with corresponding English and Italian stops. We suggest that the early bilinguals approximated English phonetic norms for /b d g/ more closely than the late bilinguals did because they had received more phonetic input from NE speakers, not because they were more likely to have established new phonetic categories for English /b d g/.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Institute for Deafness and Other Communicative Disorders (DC00257). The authors thank L. Gunnin for making VOT measurements, U. Lockridge for editorial assistance, R. Lanni for providing the Italian stimuli used in experiment 4, and K. Aoyama, A. Højen, and three anonymous reviewers for comments on earlier versions of the article.

[1]Sharma and Dorman (2000) found that native Hindi speakers identified and discriminated Hindi syllables with lead and short-lag VOT values (/ba/, /pa/) more accurately than NE speakers did because such differences are phonemic in Hindi but not English. Their analyses revealed that, for both native Hindi and English participants, N1 latencies increased as a function of the duration of the voicing lead in the /ba/ stimuli whereas a robust mismatch negativity (MMN) was observed for the native Hindi but not the NE participants.

[2]English /p t k/ are usually produced with long-lag VOT values, but may be realized with short-lag VOT values in certain contexts (Whalen *et al.*, 1997).

[3]The participants' place of origin in Italy (Abruzzo—24, Calabria—12, Sicilia—8, Veneto—7, Campania—6, Basilicata—4, Lazio—3, Friuli—2, Puglia—2, Lombardia—1, Marche—1, Piemonte—1, Toscana—1) did not vary systematically across the four groups.

[4]The male talker's /b/ tokens were all prevoiced (mean = 107 ms, s.d. = 38), whereas the female talker's /b/ tokens were all produced with short-lag VOT values.

[5]We used nonwords to minimize effects of lexical frequency on the participants' identification responses. Stimuli produced by three NE males were recorded and presented during the experiment. However, we found in a preliminary analysis that, unlike the case for the phonetically trained listeners who had taken part in a pilot experiment, native English-speaking listeners sometimes misidentified /b d g/ tokens in the no-noise condition. Their errors were due almost exclusively to the /d/ token produced by just one of the three talkers, so we decided to examine the responses for stops produced by just two speakers in subsequent analyses.

[6]The S/N estimates were based on the peak intensities of the disyllables and noise segments rather than on rms values. This is because the medial and final consonants contained silent intervals (i.e., the period of supraglottal constriction for /p t k/) whereas the stimuli with initial consonants did not.

[7]The participants were told that the six response alternatives corresponded to pronunciations, not spellings. They were told that /k/ is often spelled with ''c'' at the beginning of words such as cow and with ''ck'' at the end of words such as *tack*.

[8]A mixed design ANOVA examining the percentage of errors that each participant made identifying /p t k b d g/ on the four successive presentations (no-noise followed by the 16, 10, and 4 dB S/N levels) yielded significant main effects of group [$F(4,85)=5.2$; $p<0.01$] and presentation [$F(3,255)=285.2$; $p<0.01$] but a nonsignificant two-way interaction [$F(12,255)=1.4$; $p>0.10$]. Another ANOVA was carried out to determine if greater native versus non-native differences existed for stops presented in the with-noise conditions than in the no-noise condition. It too yielded significant main effects of group [$F(4,85)=6.08$; $p<0.01$] and condition [$F(1,85)=192.3$; $p<0.01$] but a nonsignificant interaction [$F(4,85)=1.11$; $p>0.10$].

[9]There were too few responses for the stops in the no-noise condition to support an analysis. The voicing error scores computed for both /p t k/ and /b d g/ were again based on 3 stops×2 talkers×3 S/N levels=18 judgments.

[10]On this view, the bilingual's sensory experience with a class of English phones such as [p^h] caused her to establish a new potential function. This new function was ''incorporated into...the original potential function'' for Portuguese [p] because it was based on ''far less experience'' than for Portuguese [p] phones. As a result of exposure to ''corresponding'' classes of phones in the L1 and L2 (see also Flege, 1987), the bilingual developed a new intrinsic coordinative dynamic, /p/, that was comprised of two attractors in ''close proximity'' to one another (viz., [p^h] and [p]).

[11]As mentioned in the Introduction, Flege (1987) found that French–English and English–French bilinguals tended to produce /t/ in their L2 with VOT values that were intermediate to the VOT values that are typical for French and English. The bilinguals' production of L1 stops also changed so as to partially resemble those of corresponding L2 stops.

Bever, T. (**1981**). ''Normal acquisition processes explain the critical period for language learning,'' in *Individual Differences in Language Learning Aptitude*, edited by K. Diller (Newbury House, Rowley, MA), pp. 176–198.

Bohn, O.-S., and Flege, J. E. (**1993**). ''Perceptual switching in Spanish/English bilinguals: Evidence for universal factors in stop voicing judgments,'' J. Phonetics **21**, 267–290.

Bosch, L., Costa, A., and Sebastian-Gallés, N. (**2000**). ''First and second language vowel perception in early bilinguals,'' European J. Cognitive Psychology **12**, 189–221.

Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., and Carbone, E. (**1973**). ''The acquisition of a new phonological contrast: The case of stop consonants in French–English bilinguals,'' J. Acoust. Soc. Am. **54**, 421–428.

Cho, T., and Ladefoged, P. (**1999**). ''Variation and universals in VOT: evidence from 18 languages,'' J. Phonetics **27**, 207–229.

Elman, J., Diehl, R., and Buchwald, S. (**1977**). ''Perceptual switching in bilinguals,'' J. Acoust. Soc. Am. **62**, 971–974.

Flege, J. E. (**1987**). ''The production of ''new'' and ''similar'' phones in a foreign language: Evidence for the effect of equivalence classification,'' J. Phonetics **15**, 47–65.

Flege, J. E. (**1991**). ''Age of learning affects the authenticity of voice onset time (VOT) in stop consonants produced in a second language,'' J. Acoust. Soc. Am. **89**, 395–411.

Flege, J. E. (**1995**). ''Second-language speech learning: Theory, findings, and problems,'' in *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, edited by W. Strange (York, Timonium, MD), pp. 233–272.

Flege, J. E., and Eefting, W. (**1986**). ''Linguistic and developmental effects on the production and perception of stop consonants,'' Phonetica **43**, 155–171.

Flege, J. E., and Eefting, W. (**1987**). ''The production and perception of English stops by Spanish speakers of English,'' J. Phonetics **15**, 67–83.

Flege, J. E., and Eefting, W. (**1988**). ''Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation,'' J. Acoust. Soc. Am. **83**, 729–740.

Flege, J. E., and Hammond, R. (**1982**). ''Mimicry of non-distinctive phonetic differences between language varieties,'' Studies in Second Lang. Acquis. **5**, 1–18.

Flege, J. E., and Hillenbrand, J. (**1984**). ''Limits on pronunciation accuracy in adult foreign language speech production,'' J. Acoust. Soc. Am. **76**, 708–721.

Flege, J. E., MacKay, I. R. A., and Meador, D. (**1999a**). ''Native Italian speakers' production and perception of English vowels,'' J. Acoust. Soc. Am. **106**, 2973–2987.

Flege, J. E., McCutcheon, M., and Smith, S. (**1987**). ''The development of skill in producing word-final English stops,'' J. Acoust. Soc. Am. **82**, 433–447.

Flege, J. E., Munro, M. J., and MacKay, I. R. A. (**1995**). ''The effect of age

of second language learning on the production of English consonants,'' Speech Commun. **16**, 1–26.
Flege, J. E., Munro, M., and Skelton, L. (**1992**). ''Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin and Spanish,'' J. Acoust. Soc. Am. **92**, 128–143.
Flege, J. E., Schmidt, A. M., and Wharton, G. (**1996a**). ''Age of learning affects rate-dependent processing of stops in a second language,'' Phonetica **53**, 143–161.
Flege, J. E., Takagi, N., and Mann, V. (**1996b**). ''Lexical familiarity and English-language experience affect Japanese adults' perception of /ɹ/ and /l/,'' J. Acoust. Soc. Am. **99**, 1161–1173.
Flege, J. E., Yeni-Komshian, G. H., and Liu, S. (**1999b**). ''Age constraints on second language learning,'' J. Mem. Lang. **41**, 78–104.
Grosjean, F. (**1999**). ''Studying bilinguals: Methodological and conceptual issues,'' Bilingualism: Lang. Cogn. **1**, 117–130.
Hallé, P. A., Best, C. T., and Levitt, A. (**1999**). ''Phonetic vs phonological influences on French listeners' perception of American English approximants,'' J. Phonetics **27**, 281–306.
Hazan, V., and Boulakia, G. (**1993**). ''Perception and production of a voicing contrast by French-English bilinguals,'' Lang. Speech **36**, 17–38.
Jia, G., and Aaronson, D. (**1999**). ''Age differences in second language acquisition: The dominant language switch and maintainance hypothesis,'' in *Proceedings of the 23rd Annual Boston University Conference on Language Development*, edited by A. Greenhill, H. Littlefield, and C. Tano editors (Cascadilla, Somerville, MA), pp. 301–312.
Keating, P. (**1984**). ''Phonetic and phonological representations of stop consonant voicing,'' Lang. **60**, 286–319.
Kluender, K., Lotto, A., Holt, L., and Bloedel, S. (**1998**). ''Role of experience for language-specific functional mappings of vowel sounds,'' J. Acoust. Soc. Am. **104**, 3568–3582.
Kuhl, P. (**2000**). ''Language, mind, and brain: Experience alters perception,'' in *The New Cognitive Neuroscience*, 2nd ed., edited by M. S. Gazzaniga (MIT, Cambridge, MA), pp. 99–115.
Lisker, L., and Abramson, A. (**1964**). ''A cross-language study of voicing in initial stops: Acoustical measurements,'' Word **20**, 384–422.
MacKay, I. R. A., Meador, D., and Flege, J. E. (**2001**). ''The identification of English consonants by native speakers of Italian,'' Phonetica **58**, 103–125.
Magno-Caldognetto E., Abati, A., and Dossi, L. (**1971**). *Consonanti occlusive sorde e sonore della lingua italiana* (Riccardo Pàtron, Bologna).
Magno-Caldognetto, E., Ferrero, F., Vagges, K., and Bagno, M. (**1979**). ''Indici acustici e indici percettivi nel riconoscimento dei suoni linguistici,'' Acta Phoniatrica Latina **2**, 219–249.
Major, R. (**1992**). ''Losing English as a first language,'' Mod. Lang. J. **76**, 190–208.
Meador, D., Flege, J. E., and MacKay, I. A. R. (**2000**). ''Factors affecting the recognition of words in second language,'' Bilingualism: Lang. and Cogn. **3**, 55–67.
Munro, M. J., Derwing, T. M., and Flege, J. E. (**2000**). ''Canadians in Alabama: A perceptual study of dialect acquisition in adults,'' J. Phonetics **27**, 385–403.
Nathan, G. S. (**1987**). ''On second-language acquisition of voiced stops,'' J. Phonetics **15**, 313–322.
Polivanov, E. (**1931**). ''La perception des sons d'une langue étrangère,'' Travaux du Cercle Linguistique de Prague **4**, 79–96.
Port, R. F., and Mitleb, F. M. (**1983**). ''Segmental features and implementation in acquisition of English by Arabic speakers,'' J. Phonetics **11**, 219–229.
Sancier, M. and Fowler, C. (**1997**). ''Gestural drift in a bilingual speaker of Brazilian Portuguese and English,'' J. Phonetics **25**, 421–437.
Scovel, T. (**1988**). *A Time to Speak. A Psycholinguistic Inquiry into the Critical Period for Human Speech* (Newbury House, Cambridge, MA).
Sebastián-Gallés, N., and Soto-Faraco, S. (**1999**). ''On-line processing of native and non-native phonemic contrasts in early bilinguals,'' Cognition **72**, 111–123.
Sharma, A., and Dorman, M. (**2000**). ''Neurophysiological correlates of cross-language phonetic perception,'' J. Acoust. Soc. Am. **107**, 2697–2703.
Stevens, G. (**1999**). ''Age at immigration and second language proficiency among foreign-born adults,'' Lang. in Soc. **28**, 555–578.
Trubetzkoy, N. S. (**1939/1969**). *Principles of Phonology*, translated by C. A. Baltaxe (Univ. of California, Berkeley, CA).
Werker, J., and Logan, J. (**1985**). ''Cross-language evidence for three factors in speech perception,'' Percept. Psychophys. **37**, 35–44.
Whalen, D. H., Best, C. T., and Irwin, J. R. (**1997**). ''Lexical effects in the perception and production of American English /p/ allophones,'' J. Phonetics **25**, 421–436.
Williams, L. (**1977a**). ''The voicing contrast in Spanish,'' J. Phonetics **5**, 169–184.
Williams, L. (**1977b**). ''The perception of stop consonant voicing by Spanish-English monolinguals,'' Percept. Psychophys. **21**, 289–297.
Williams, L. (**1979**). ''The modification of speech perception and production in second-language learning,'' Percept. Psychophys. **26**, 95–104.
Yeni-Komshian, G. H., Flege, J. E., and Liu, S. (**2000**). ''Pronunciation proficiency in the first and second languages of Korean-English bilinguals,'' Bilingualism: Lang. Cognition **3**, 131–150.