Original Paper

Phonetica 2002;59:49–71

Received: July 1, 1999
Accepted: May 14, 2001

The Production of English Vowels by Fluent Early and Late Italian-English Bilinguals

Thorsten Piske[a] James Emil Flege[b] Ian R.A. MacKay[c]
Diane Meador[d]

[a]Department of English, University of Kiel, Germany; [b]Department of Rehabilitation Sciences, University of Alabama at Birmingham, Ala., USA; [c]Department of Linguistics, University of Ottawa, Ont., Canada; [d]Speech and Language Sciences Research Laboratory, Juneau, Alaska, USA

Abstract

The primary aim of this study was to determine if fluent early bilinguals who are highly experienced in their second language (L2) can produce L2 vowels in a way that is indistinguishable from native speakers' vowels. The subjects were native speakers of Italian who began learning English when they immigrated to Canada as children or adults ('early' vs. 'late' bilinguals). The early bilinguals were subdivided into groups differing in amount of continued L1 use (early-low vs. early-high). In experiment 1, native English-speaking listeners rated 11 English vowels for goodness. As expected, the late bilinguals' vowels received significantly lower ratings than the early bilinguals' vowels did. Some of the early-high subjects' vowels received lower ratings than vowels spoken by a group of native English (NE) speakers, whereas none of the early-low subjects' vowels differed from the NE subjects' vowels. Most of the observed differences between the NE and early-high groups were for vowels spoken in a nonword condition. The results of experiment 2 suggested that some of these errors were due to the influence of orthography.

Introduction

The ability of bilinguals to perceive vowels in a second language (L2) has been examined in several studies [e.g., Flege, 1992a, b; Best et al., 1996; Flege et al., 1997a, 1999]. Other studies have examined bilinguals' productions of L2 vowels [e.g., Major, 1987; Flege, 1992a, b; Busà, 1992, 1995; Munro, 1993; Jun and Cowie, 1994; Munro et al., 1996]. The results of these studies indicate that individuals who began learning their L2 in childhood (often called 'early' bilinguals) are likely to produce and perceive certain L2 vowels more like L2 native speakers than are individuals who began learning their L2 in late adolescence or early adulthood ('late' bilinguals). However, it has not been resolved whether the performance of fluent early bilinguals is indistinguishable from that of monolingual native speakers of the target L2.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0031–8388/02/0591–0049
$18.50/0
Accessible online at:
www.karger.com/journals/pho

Thorsten Piske
Englisches Seminar der Universität Kiel
Olshausenstrasse 40, D–24098 Kiel, Germany
Tel. +49-431-880-2680, Fax +49-431-880-1512
E-Mail piske@anglistik.uni–kiel.de

Four recent studies examined the perception of the Catalan vowels /e/ and /ɛ/ by native speakers of Spanish who began to learn Catalan by school age [Pallier et al., 1997; Pallier et al., 1999; Sebastián-Gallés and Soto-Faraco, 1999; Bosch et al., 2000]. The early Spanish-Catalan bilinguals examined in these studies were described as fluent and frequent speakers of Catalan who had learned their L2, Catalan, in ideal conditions in Barcelona. The studies carried out in Barcelona used different perceptual testing techniques,[1] but all four revealed differences between the early bilinguals and native Catalan speakers. Sebastián-Gallés and Soto-Faraco [1999, p. 120] interpreted their results to mean that 'severe' limitations exist on the 'malleability of the initially acquired L1 phonemic categories, even under conditions of early and extensive exposure'.

More recently, Flege and MacKay [submitted] examined the categorial discrimination of pairs of naturally produced English vowels. The Italian-English bilinguals who participated were assigned to one of four groups based on their age of arrival (AOA) to Canada from Italy, and amount of self-reported use of Italian. Both AOA and amount of L1 use were found to affect the Italian-English bilinguals' discrimination of English vowels. More importantly for the present study, a group of early bilinguals who continued to use Italian often (called 'early-high') differed from a group of age-matched native English (NE) monolinguals in discriminating certain pairs of English vowels. However, early bilinguals who seldom used Italian ('early-low') were never found to differ from the NE monolinguals. Other recent studies have also shown an effect of amount of L1 use on performance in an L2 [Flege et al., 1997b; Guion et al., 1999, 2000; Piske and MacKay, 1999; Meador et al., 2000; MacKay et al., 2001; Piske et al., 2001].

The primary question raised by this study was whether the same pattern of results obtained by Flege and MacKay [submitted] for vowel perception would also hold true for the *production* of L2 vowels. More specifically, we wanted to learn if an early-high group would be found to differ from NE monolinguals in producing English vowels whereas an early-low group would not. Previous studies that have reported an effect of AOA have not controlled for AOA-related variation in amount of L1 use. A secondary question raised by this study, therefore, was whether early and late bilinguals who were roughly matched for amount of L1 use would be found to differ in L2 vowel production accuracy.

An answer to the questions raised here concerning early bilinguals is important theoretically. It is often assumed that the age of first exposure to an L2 (i.e., the early vs. late distinction) is an important determinant of the extent to which bilinguals can maintain a functional separation, or independence, between their two languages. Degree of independence, in turn, is often thought to influence the extent to which the L1 and L2 systems will mutually influence one another [e.g., Lambert and Rawlings, 1969; Paradis, 1978]. According to Grosjean [1997, 1999], early bilinguals' production of L2 vowels might be expected to vary as a function of several factors. These include variation in the strength of the L1 system, the language(s) spoken by the interlocutor(s), and the relative degree of activation of the L1 and L2 systems in a particular speaking context. Given that the L1 system is seldom if ever completely deactivated

[1] Sebastián-Gallés and Soto-Faraco [1999]: a modified version of the gating procedure; Pallier et al. [1999]: the repetition priming paradigm; Pallier et al. [1997]: identification and discrimination of a synthetic /e/-to-/ɛ/ continuum; Bosch et al. [2000]: the 'perceptual magnet' paradigm.

[e.g., Soares and Grosjean, 1984], this approach leads one to expect that at least modest differences will be found to exist between the L2 vowels spoken by early bilinguals and monolingual native speakers of the L2.

A study by Mack [1984] supported the hypothesis that early bilinguals will show modest differences from monolinguals. Mack [1984] examined fluent, English-dominant bilinguals who had learned both of their languages (English and French) prior to the age of 8 years. The early bilinguals differed from NE monolinguals on tests assessing syntactic and semantic aspects of English. They also differed from NE monolinguals in producing and perceiving the vowel /i/. (The differences in production, however, were confined to amount of formant frequency movement at the end of the vowels, not in vowel duration or absolute formant frequency values.) These findings led Mack [1984, p. 173] to conclude that a difference between monolinguals and early bilinguals is an 'inevitable' consequence of bilingualism [see also Grosjean, 1989].

A similar conclusion can be derived from the Speech Learning Model (SLM) [Flege, 1995; see also Paradis, 1978]. According to the SLM, the phonic elements making up the L1 and L2 phonetic systems of a bilingual exist in a 'common phonological space', and so will mutually influence one another. One hypothesis of the SLM is that the formation of new categories for L2 vowels (and consonants) not found in the L1 becomes less likely as the categories making up the L1 phonetic system develop. The SLM posits that even if early bilinguals establish a category for an L2 vowel, they may not produce and perceive it exactly like monolinguals. The SLM proposes that early bilinguals' representations for L2 vowels will 'dissimilate' from those for neighboring L1 vowels to ensure phonetic contrast [see, e.g., Mack, 1989]. The role of subsegmental 'features' in L2 learning is uncertain at present. An inability to modify the feature weights established for L1 vowels might prevent early bilinguals from perceiving [Sebastián-Gallés and Soto-Faraco, 1999] and producing L2 vowels accurately. The SLM proposes that the nonuse of a feature needed to distinguish L2 vowels might block category formation [see Fox et al., 1995; McAllister et al., 1999, submitted]. However, it is not known at present if such a limitation affects early and late bilinguals to the same extent.

Neither of two previous studies that examined the production of English vowels by early Italian-English bilinguals revealed a difference between early bilinguals and NE monolinguals. Munro et al. [1996] had NE-speaking listeners rate vowels spoken by native Italian adults whose AOAs ranged from 2 to 22 years. The ratings obtained from all 10 listeners decreased as the subjects' AOAs increased. This meant that the later the native Italian subjects were first exposed to English, the more their vowels differed from NE monolinguals' vowels. Although the strength of the AOA effect varied across listeners, few listeners judged any vowel spoken by subjects with an AOA less than 10 years to differ from the NE monolinguals' vowels. Importantly, most vowels spoken by early bilinguals received ratings that fell within 2.0 standard deviations (SDs) of the mean rating obtained for NE monolinguals' vowels. This means that the early bilinguals reached the criterion of native-like performance that has been used in several L2 studies [e.g., Flege et al., 1995; Bongaerts et al., 1997; Piske and MacKay, 1999].

Flege et al. [1999] examined English vowels spoken by late bilinguals who used their L1 (Italian) often and two groups of early bilinguals who were matched for AOA (7 years) but differed according to how much they used Italian (early-low = 8%, early-high = 32% of the time). Six of 10 vowels spoken by the late bilinguals (/i ɪ u ʊ o ʌ/)

were correctly identified by NE-speaking listeners significantly less often than vowels spoken by NE monolinguals. However, none of the vowels spoken by either group of early bilinguals differed significantly from the NE monolinguals' vowels.

The lack of significant differences between early Italian-English bilinguals' and NE monolinguals' vowels in the two studies just cited can probably not be attributed to the use of listeners' judgments to assess L2 vowel production accuracy. Flege [1992a] used both acoustic analyses and listeners' judgments to assess the production of /i/-/ɪ/ and /ɛ/-/æ/ by early and late Spanish-English bilinguals. Both groups produced native-like duration differences between the two pairs of vowels. The early but not the late bilinguals produced spectral differences between /i/-/ɪ/ and /ɛ/-/æ/ that closely resembled the spectral differences observed for NE monolinguals. The acoustic measurements agreed with the listeners' judgments of the same vowels in suggesting that the late bilinguals tended to substitute Spanish /i/ for both English /ɪ/ and /i/, and to exaggerate the difference between /ɛ/-/æ/ by substituting two Spanish vowels (the [ɛ] allophone of Spanish /e/ and Spanish /a/) for /ɛ/ and /æ/, respectively.

The subjects examined in experiment 1 of this study were the same subjects whose English vowel production was examined by Flege et al. [1999]. As mentioned, that study did not reveal a difference between NE monolinguals and either of two groups of early bilinguals (early-low, early-high). However, the possibility existed that experiment 1 of the present study would yield different results. First, /ɚ/ was examined in addition to the 10 vowels examined previously (i.e., /i ɪ eˈ ɛ æ u ʊ o ʌ ɒ/). Second, the dependent variable was listeners' ratings of degree of goodness rather than intelligibility (i.e., scores indicating how frequently vowels were heard as intended). Also, experiment 1 of the present study examined the production of vowels in two elicitation conditions. In one, vowels were spoken in words following the auditory presentation of native speaker models. In the second condition, vowels were spoken in nonwords long after the native speaker models had been presented.

The results of experiment 1 revealed the predicted difference between NE monolinguals and subjects in the early-high but not the early-low group. The differences between the early-high and NE groups were largely confined to the 'nonword' condition. Experiment 2 was therefore carried out to provide insight into the nature of the vowel errors that contributed to the observed between-group differences. It did so by examining the NE-speaking listeners' classification of the vowels that had previously been rated in experiment 1.

Experiment 1

This experiment examined the production of 11 Canadian English vowels (/i ɪ eˈ ɛ æ u o ʌ ɒ ʊ ɚ/) in order to test two hypotheses. The first hypothesis was that early Italian-English bilinguals would produce English vowels more accurately than late bilinguals would. The second hypothesis was that early bilinguals who used Italian often (early-high), but not early bilinguals who used Italian seldom (early-low), would differ significantly from NE monolinguals in producing English vowels.

Flege [1995] hypothesized that the more dissimilar an L2 vowel is from the closest L1 vowel, the more likely it is that a category will be established for it. Flege [1995] also hypothesized that category formation is needed for the accurate production of L2 vowels. (The second hypothesis assumes, of course, that the substitution without

modification of the nearest L1 vowel for a particular L2 vowel would be audible to native speakers of the L2 being learned.) These hypotheses, which form part of Flege's [1995] SLM, could be used to generate predictions concerning the relative degree of accuracy with which native speakers of Italian will produce various English vowels. However, that was not possible in the present study because of uncertainty concerning the perceived relation between vowels in the native Italian subjects' L1 systems and those in English at the time they first arrived in Canada. Standard Italian is described as having seven vowels, /i e ɛ a o ɔ u/ [e.g., Agard and DiPietro, 1964]. However, the vowels found in various varieties and dialects of Italian may differ in number and/or phonetic quality [e.g., Romito and Trumper, 1989; Trumper, 1995]. The native Italian speakers who participated in this study came from different provinces in Italy. It is, therefore, likely that their L1 vowels perceptually assimilated Canadian English vowels in a variety of ways. It was not practical to determine the perceived relation between the bilinguals' L1 (Italian) and L2 (English) vowels when they were tested because, after an average of 35 years of residence in Canada, their long-term memory representations for Italian vowels may have changed due to their exposure to the vowels of English [see Flege et al., 1999] or those of other varieties and/or dialects of Italian [see Milani, 1996, p. 480].

Vowel production accuracy was evaluated auditorily by NE-speaking listeners who were drawn from the same community as the subjects who produced the vowels. Several factors led us to use listener judgments rather than acoustic measurements. Differences in gender makeup or vocal tract size might have created spurious between-group differences in formant frequency or fundamental frequency. A single set of formant measurements (at, say, the vowel midpoint) might have generated spurious conclusions if between-group differences existed in the extent and/or the direction of formant movement [see Mack, 1989]. We might have made frequency measurements at multiple measurement locations. It is possible to relate such measurements, along with vowel duration, to the listeners' classification of vowels in their native language [e.g., Hillenbrand et al., 1995]. However, we did not adopt this approach because of the difficulty inherent in relating multiple time-varying acoustic measurements to listeners' normative judgments of vowels as instances of particular vowel categories.

Relatively little is known at present concerning the effect of lexical factors on the production of phonetic segments in an L2. Hammarberg [1993] observed that German learners of Swedish produced certain Swedish vowels more accurately in Swedish words without a German equivalent than in Swedish words that had a German equivalent (or 'cognate'). He suggested that, in cognates, the learner may apply 'already existing knowledge rather than observing the phonetic input' when producing known words (p. 455). On the other hand, Flege et al. [1998] found that late Spanish/English bilinguals generally produced word-initial English /t/ tokens with shorter VOT values than NE speakers did. The VOT values did not vary systematically as a function of the subjective familiarity or the text frequency of the words containing the measured /t/ tokens. Nor did the VOT values depend on whether the English words did or did not have a cognate in Spanish. We assessed the effect of lexical status in the present experiment by examining the accuracy with which vowels were produced in familiar English words and in nonwords.

Table 1. Mean characteristics of the four groups of subjects

	Gender	Age	AOA	LOR	% Italian
NE	9 m, 9 f	48 (7)	–	–	–
Early-low	9 m, 9 f	48 (5)	7 (3)	40 (5)	8 (6)
Early-high	8 m, 10 f	47 (6)	7 (2)	40 (6)	32 (16)
Late-high	8 m, 10 f	48 (6)	19 (1)	28 (5)	41 (23)

Age = Chronological age, in years; AOA = subjects' age of arrival, in years; LOR = subjects' length of residence in Canada, in years; % Italian = subjects' self-estimated percentage use of Italian. SDs are in parentheses.

Method

Subjects

The mean age of the 64 subjects who participated in this experiment was 48 years. All of the subjects passed a pure-tone hearing screening prior to participating; none reported a history of auditory disorder. As described previously in detail [Flege et al., 1999; MacKay et al., 2001], three groups of Italian-English bilinguals and a group of NE speakers (18 per group) were recruited in Ottawa, Ontario.[2] The native Italian subjects had been living in Canada for a minimum of 18 years at the time of testing (mean = 35 years). As summarized in table 1, the subjects in the 'late' group had arrived in Canada later in life (mean = 19 years) than the two groups of early bilinguals had (mean = 7 years for both). The early bilinguals in the 'early-low' group reported using Italian much less often (mean = 8%) than the early bilinguals in the 'early-high' group did (mean = 32%). An ANOVA examining the native Italian subjects' self estimates of percentage L1 use was significant $F(2, 51) = 18.4$, $p < 0.01$. A Tukey's test revealed that, as intended by the design, the early-high and late-high groups used Italian more than the early-low group ($p < 0.01$), whereas the early-high and late-high groups did not differ significantly ($p > 0.05$).

Procedure

The subjects were tested one at a time in a small room located in a predominantly Italian Roman Catholic parish church in Ottawa. The subjects repeated English and Italian sentences prior to producing the speech materials, i.e., words and nonwords, examined here. (The English sentences were rated for degree of foreign accent in a study by Meador et al. [2000]; the Italian sentences have not yet been examined.) The words and nonwords contained 11 vowels, i.e. /i ɪ eˈ ɛ æ u o ʌ ɒ ʊ ɚ/. To elicit word production, a native speaker of English recorded the four-word sequences shown in table 2. His productions were digitized and then normalized for peak intensity. The subjects were provided with a written list of the sequences. They simply repeated each word a single time after hearing the four words in each sequence via a loudspeaker. When the same four words were presented a second time, the subjects inserted the vowel common to all four words (e.g., /i/ in *read, deed, heed, bead*) into a /b_do/ frame, creating a nonword. After the subjects produced a nonword (e.g., /bido/ in the example given), a carrier phrase containing an internal pause (*I say . . . again and again*) was presented via the loudspeaker. The subjects were told to insert the nonword into the carrier phrase (saying, e.g., *I say* /*bido*/ *again and again*). This meant that a relatively long speech-filled interval occurred between

[2] Vowels spoken by 18 native speakers of Italian with a mean AOA of 14 years were also elicited and evaluated by the listeners. The mean ratings obtained for these subjects' vowels were generally intermediate to those obtained for the early and late bilinguals' vowels. In the interest of economy, these results will not be reported here.

Table 2. The four English words repeated by the subjects in experiment 1

Target vowel	Test words			
/i/	read	deed	heed	bead
/ɪ/	rid	did	hid	bid
/eɪ/	raid	shade	made	bade
/ɛ/	red	dead	said	bed
/æ/	mad	dad	had	bad
/u/	rude	food	sued	booed
/o/	road	code	hoed	bode
/ʌ/	cud	mud	dud	bud
/ɒ/	rod	sod	cod	god
/ʊ/	good	could	would	hood
/ɚ/	third	word	heard	bird

The final word in each sequence was examined in the 'word' condition (see text).

the nonwords that the subjects produced in the carrier phrase and the native speaker's model of the vowels in the nonwords.

The final word in the four-word sequences (i.e., *bead, bid, bade, bed, bad, booed, bode, bud, god, hood, bird*) and the nonwords produced in the carrier phrases were later digitized (22.05 kHz) and normalized for peak intensity. The final /o/ in the nonwords was removed by deleting all portions of the signal following the complete constriction of the postvocalic /d/. ('Complete constriction' was defined on the basis of a drop in amplitude and/or a change in the shape of the waveform.) This procedure yielded 22 /CVd/ syllables per subject (11 vowels × 2 conditions).

As mentioned earlier, the possibility existed that lexical factors might influence how accurately the native Italian subjects produced English vowels. The possibility also existed that lexical factors might influence the judgments of the NE-speaking listeners who were called upon to judge vowel production accuracy. The /CVd/ stimuli just described differed in lexical status. For example, the /bɚd/ derived from /bɚdo/ is an English word ('bird') but not the /bʊd/ derived from /bʊdo/. The stimuli were also likely to have differed in subjective familiarity to the NE-speaking listeners (compare, for example, *bed* vs. *bade*). Moreover, there was some variation in the initial consonants. All of the nonwords and nine of the 11 real words began with /b/, but one real word began with /h/ (hood), and one began with /g/ (god).

The stimuli were edited to minimize the possibility that lexical factors would influence the NE-speaking listeners' judgments of vowel production accuracy. The aim of the editing was to obscure the identity of the initial consonant (and thus lexical identity) while leaving cues to vowel identity and goodness intact insofar as possible. This was done by applying a weighting function to an interval ranging from 30 to 60 ms at the beginning of each stimulus. The duration of the weighting function that was used varied as a function of stimulus duration. It attenuated the signal to zero over the first one half of the interval, and from 0 to 100% of amplitude over the second half of the selected interval.[3]

Auditory Evaluation

The 1,584 vowel stimuli (4 groups × 18 subjects × 11 vowels × 2 conditions) were presented to 11 native speakers of English (5 male, 6 female) with a mean age of 32 years. All of the listeners had

[3] Pilot work indicated that the shortest interval yielding unidentifiable initial consonants varied as a function of word duration, apparently as the result of rate-dependent variation in the duration of formant transitions. The weighting function was therefore applied over just a 30-ms interval for stimuli that were shorter than 140 ms (29 words, 276 nonwords), over a 46-ms interval for stimuli that were 140–300 ms in duration (737 words, 686 nonwords), and over a 60-ms interval for the stimuli that were longer than 300 ms (224 words, 28 nonwords). Any prevoicing in the /b/ and /d/ tokens, as well as the aspiration noise in /h/, were removed before the weighting function was applied.

been born and raised in the Ottawa region. Some of them knew French, but none was proficient in French or in any language other than English. The listeners were tested one at a time at the Phonetics Laboratory of the University of Ottawa after passing a pure-tone hearing screening.

The vowels spoken in words and nonwords were presented via a loudspeaker in separate, counterbalanced blocks using a notebook computer. The order of presentation of the 11 target vowels was counterbalanced across listeners within each block. The listeners were told the identity of each target vowel before the 72 tokens of it were presented in the word and nonword conditions. The listeners used a scale that ranged from 1 ('very strong foreign accent') to 5 ('no foreign accent') to rate each token. The listeners were told to click a sixth button that was marked 'wrong' if they judged the vowel in a stimulus to be an instance of some other vowel category.

The listeners received no training on the judgment task. However, three extra stimuli were presented for practice at the beginning of each set. Responses to these vowels were not analyzed. The listeners were required to judge the vowel in each stimulus, and were told to make their best guess if uncertain. The stimuli could be replayed, but a response could not be changed once given. The interval between each response and presentation of the next stimulus was 1 s.

Analyses

Twenty-two 'talker-based' ratings (11 vowels × 2 conditions) were calculated for each subject by averaging over the judgments of each vowel token that were given by the 11 NE-speaking listeners. In computing these ratings, any judgements of a vowel token as the 'wrong vowel' were assigned a value of 0. This procedure might have underestimated the magnitude of vowel production errors in instances where a target vowel was replaced by a vowel that was distant in the vowel space from the target vowel. Such an underestimation, if it occurred, would be more likely for late than early bilinguals, and so might reduce the expected effect of AOA on the native Italian subjects' production of English vowels.

Averaging across listeners was deemed appropriate because preliminary analyses had revealed that the listeners evaluated the vowel stimuli in a similar way. Despite an acceptably high level of interrater agreement,[4] there was reason for caution in accepting the results of analyses based solely on ratings averaged across listeners. Flege et al. [1995] examined NE-speaking listeners' ratings of English sentences that had been produced by NE speakers and Italian-English bilinguals differing in AOA to Canada. All of the listeners gave significantly lower ratings to sentences spoken by late bilinguals than by NE speakers, which indicated that the late bilinguals spoke English with detectable foreign accents. However, only some of the listeners gave significantly lower ratings to sentences spoken by early bilinguals than by NE speakers [see also Munro et al., 1996]. Flege [1998] reported that a strong correlation existed between the accuracy with which Italian-English bilinguals produced English vowels and their overall degree of foreign accent in English sentences. The possibility therefore existed that some but not all of the NE-speaking listeners who participated in this experiment may have detected a difference between vowels produced by the early Italian-English bilinguals and the NE group.

The foregoing considerations led us to test for differences in 'listener-based' ratings in addition to testing for between-group differences in the 'talker-based' ratings described earlier. A total of 968 'listener-based' ratings (11 listeners × 4 groups × 11 vowels × 2 conditions) were computed by averaging over the judgments given by each listener to the productions of each target vowel by the 18 subjects in a group. For example, 11 mean ratings (one for each listener) were computed for the early-high group's productions of /i/ in words, and 11 mean ratings (one per listener) were computed for the late group's productions of /ɛ/ in nonwords.

[4] ANOVAs indicated that the 11 listeners did not differ significantly in judging the vowels in either words [$F(10, 9890) = 354.5$, $p > 0.10$] or nonwords [$F(10, 9890) = 190.8$, $p > 0.10$]. Intraclass correlations were acceptably high in both words ($R = 0.86$) and nonwords ($R = 0.92$, $p < 0.0001$ in both instances). The same held true when mean scores obtained for each group (11 vowels × 4 groups = 44 scores per listener) were examined (words: $R = 0.89$, nonwords, $R = 0.93$; $p < 0.0001$).

Table 3. The mean ratings obtained in experiment 1 for 11 English vowels spoken by the subjects in four groups in two conditions

	Condition	Group			
		NE	EL	EH	LH
/i/	W	3.6 (0.5)	3.6 (0.4)	3.5 (0.5)	3.1 (1.0)
	NW	3.7 (0.4)	3.8 (0.6)	3.4 (1.0)	3.3 (1.1)
/ɪ/	W	3.9 (0.6)	3.5 (0.6)	3.5 (0.4)	2.1 (1.2)
	NW	3.9 (0.5)	2.8 (1.7)	1.7 (1.5)	1.9 (1.6)
/eᴵ/	W	3.6 (0.7)	3.7 (0.6)	3.8 (0.5)	2.6 (1.2)
	NW	3.6 (0.6)	3.5 (1.2)	3.3 (1.1)	2.7 (1.3)
/ɛ/	W	3.6 (0.7)	3.8 (0.5)	3.3 (0.9)	2.6 (0.8)
	NW	3.9 (0.5)	3.4 (1.3)	2.7 (1.5)	2.6 (1.3)
/æ/	W	3.7 (0.6)	3.5 (0.6)	3.4 (0.8)	2.6 (0.9)
	NW	3.6 (0.7)	3.5 (0.8)	3.0 (1.2)	3.2 (1.2)
/ɒ/	W	3.3 (0.6)	3.5 (0.5)	3.2 (0.9)	2.5 (1.0)
	NW	3.8 (0.6)	3.4 (1.0)	2.7 (1.6)	2.2 (1.1)
/ɚ/	W	4.2 (0.4)	4.0 (0.4)	3.6 (0.5)	2.1 (1.1)
	NW	3.6 (1.6)	2.5 (1.8)	1.9 (1.8)	1.4 (1.4)
/ʌ/	W	3.7 (0.6)	3.5 (0.7)	3.3 (0.6)	1.7 (1.4)
	NW	4.0 (0.6)	3.2 (1.6)	2.9 (1.9)	1.7 (1.5)
/o/	W	4.1 (0.5)	3.6 (0.6)	3.6 (0.8)	1.9 (1.3)
	NW	3.6 (0.7)	3.5 (1.2)	3.2 (1.1)	2.3 (1.4)
/ʊ/	W	4.1 (0.5)	3.5 (0.8)	3.3 (0.9)	1.7 (1.3)
	NW	3.6 (1.2)	2.7 (1.4)	1.7 (1.2)	1.6 (1.1)
/u/	W	4.1 (0.4)	3.6 (0.7)	3.5 (0.7)	2.2 (1.2)
	NW	3.8 (0.5)	3.4 (1.0)	2.9 (1.1)	2.0 (1.1)

W = Word condition; NW = nonword condition; EL = early-low; EH = early-high; LH = late-high. Each mean is based on the average rating assigned by 11 listeners to each of 18 subjects per group. SDs are in parentheses.

Results

The mean talker-based ratings are presented in table 3. The listener-based scores yielded the same group mean values but were, of course, associated with different SDs. Averaged over the 11 vowels and two conditions, the ratings obtained for the four groups were as follows: NE = 3.76, early-low = 3.43, early-high = 3.07, late-high = 2.26. Averaged over the four groups and two elicitation conditions (word vs. nonword), ratings for the 11 vowels ranged from a low of 2.77 for /ʊ/, to a high of 3.50 for /i/. The ratings were higher on average for vowels spoken in words (mean = 3.29) than in nonwords (mean = 2.97). However, the size of the word versus nonword difference was greater for the two groups of early bilinguals (early-low: 3.62 vs. 3.25, early-high group: 3.46 vs. 2.68) than for the NE (3.80 vs. 3.73) or the late-high group (2.27 vs. 2.24).

The talker-based and listener-based ratings were examined in separate (4) Group × (2) Condition ANOVAs. The aim of these analyses was to determine if subjects in the four groups differed in producing any of the English vowels and, if so, whether the pattern of between-group differences was the same for vowels spoken in words and nonwords. We reasoned that a between-group difference in talker-based ratings might

Table 4. Summary of F tests obtained in two-way ANOVAs examining the 'talker-based' and 'listener-based' scores obtained in experiment 1 (see text)

	Analysis	Factors		
		Group	Condition	G × C
/i/	talker	2.9*	0.6	0.4
	listener	0.7	0.7	0.5
/ɪ/	talker	17.2*	16.2*	5.8*
	listener	28.4*	39.6*	14.1*
/eɪ/	talker	7.4*	1.4	1.1
	listener	6.2*	1.3	1.0
/ɛ/	talker	9.3*	1.4	1.9
	listener	7.5*	3.6	4.8*
/æ/	talker	3.6*	0.0	3.1*
	listener	3.6*	0.0	2.5
/ɒ/	talker	9.7*	0.5	2.4
	listener	10.3*	1.6	7.4*
/ɚ/	talker	17.8*	29.5*	1.8
	listener	36.1*	101.6*	6.3*
/ʌ/	talker	19.7*	0.2	0.6
	listener	31.5*	1.0	2.4
/o/	talker	16.2*	1.5	2.0
	listener	24.7*	2.8	3.7*
/ʊ/	talker	25.1*	17.0*	3.1*
	listener	48.7*	40.4*	7.3*
/u/	talker	20.6*	8.3*	0.8
	listener	14.4*	20.0*	2.0

An asterisk indicates significance at the 0.05 level.

generalize to other groups of Italian-English bilinguals having characteristics similar to those examined here. However, the differences in talker-based ratings might not generalize to another panel of NE-speaking listeners. Conversely, between-group differences in listener-based ratings might generalize to another panel of NE listeners but not to other groups of Italian-English bilinguals. We therefore considered a between-group difference to be significant only if it was significant in the analyses of both the talker-based and listener-based ratings.

The results of the ANOVAs are summarized in table 4. A significant ($p < 0.05$) effect of Group was obtained for all but one of the 11 vowels that were examined. The effect for /i/ was considered nonsignificant because the main effect of Group was significant in the analysis of talker-based but not listener-based ratings (see above). For four vowels, /ɪ ɚ ʊ u/, significantly ($p < 0.05$) higher ratings – indicating more accurate production – were obtained for words than nonwords (/ɪ/: 4.26 vs. 3.58; /ɚ/: 4.46 vs. 3.36; /ʊ/: 4.13 vs. 3.40; /u/: 4.32 vs. 4.00). Finally, a significant Group × Condition interaction was obtained for two vowels, /ɪ/ and /ʊ/ ($p < 0.05$).

A series of five a posteriori t tests was carried out to test for between-group differences in the production of each vowel. The first test compared the ratings obtained for vowels spoken by the early-high and late-high groups, and the second test compared the early-low and early-high groups. The remaining tests compared the three native Italian groups to the NE group. In cases where the ANOVA examining a vowel yielded

Table 5. Summary of a posteriori comparisons (t tests) carried out to test for between-group differences in experiment 1

Vowel	Comparison				
	EH>LH	EL>EH	NE>EL	NE>EH	NE>LH
/i/	–	–	–	–	–
/ɪ/	W	–	–	NW	W, NW
/eɪ/	W, NW	–	–	–	W, NW
/ɛ/	–	–	–	NW	W, NW
/æ/	–	–	–	–	W
/ɒ/	–	–	–	–	NW
/ɚ/	W, NW	–	–	NW	W, NW
/ʌ/	W, NW	–	–	W, NW	W, NW
/o/	W, NW	–	–	–	W, NW
/ʊ/	W	–	–	NW	W, NW
/u/	W, NW	–	–	–	W, NW

EL = Early-low; EH = early-high; LH = late-high; W = vowels spoken in the word condition differed significantly (Bonferroni $p < 0.05$) in both the talker and listener analyses; NW = a significant difference for vowels in the nonword condition.

a nonsignificant Group × Condition interaction, the a posteriori tests examined the average of ratings obtained for that vowel in words and nonwords. A pairwise between-group difference was considered significant only if a difference was observed (Bonferroni $p < 0.05$) in analyses examining both the listener-based and talker-based ratings.

The results of the a posteriori tests are summarized in table 5. These tests revealed that AOA affected L2 vowel production accuracy. Significantly lower ratings were obtained for the late-high than the early-high group for seven vowels, /ɪ eɪ ɚ ʌ o ʊ u/. (The late-high vs. early-high differences for /ɪ/ and /ʊ/ were confined to words.) Given that the two groups' amount of L1 use was comparable, these results indicate that an early exposure to the L2 promotes a relatively accurate production of L2 vowels.

In no case did the differences between the early-low and early-high groups reach significance. The lack of significant differences between these groups would seem to contradict the results of other recent studies examining the effect of L1 use on Italian-English bilinguals' performance in English [Flege et al., 1997b; Piske and MacKay, 1999; Meador et al., 2000; Piske et al., 2001; Flege and MacKay, submitted]. However, the remaining a posteriori tests indirectly suggested an effect of L1 use by showing more differences between the early-high and NE groups than between the early-low and NE groups. Significantly lower ratings were obtained for five vowels (/ɪ ɛ ɚ ʌ ʊ/) spoken by the early-high than NE group, whereas no significant differences between the early-low and NE groups were observed. (The differences between the early-high and NE groups for /ɪ ɛ ɚ ʊ/ were confined to vowels in nonwords.) The lack of differences between the early-low and NE groups cannot be attributed to a lack of sensitivity of the rating procedure. This is because all of the late-high group's vowels except /i/ received significantly lower ratings than the vowels produced by the NE group. (The difference between the late-high and NE groups for /æ/ was confined to words, and the /ɒ/ difference was confined to nonwords.)

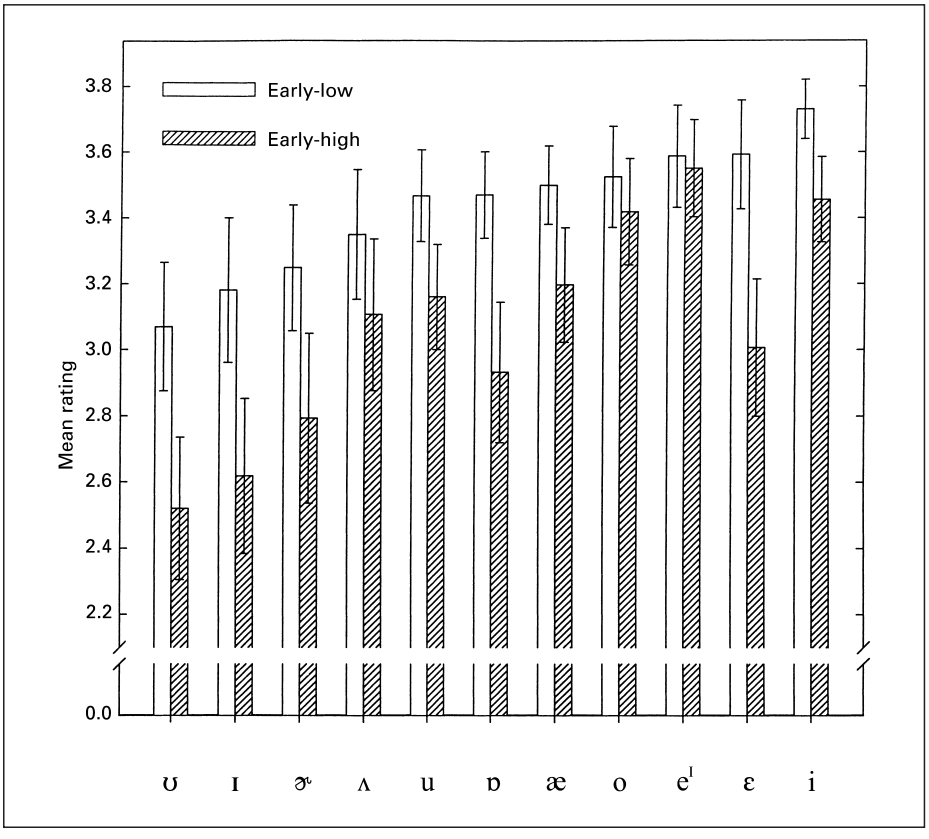

Fig. 1. The mean ratings obtained for 11 English vowels spoken by early Italian-English bilinguals who seldom ('early-low') or relatively often ('early-high') used Italian. The error bars enclose ± 2.0 SE for 'talker-based' means (see text).

Given the indirect evidence of an L1 use effect just presented, we decided to directly compare the early-low and early-high groups. Figure 1 shows the mean ratings obtained for vowels spoken by these two groups. (The error bars in the figure are based on means of the talker-based ratings.) The ratings for all 11 vowels were higher for the early-low than the early-high group.

The talker-based ratings obtained for the early-low and early-high groups were submitted to a (2) Group × (2) Condition × (11) Vowel ANOVA. Significantly higher ratings were obtained for the early-low than the early-high group [$F(1, 34) = 8.7$, $p < 0.01$]. The Group factor did not interact significantly with Vowel [$F(10, 340) = 0.7$, $p > 0.10$] or Condition [$F(1, 34) = 2.8$, $p > 0.10$], or enter into a significant three-way interaction [$F(10, 340) = 0.5$, $p > 0.10$].

These results suggested that the subjects in the early-low group produced all 11 English vowels more accurately than the subjects in the early-high group did. However, the results obtained in an analysis of the listener-based ratings suggested that a reliable between-group difference was confined to just five vowels. This three-way ANOVA

yielded a nonsignificant main effect of Group [$F(1, 20) = 2.4$, $p > 0.10$], a nonsignificant Group × Condition interaction [$F(1, 20) = 3.9$, $p > 0.05$] and a nonsignificant three-way interaction [$F(10, 200) = 1.6$, $p > 0.10$]. However, a significant Group × Vowel interaction was obtained [$F(10, 200) = 2.5$, $p < 0.01$]. Tests of simple main effects revealed that the early-low subjects produced /ɪ ɛ ɒ ɚ ʊ/ more accurately than the early-high subjects did ($p < 0.05$), but did not differ significantly from the early-high subjects for /i eᴵ æ ʌ o u/.

Discussion

The results obtained here agreed with the findings of Munro et al. [1996] in showing that early bilinguals produced English vowels more accurately than late bilinguals. The early-high subjects produced seven of the 11 vowels examined here (/ɪ eᴵ ɚ ʌ o ʊ u/) more accurately than the late-high subjects did. However, this is apparently the first study that has compared groups of early and late bilinguals who were roughly matched for amount of L1 use. AOA and amount of L1 use are typically confounded in studies examining immigrants to North America [see, e.g., Flege et al., 1995]. Thus the difference between subjects in the early-high and late-high groups was probably due to some factor associated with the subjects' age at the time of first extensive exposure to English rather than to a confounded difference in amount of L1 use.

Our results confirmed the hypothesis that L2 vowel production accuracy is influenced by amount of L1 use. The early-low subjects produced five of the 11 vowels examined here (i.e., /ɪ ɛ ɒ ɚ ʊ/) more accurately than the early-high subjects did. Also, the early-high group produced five vowels (/ɪ ɛ ɚ ʌ ʊ/) less accurately than the NE subjects did whereas the early-low group did not differ significantly from the NE subjects for any vowel.

These findings agree in part with recent studies examining overall degree of foreign accent in English sentences. These studies [Flege et al., 1997b; Piske and MacKay, 1999; Piske et al., 2001] showed that early bilinguals who used Italian seldom had milder foreign accents in English than early bilinguals who used Italian often did. In the foreign accent studies, sentences spoken by subjects in both early-high and early-low groups received significantly lower ratings than sentences spoken by NE speakers did. However, in the present study, only vowels produced by subjects in the early-high group received significantly lower ratings than vowels spoken by NE speakers did. The difference can probably be attributed to the length of the speech materials that were examined (/CVd/ syllables vs. whole sentences). That is, the presence of a detectable foreign accent in sentences but not individual vowels spoken by the early-low subjects might have arisen from the larger number of vowels and consonants in the sentences, as well as from prosodic dimensions.

The results obtained here for one group of early bilinguals (early-high) differ from the results obtained in two previous studies. Flege [1992a] did not observe vowel production differences between NE speakers and early Spanish-English bilinguals. It is important to note, however, that the Spanish-English bilinguals' use of the L1 (Spanish) was not assessed. Perhaps they did not differ from NE speakers because they used their L1 (Spanish) as infrequently as the early-low group examined in this study used their L1 (Italian). Flege et al. [1999] did not observe differences between the early-high and NE groups, whereas we observed differences between these same groups

for /ɪ ɛ ɚ ʌ ʊ/ in the present study. The difference between the studies was probably methodological. The listeners in the Flege et al. [1999] study classified English vowels using keywords, whereas the listeners in the present study rated vowels for goodness, allowing them to respond to within-category differences in L2 vowel production accuracy.

The differences observed here between the early-high and NE subjects' productions of /ɪ/, /ʊ/ and /ɚ/ are notable. As far as we know, none of these vowels has a phonetic counterpart in any variety or dialect of Italian. These findings might, therefore, be regarded as disconfirming the hypothesis that early bilinguals establish phonetic categories for certain 'new' L2 vowels and, as a result, produce such vowels in a native-like way [see Flege, 1992a, for discussion].[5] It is important to note, however, that the early-high versus NE differences for /ɪ ʊ ɚ/ (and also /ɛ/, which does have a counterpart in Italian) were confined to nonwords. The current version of the SLM [Flege, 1995] has abandoned the notion of 'new' as a discrete classification for L2 vowels, in favor of the hypothesis that the likelihood of category formation for L2 vowels varies as a continuous function of their perceived dissimilarity from the closest L1 vowel. Perhaps the hypothesis regarding 'new' vowels [Flege, 1988, 1992a, b] holds true only for the production of vowels in familiar words, or for vowels that exceed some as yet undetermined 'threshold' of dissimilarity. Alternatively, the nonword production task might have stressed the speech production system in a way that revealed underlying differences between native and nonnative speakers more clearly.

Experiment 2

The aim of this experiment was to provide additional insight into native versus nonnative differences that were observed in experiment 1. We did this by considering an aspect of the scaling procedure that was used in experiment 1, and by seeking to determine if orthography may have led to some of the vowel production errors that were observed. Both analyses were based on NE-speaking listeners' classifications of the same vowels that were rated in experiment 1.

The data presented here have been drawn from an experiment reported by Flege et al. [1999], where 6 of the NE listeners who had rated vowels in experiment 1 later classified the same vowels using keywords. Flege et al. [1999] tested for between-group differences in vowel 'intelligibility', that is, the percentage of times that each vowel token was identified as an instance of its intended category. Here we focused on the *nature* of the errors the native Italian subjects made when producing English vowels by tabulating how their productions were classified. Given our interest in the accuracy with which the early bilinguals produced English vowels, just the results obtained for three groups (NE, early-low, early-high) will be presented.

In experiment 1, significantly lower ratings were obtained for productions of /ɪ ɛ ɚ ʌ ʊ/ by the early-high than the NE group, whereas no significant differences were observed between the early-low and NE groups. However, it is important to recall that the procedure used in experiment 1 diverged from conventional scaling techniques in

[5] Although the early-high and NE subjects differed for /ʌ/, this vowel might not be considered 'new' because some varieties of Italian apparently possess an /ʌ/-quality vowel [see Trumper, 1995].

an important way. Therefore, one aim of experiment 2 was to consider data that might cause us to reevaluate our conclusion that the early-low subjects did not differ from the NE subjects in producing English vowels.

When an equal-appearing interval scale is used all of the stimuli being evaluated usually differ along a shared dimension. It is this shared dimension that is rated. However, a nonnative speaker's productions of an L2 vowel may diverge from the phonetic norms of the L2 in two distinct ways: categorical identity and degree of goodness of fit. With this in mind, the decision was made in experiment 1 to present productions of each target vowel in separate blocks. The identity of the target vowel being judged in each block was known beforehand to the NE listeners. They were told to rate each vowel token as an instance of its intended category using a scale that ranged from 1 ('very strong foreign accent') to 5 ('no foreign accent'). They were told to click a sixth button marked 'wrong' if they heard some other vowel.

The 'wrong vowel' judgments were assigned a value of 0 when mean ratings were computed in experiment 1. However, the perception of a categorical difference between two vowels augments the degree of perceived dissimilarity between them [Flege et al., 1994]. It is logically possible that the early-low subjects produced more vowels that were not identified as an instance of the intended (target) vowel category than the early-high subjects did. If so, then we may have underestimated the magnitude of their vowel production errors. Were this to hold true, it would not be reasonable to maintain that the early-high subjects but not the early-low subjects differed from the NE subjects in producing English vowels.

Another aim of experiment 2 was to evaluate the role of orthography in L2 vowel production. Experiment 1 yielded more native versus nonnative vowel production differences in nonwords than in words. One possible explanation for the effect of lexical status is that some vowel errors in nonwords were due to the Italian-English bilinguals' reliance on orthography.

As described in experiment 1, the subjects produced target vowels in nonwords after repeating a sequence of English words that were presented aurally and also visually, via a written list. A long delay occurred between the aural models of an English vowel and its attempted production in a nonword. As a result, subjects who did not have a robust representation for an English vowel may not have been able to produce it in a /b_do/ context. In such an instance, they may have resorted to producing the English vowel as the letter representing it is pronounced in Italian. For example, English /ɪ/ might have been produced as an [i]-quality vowel in /b_do/ nonwords because the real English words containing /ɪ/ were all spelled with 'i' *(rid, did, hid, bid)*, which is pronounced /i/ in Italian. For such a conclusion to be accepted, it would be necessary to observe the presence of 'spelling' errors in the nonword condition, where an aural model was not heard recently, but not in the word condition, where English vowels were produced soon after the presentation of an aural model.

Method

As described by Flege et al. [1999], 6 NE-speaking listeners used keywords to identify all of the vowels examined in experiment 1 except /ɚ/. Vowels spoken in words and nonwords were presented to the listeners in counterbalanced blocks. Within these two blocks, 2 subsets of vowels were presented (also in counterbalanced order) to reduce the number of keywords that were needed to unambiguously classify the vowels. *Heed, hid, hayed, head, had, hot* and *hut* were the keywords offered as responses

Table 6. Native English listeners' classifications of 10 English vowels spoken by the subjects in three groups (percentages)

Group	Vowel	Word condition	Nonword condition
NE	/i/	i(99)	i(95) ɪ(4)
EL	/i/	i(100)	i(87) ɪ(13)
EH	/i/	i(95) eᶦ(3)	i(88) ɪ(6) ɒ(3)
NE	/ɪ/	ɪ(95) ɛ(5)	ɪ(95) ɛ(4)
EL	/ɪ/	ɪ(92) ɛ(7)	ɪ(67) i(31)
EH	/ɪ/	ɪ(97) ɛ(3)	ɪ(35) i(57)
NE	/eᶦ/	eᶦ(95) i(5)	eᶦ(91) i(4) ɛ(3) ɪ(3)
EL	/eᶦ/	eᶦ(100)	eᶦ(81) i(10) æ(5) ɪ(4)
EH	/eᶦ/	eᶦ(99)	eᶦ(85) i(6) ɛ(5) ɪ(4)
NE	/ɛ/	ɛ(91) æ(6)	ɛ(92) ɪ(6)
EL	/ɛ/	ɛ(97)	ɛ(71) eᶦ(16) ɪ(8)
EH	/ɛ/	ɛ(86) æ(8) ɪ(4)	ɛ(55) eᶦ(31) ɪ(11)
NE	/æ/	æ(84) ɛ(13)	æ(73) ɛ(15) ʌ(9) ɒ(3)
EL	/æ/	æ(99)	æ(83) ɛ(6) ʌ(6) ɒ(6)
EH	/æ/	æ(89) ɛ(8)	æ(68) ɒ(13) ɛ(10) ʌ(4) eᶦ(3)
NE	/ɒ/	ɒ(94) ʌ(6)	ɒ(85) ʌ(15)
EL	/ɒ/	ɒ(94) ʌ(6)	ɒ(82) ʌ(9) o(5) ʊ(3)
EH	/ɒ/	ɒ(83) ʌ(16)	ɒ(60) o(20) ʌ(13) æ(3)
NE	/ʌ/	ʌ(86) ɒ(7) ʊ(6)	ʌ(89) ʊ(7) ɒ(4)
EL	/ʌ/	ʌ(79) ʊ(13) ɒ(8)	ʌ(61) ʊ(19) u(12) ɒ(6)
EH	/ʌ/	ʌ(80) ʊ(16) ɒ(4)	ʌ(51) ɒ(19) ʊ(14) u(10) o(6)
NE	/o/	o(99)	o(90) ʊ(4) ʌ(4)
EL	/o/	o(100)	o(91) ʊ(6)
EH	/o/	o(100)	o(81) ɒ(7) ʊ(6) ʌ(4)
NE	/ʊ/	ʊ(97) ʌ(3)	ʊ(74) u(15) ʌ(7) o(3)
EL	/ʊ/	ʊ(92) ʌ(6)	ʊ(45) u(43) o(6) ʌ(4)
EH	/ʊ/	ʊ(94) ʌ(5)	o(34) ʊ(31) u(30) ɒ(3)
NE	/u/	u(94) ʊ(4)	u(88) ʊ(12)
EL	/u/	u(97)	u(92) ʊ(8)
EH	/u/	u(98)	u(85) ʊ(7) o(4)

EL = Early-low; EH = early-high; LH = late-high. Only percentages greater than 2% are shown.

for /i ɪ eᶦ ɛ æ/. *Who'd, hood, hoed, hut, hot, head* and *had* were used to classify productions of the target vowels /u ʊ o ʌ ɒ/. The listeners were required to classify the vowel in each stimulus by clicking one of the seven keywords shown on a computer screen. Each vowel token was then assigned a phonetic symbol based on the keyword that was selected. For example, selection of the keyword *who'd* (/hud/) was taken to mean that a listener had heard a vowel that was closer to Canadian English /u/ than to any other Canadian English vowel. The percentage of times that each phonetic symbol was used to classify each vowel was computed for each group.

Results

The results shown in table 6 summarize the listeners' classifications of vowels produced by the subjects in three groups. The numbers in parentheses indicate the percentage of times, based on a maximum number of 108 forced-choice judgments

(18 subjects per group × 6 listeners), that the productions of a particular vowel were classified as instances of various English vowels. The classification of a target vowel as itself (e.g., productions of /ɪ/ heard as /ɪ/) corresponded to the intelligibility data reported previously by Flege et al. [1999].

The first aim of this experiment was to determine if early-low subjects produced more vowels that were misidentified than the early-high subjects did. If this were so, then the magnitude of the early-low subjects' vowel errors may have been underestimated in experiment 1. However, a consideration of the data in table 6 indicates that such an artifact did not affect the results of experiment 1. There are five relevant cases in table 6 to consider (i.e., the cases where an NE versus early-high difference was observed in experiment 1 but not an NE versus early-low difference). In four of these cases, the early-low subjects produced fewer rather than more misidentified vowels than the early-high subjects did (/ɪ/ in nonwords: 31 vs. 57%; /ɛ/ in nonwords: 29 vs. 45%; /ʌ/ in nonwords: 39 vs. 49%; /ʊ/ in nonwords: 55 vs. 69%). This suggested that the lack of a difference between the early-low and NE groups was not due to an underestimation of the early-low groups' errors.

The remaining case of interest involved the production of /ʌ/ in words. The percentage of times that the productions of this vowel by the early-low and early-high groups were misidentified differed little (21 vs. 20%). Moreover, the pattern of misidentifications for /ʌ/ in words is virtually identical. In this case, it seems reasonable to conclude that the early-high subjects showed a greater degree of within-category difference from the phonetic norms of English for /ʌ/ than the early-low subjects did. Taken together, the results suggest that vowels spoken by the early-low subjects were not misidentified more often than the vowels spoken by the early-high subjects. If anything, the reverse held true.

It might be objected that the foregoing analysis did not include all 10 of the listeners who rated vowels in experiment 1, nor did it include results for /ɚ/, which was not examined in experiment 2. Accordingly, the percentage of times that the 'wrong vowel' button was selected in experiment 1 has been tabulated for all 11 vowels examined. (For the sake of comparison, results have also been presented for the late-high group.) Each mean value shown in table 7 was based on a maximum of 180 possible judgments (10 listeners × 18 subjects per group). The six relevant cases in the table have been boldfaced. In each case, a larger percentage of vowels spoken by the early-high group than the early-low group was classified as the 'wrong' vowel. These findings also suggest that the absence of differences for the early-low and NE groups was not due to an underestimation of the early-low subjects' errors.

The second aim of experiment 2 was to evaluate the role of orthography. The pattern of data in table 6 suggests that some of the early bilinguals' vowel production errors in nonwords were due, at least in part, to the influence of orthography. For example, the ratings obtained in experiment 1 for the early-high and NE groups' productions of /ɪ/ differed significantly in nonwords but not words. The subjects produced four words with /ɪ/ *(rid, did, hid, bid)* after hearing a native English speaker's productions and seeing these words on a written list. The /ɪ/s produced by the early-low and early-high subjects in words were never classified as /i/ whereas their productions of /ɪ/ in nonwords were often classified as /i/ (means = 31 and 57%, respectively). The early bilinguals probably accessed phonological representations for words stored in their mental lexicon when asked to produce the four English words containing /ɪ/. However, some of them – especially subjects in the early-high group – may have pronounced the

Table 7. The mean percentage of times that the 10 listeners in experiment 1 classified vowel tokens as 'wrong' (i.e., not an instance of the intended category)

	Word condition				Nonword condition			
	NE	EL	EH	LH	NE	EL	EH	LH
/i/	2.5	3.5	3.0	8.6	1.0	0.5	6.1	7.1
/ɪ/	1.5	2.5	0.0	23.7	0.5	**23.7**	**37.9**	37.4
/eᶦ/	2.0	0.5	0.0	15.7	2.5	8.1	8.1	16.2
/ɛ/	3.5	1.5	6.6	15.2	0.5	**10.1**	**17.7**	17.2
/æ/	1.5	0.5	3.0	8.1	3.0	2.5	13.1	10.1
/ɒ/	2.0	1.0	6.1	12.1	1.5	4.0	20.2	15.7
/ʌ/	2.5	**4.0**	**5.6**	34.8	0.5	**15.7**	**24.2**	36.9
/o/	0.5	0.5	0.5	23.5	3.5	4.0	6.6	19.2
/ʊ/	1.0	5.1	5.6	33.8	8.6	**18.7**	**33.8**	35.4
/u/	1.0	2.0	1.5	18.2	2.0	5.6	10.1	23.7
/ɚ/	2.0	2.0	3.5	12.1	10.6	**28.8**	**35.4**	43.4

EL = Early-low; EH = early-high; LH = late-high.

letter 'i' in *rid, did, hid* and *bid* as it is pronounced in Italian (i.e., /i/) when asked to insert the vowel common to all four words into a /b_do/ frame.

A second pattern suggesting an influence of orthography involved the vowel /ɛ/. The early-high and NE subjects' productions of /ɛ/ in nonwords but not words received significantly different ratings in experiment 1. Three of the four English words with /ɛ/ were spelled with 'e' *(red, dead, bed),* which is often pronounced /e/ in Italian. An unpublished study by Flege and Fox revealed that Italian /e/ is typically heard as /eᶦ/ by native speakers of English. The early-high groups' productions of /ɛ/ in words were never classified as /eᶦ/ in this experiment; however, their /ɛ/ productions in nonwords were classified as /eᶦ/ in 31% of instances.

A third pattern suggesting an orthographic influence involved /ʊ/. The early-high and NE subjects' /ʊ/s differed in nonwords but not words in experiment 1. All four words with /ʊ/ were spelled with at least one 'o' *(good, could, would, hood),* which is often pronounced /o/ in Italian. Italian /o/ is typically heard as /o/ by native speakers of English. The early-high groups' productions of /ʊ/ in words were never classified as /o/ whereas the /ʊ/s they produced in nonwords were heard as /o/ in 34% of instances.

Other similar patterns existed. All four of the English words containing /ɒ/ were spelled with 'o' *(rod, sod, cod, god).* The early-high subjects' productions of /ɒ/ in words were never heard as /o/, but the /ɒ/s they produced in nonwords were heard as /o/ in 20% of instances. All of the words with /ʌ/ were spelled with 'u' *(cud, mud, dud, bud),* which is pronounced /u/ in Italian. The early-high subjects' productions of /ʌ/ in words were never heard as /u/, but the /ʌ/s they produced in nonwords were heard as /u/ in 10% of instances. The early-high subjects' production of /æ/ as /ɒ/ (13% of instances) might be attributed to the fact that all four of the words with /æ/ were spelled with 'a' *(mad, dad, had, bad).* Italian /a/ might be heard as Canadian English /ɒ/. Finally, the early-high subjects' production of /i/ as /ɒ/ in nonwords (3%) might be attributed to the fact that two of the words with /i/ were spelled with 'a' *(read, bead).*

Discussion

The first important finding of this experiment was that the early-low subjects' productions of certain English vowels were not misidentified more frequently than the early-high subjects' productions were. Rather, the opposite usually held true. This conclusion was based on the classification data here, and a consideration of how often vowels were classified as the 'wrong' vowel in experiment 1. This finding suggests that an important conclusion drawn from experiment 1 – namely, that the early-high but not the early-low subjects' vowels differed from the NE speakers' vowels – was not the result of an artifact in the scaling procedure used in experiment 1. Specifically, we did not miss finding a difference between the early-low and NE groups because we underestimated the magnitude of vowel production errors made by the early-low group. The results also showed that some of the native versus nonnative differences observed in experiment 1 were the result of categorical changes in vowels (e.g., /ɪ/s heard as /i/) whereas others were the result of within-category divergences from the phonetic norms of English that did not cause listeners to hear some other vowel.

The results obtained here strongly suggested that orthography affected early bilinguals' production of vowels in nonwords. However, it is important to note that orthographic effects cannot explain all of the early bilinguals' nonword vowel errors. Productions of /ɛ/ as /ɪ/ (early-low: 8%, early-high: 11%) can probably not be explained by orthography, nor can productions of /ʊ/ as /u/ (early-low: 43%, early-high: 30%). To take another example, the early-high subjects' production of /o/ as /ʊ/ (6%) and as /ʌ/ (4%) in nonwords cannot be explained by the spelling of the words with /o/ *(road, code, hoed, bode)*.

Moreover, we cannot explain with certainty why the early-high subjects seemed to have been influenced by orthography to a greater extent than the early-low subjects were. Producing vowels in nonwords (i.e., in a /b_do/ frame) was a psycholinguistically complex task. The Italian-English bilinguals first had to identify the vowel common to the four real words. They had to retain a code for the identified vowel in working memory while listening to a carrier phrase, then encode the represented vowel and output it through motor commands. Although vowels spoken in the /b_do/ frame resulted in nonoccurring words in English, the nonwords resembled Italian words [Carlson et al., 1985]. Perhaps pronunciations specified in an Italian lexicon influenced the early-high subjects' productions of the nonwords to a greater extent than the early-low subjects' productions because the early-high subjects' Italian lexicon was more strongly activated, on a long-term basis.

Still another possibility is a difference in the retention of auditory short-term memory information by the early-low and the early-high subjects. Several studies have revealed measurable differences in vowel perception between early bilinguals and L2 monolinguals [Mack, 1989; Pallier et al., 1997, 1999; Sebastián-Gallés and Soto-Faraco, 1999]. If a greater mismatch existed between the native-speaker models and the early-high groups' representations of English vowels, it might have slowed their coding of vowels during the nonword production task, or lessened the durability of codes they generated when hearing the target vowel. This, in turn, may have affected how well early-high subjects translated short-term memory codes into a mode suitable for producing /ɪ ɛ ɚ ʌ ʊ/ in a /b_do/ frame.

General Discussion

As expected from previous work [Flege, 1992a, b; Munro et al., 1996; Flege et al., 1999], experiment 1 revealed an effect of AOA on Italian-English bilinguals' accuracy in producing English vowels. Significantly lower ratings were obtained for seven (of 11) vowels spoken by subjects in the late-high than in the early-high group. The primary question addressed by this study, however, was whether the subjects in one or both of the two early bilingual groups examined here would produce English vowels that differed significantly from vowels spoken by NE monolinguals. The answer to this question depended on the early bilinguals' amount of continued L1 (Italian) use. Five vowels spoken by subjects in the early-high group (i.e., /ɪ ɛ æ ʌ ʊ/) received significantly lower ratings than vowels spoken by the NE monolinguals did. However, none of the vowels spoken by the subjects in the early-low group were found to differ from the NE monolinguals' vowels.

The finding obtained here for L2 vowel production agrees with the results of a recent study examining Italian-English bilinguals' categorial discrimination of English vowels. Flege and MacKay [submitted] found that subjects in an early-high group, but not those in an early-low group, differed significantly from NE monolinguals in discriminating certain pairs of English vowels. When taken together, the results obtained here and by Flege and MacKay [submitted] for early-low groups do not support the conclusion [Sebastián-Gallés and Soto-Faraco, 1999, p. 120] that 'severe' limitations exist on the 'malleability of the initially acquired L1 phonemic categories'. In fact, the lack of a difference between the early-low and NE groups suggests that, under certain circumstances, it is possible for early bilinguals to produce and perceive English vowels in a manner that is functionally equivalent to the performance of L2 monolinguals.

The difference we observed between the early-high and NE groups confirms theoretical perspectives [e.g., Paradis, 1978; Mack, 1984; Flege, 1995; Grosjean, 1999] leading to the expectation that even fluent early bilinguals will not produce all L2 vowels in a manner that is indistinguishable from L2 monolinguals. Our results suggest that, under certain conditions of language use and/or L1 maintenance, a native-like performance in the L2 is unlikely to arise.

Our observation of significant differences between the early-high and NE groups in this study differs from the results obtained in vowel production studies reviewed in the 'Introduction'. A difference may have been observed in the present study because a more fine-grained form of auditory evaluation was used here than in one previous study [Flege et al., 1999], because the early bilinguals in this study were more experienced in English than those examined in another previous study [Flege, 1992a], or because the early bilinguals examined here used their L1 more frequently than did the bilinguals in still another previous study [Munro et al., 1996].

As mentioned, Flege and MacKay [submitted] observed a strong effect of L1 use on Italian-English bilinguals' categorial discrimination of English vowels. Other studies have shown an effect of amount of L1 use on Italian-English bilinguals' overall degree of foreign accent in English sentences [Flege et al., 1997b; Piske and MacKay, 1999; Piske et al., 2001], their identification of English consonants [MacKay et al., 2001], as well as their recognition of English words [Meador et al., 2000].

The basis for L1 use effects on performance in an L2 is uncertain at present. What we have referred to as an effect of 'Ll use' may actually be an effect arising from differences in amount of L2 input. This is because, in bilinguals, the frequency of L1 and

L2 use are inversely related. Another possibility is that the bilinguals who continued to use Italian relatively often (early-high) were more likely to hear Italian-accented English than were those who used Italian seldom (early-low). Still another possibility that should be examined in future research is that a frequent continued use of the L1 increases the influence of the L1 phonetic subsystem on the phonic elements making up the L2 subsystem. If so, then it will be important to determine whether the L1 system of fluent early bilinguals affects L2 performance indirectly, as the result of effects on representations developed slowly over time for L2 vowels, or directly affects the real-time regulation of L2 speech.

This study also suggested a difference in the accuracy with which Italian-English bilinguals produced English vowels in two conditions. In one condition, vowels were spoken in words following the auditory presentation of native speaker models. In the other condition, vowels were spoken in nonwords long after the native speaker models had been presented. Four of the five differences observed between the early-high and NE groups were confined to vowels spoken in the nonword condition. The basis for the difference between vowels spoken in the two conditions is uncertain. Perhaps the early-high subjects were less able than the early-low subjects to retain information for English vowels in auditory short-term memory in the nonword condition because the native speaker models diverged more from the early-high than the early-low groups' long-term memory representations for vowels [see Mack, 1989; Sebastián-Gallés and Soto-Faraco, 1999]. Experiment 2 revealed that many of the early-high groups' vowel production errors in nonwords showed an influence of orthography. In some instances, subjects seem to have produced an English vowel as the letter representing it would be pronounced in Italian. One hypothesis that should be investigated is that bilinguals who have developed long-term representations for vowels that differ from those of L2 monolinguals are more likely to be influenced by L1-inspired spelling conventions [Carlson et al., 1985] than are bilinguals who have developed native-like representations for L2 vowels. Another possibility that should be considered is that the L1 lexicon of bilinguals who continue to use the L1 frequently is more strongly activated, on a long-term basis, than is the L1 lexicon of those who use the L1 seldom.

In summary, the results obtained here support the view [e.g., Paradis, 1978; Mack, 1984; Flege, 1995; Grosjean, 1999] that fluent early bilinguals may differ from monolingual native speakers of an L2 in producing L2 vowels. However, differences between L2 monolinguals and early bilinguals may not be inevitable. The fact that early bilinguals who seldom used their L1 (early-low) never differed significantly from NE monolinguals argues against the conclusion that 'severe' limitations exist on the malleability of the phonetic system once the L1 phonetic inventory has been established. Additional research is needed to determine why native versus nonnative differences are more likely to be observed for bilinguals who continue to use their L1 often, and why they are more likely to be observed in certain elicitation conditions than in others.

Acknowledgments

This study was supported by grant DC00257 from the National Institute for Deafness and Other Communicative Disorders. The authors thank J. Prosperine and M. Pearse for their help recruiting subjects, St. Anthony's parish in Ottawa, all of the participants, C. Schirru for help with data reduction, and K. Aoyama, G. Busà, C. Best, S. Guion, A. Højen, H. Winitz and an anonymous reviewer for their comments on an earlier version of this paper.

References

Agard, F.; DiPietro, R.: The sounds of English and Italian (University of Chicago Press, Chicago 1964).

Best, C.; Faber, A.; Levitt, A.: Perceptual assimilation of non-native vowel contrasts to the American English vowel system (Abstract). J. acoust. Soc. Am. *99:* 2602 (1996).

Bongaerts, T.; van Summeren, C.; Planken, B.; Schils, E.: Age and ultimate attainment in the pronunciation of a foreign language. Stud. sec. Lang. Acquis. *19:* 447–465 (1997).

Bosch, L.; Costa, A.; Sebastián-Gallés, N.: First and second language vowel perception in early bilinguals. Eur. J. cogn. Psychol. *12:* 189–221 (2000).

Busà, M.G.: On the production of English vowels by Italian speakers with different degrees of accent; in Leather, James, New Sounds 92. Proc. 1992 Amsterdam Symp. on the Acquisition of Second-Language Speech, pp. 47–63 (Department of English, University of Amsterdam, Amsterdam 1992).

Busà, M.G.: L'inglese degli Italiani (Unipress, Padua 1995).

Carlson, R., Elenius, K.; Granström, B.; Hunnicutt, S.: Phonetic and orthographic properties of the basic vocabulary of five European languages. Q. Prog. Status Rep., Speech Transm. Lab., No. 1, pp. 63–94 (1985).

Flege, J.E.: The production and perception of speech sounds in a foreign language; in Winitz, Human communication and its disorders, a review 1988 (Ablex, Norwood 1988).

Flege, J.E.: Speech learning in a second language; in Ferguson, Menn, Stoel-Gammon, Phonological development: models, research, and implications (York Press, Timonium 1992a).

Flege, J.E.: The intelligibility of English vowels spoken by British and Dutch talkers; in Kent, Intelligibility in speech disorders (Benjamins, Amsterdam 1992b).

Flege, J.E.: Second-language speech learning: theory, findings, and problems; in Strange, Speech perception and linguistic experience: issues in cross-language research (York Press, Timonium 1995).

Flege, J.E.: The role of subject and phonetic variables; in Gruber, Higgins, Olsen, Wysocki, Papers 34th Annu. Meet. Chicago Ling. Soc., Vol. II (Chicago Linguistic Society, Chicago 1998).

Flege, J.E.; Bohn, O.-S.; Jang, S.: Effects of experience on nonnative subjects' production and perception of English vowels. J. Phonet. *25:* 437–470 (1997a).

Flege, J.E.; Frieda; E.M.; Nozawa, T.: Amount of native-language (L1) use affects the pronunciation of an L2. J. Phonet. *25:* 169–186 (1997b).

Flege, J.E.; Frieda, E.M.; Walley, A.C.; Randazza, L.: Lexical factors and segmental accuracy in second-language speech production. Stud. Sec. Lang. Acquis. *20:* 155–188 (1998).

Flege, J.E.; MacKay, I.R.A.: Malleability of the speech perception system: evidence from second-language acquisition (submitted).

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

Flege, J.E.; Munro, M.J.; Fox, R.A.: Auditory and categorical effects on cross-language vowel perception. J. acoust. Soc. Am. *95:* 3623–3641 (1994).

Flege, J.E.; Munro, M.J.; MacKay, I.RA.: Factors affecting degree of perceived foreign accent in a second language, J. acoust. Soc. Am. *97:* 3125–3134 (1995).

Fox, R.A.; Flege, J.E.; Munro, M.J.: The perception of English and Spanish vowels by native English and Spanish listeners: a multidimensional scaling analysis. J. acoust. Soc. Am. *97:* 2540–2551 (1995).

Grosjean, F.: Neurolinguists, beware! The bilingual is not two monolinguals in one person. Brain Lang. *36:* 3–15 (1989).

Grosjean, F.: Processing mixed language: issues, findings and models; in de Groot, Kroll, Tutorials in bilingualism: psycholinguistic perspectives (Erlbaum, Mahwah 1997).

Grosjean, F.: Studying bilinguals: methodological and conceptual issues. Bilingualism Lang. Cogn. *1:* 117–130 (1999).

Guion, S.G.; Flege, J.E.; Loftin, J.D.: The effect of L1 use on foreign accent ratings in Quichua-Spanish bilinguals. Proc. 14th Int. Congr. Phonet. Sci., pp. 1471–1474, San Francisco 1999.

Guion, S.G.; Flege, J.E.; Loftin, J.D.: The effect of L1 use on pronunciation in Quichua-Spanish bilinguals. J. Phonet. *28:* 27–42 (2000).

Hammarberg, B.: The course of development in second language phonology acquisition: a natural path or strategic choice? In Hyltenstam, Viberg, Progression and regression in language: sociocultural, neuropsychological and linguistic perspectives (Cambridge University Press, Cambridge 1993).

Hillenbrand, J.; Getty, L.; Clark, M.J.; Wheeler, K.: Acoustic characteristics of American English vowels. J. acoust. Soc. Am. *97:* 3099–3111 (1995).

Jun, S.-A.; Cowie, I.: Interference for 'new' and 'similar' vowels in Korean speakers of English. Ohio State Univ. Working Papers *43:* 117–130 (1994).

Lambert, W.; Rawlings, C.: Bilingual processing of mixed-language associative networks. J. verbal Learn. verbal Behav. *8:* 604–609 (1969).

McAllister, R.; Flege, J.E.; Piske, T.: The acquisition of Swedish long vs. short vowel contrasts by native speakers of English and Spanish. Proc. 14th Int. Congr. Phonet. Sci., pp. 751–754, San Francisco 1999.

McAllister, R.; Flege, J.E.; Piske, T.: The influence of L1 on the acquisition of Swedish vowel quantity by native speakers of Spanish, English and Estonian (submitted).

Mack, M.: Early bilinguals: how monolingual are they? In Paradis, Lebrun, Early bilingualism and child development (Swets & Zeitlinger, Lisse 1984).

Mack, M.: Consonant and vowel perception and production: early English-French bilinguals and English monolinguals. Percept. Psychophys. *46:* 187–200 (1989).
MacKay, I.R.A.; Meador, D.; Flege, J.E.: The identification of English consonants by native speakers of Italian. Phonetica *58:* 103–125 (2001).
Major, R.: Phonological similarity, markedness, and rate of L2 acquisition. Stud. sec. Lang. Acquis. *9:* 63–82 (1987).
Meador, D.; Flege, J.E.; MacKay, I.R.A.: Factors affecting the recognition of words in a second language. Bilingualism Lang. Cogn. *3:* 55–67 (2000).
Milani, C.: Language contact among North American people of Italian origin; in Sture Ureland, Clarkson, Language contact across the North Atlantic (Niemeyer, Tübingen 1996).
Munro, M.J.: Production of English vowels by native speakers of Arabic: acoustic measurements and accentedness ratings. Lang. Speech *36:* 39–66 (1993).
Munro, M.J.; Flege, J.E.; MacKay, I.R.A.: The effects of age of second-language learning on the production of English vowels. Appl. Psycholing. *17:* 313–334 (1996).
Pallier, C.; Bosch, L.; Sebastián-Gallés, N.: A limit on behavioral plasticity in speech perception. Cognition *64:* B9–B17 (1997).
Pallier, C.; Sebastián-Gallés, N.; Colomé, A.: Phonological representations and repetition priming. Eurospeech '99, Budapest 1999.
Paradis, M.: The stratification of bilingualism; in Paradis, Aspects of bilingualism (Hornbeam Press, Columbia 1978).
Piske, T.; MacKay, I.R.A.: Age and L1 use effects on degree of foreign accent in English. Proc. 14th Int. Congr. Phonet. Sci., pp. 1433–1436, San Francisco 1999.
Piske, T.; MacKay, I.R.A.; Flege, J.E.: Factors affecting degree of foreign accent in an L2: a review. J. Phonet. *29:* 191–215 (2001).
Romito, L.; Trumper, J.: Un problema della coarticulazione: L'isocronia rivistata. Acts 27th Natl. Conf. Ital. Acoust. Assoc., pp. 449–455 (1989).
Sebastián-Gallés, N.; Soto-Faraco, S.: On-line processing of native and non-native phonemic contrasts in early bilinguals. Cognition *72:* 111–123 (1999).
Soares, C.; Grosjean, F.: Bilinguals in a monolingual and a bilingual speech mode: the effect on lexical access. Mem. Cognition *12:* 380–386 (1984).
Trumper, J.: L'influenza di eventi macrosismici su alcune discontinuità linguistiche (Calabria); in Pellegrini, Saggi Dialettologici in Area Italo-Romanza, pp. 89–103 (CNR, Padova 1995).