The effects of age of second language learning on the production of English vowels

MURRAY J. MUNRO
Simon Fraser University

JAMES EMIL FLEGE
University of Alabama, Birmingham

IAN R. A. MACKAY
University of Ottawa

ADDRESS FOR CORRESPONDENCE
Murray J. Munro, Department of Linguistics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

ABSTRACT
This study examined the English vowel productions of 240 native speakers of Italian who had arrived in Canada at ages ranging from 2 to 23 years and 24 native English speakers from the same community. The productions of 11 vowels were rated for degree of foreign accent by 10 listeners. An increase in perceived accentedness as a function of increasing age of arrival was observed on every vowel. Not one of the vowels was observed to be produced in a consistently native-like manner by the latest-arriving learners, even though they had been living in Canada for an average of 32 years. However, high intelligibility (percent correct identification) scores were obtained for the same set of productions. This was true even for English vowels that have no counterpart in Italian.

In a comprehensive study of the relationship between age of second language learning (AOL) and pronunciation, Flege, Munro, and MacKay (1995b) examined the strength of perceived foreign accent in the speech of Italian learners of English. Standard sentences produced by a large group of talkers ($N = 240$) were rated by native English listeners for degree of perceived accent. The talkers had been living in Canada for a mean of 32 years and had almost certainly reached their ultimate level of attainment in English. Nonetheless, the listeners were quite aware of an Italian accent in the majority of the utterances, and they consistently rated sentences as less "native-like" as a function of the talkers' increasing age of arrival in Canada. One listener, in fact, gave significantly lower ratings to the subset of talkers who had arrived at 3.1 years of age than to the native English comparison group. Moreover, not a single speaker who began learning
English at or after the age of 16 years received an accent rating within the native English range. Very few speakers who began learning English between 12 and 15 years did so.

When assigning foreign accent scores, the listeners probably took into account the frequency and degree of segmental errors in the utterances. In fact, in a related study involving the same talkers, Flege, Munro, and MacKay (1995a) reported AOL effects on a number of individual consonant segments, including word-initial [t ð d p t]. However, we know of no detailed study of the effects of AOL on vowel production. The present study addresses this issue through an examination of the productions of 11 English vowels by the same group of Italian talkers. It focuses on whether the relationship between AOL and perceived accentedness in vowel production is comparable to the one seen in the Flege et al. (1995b) study of sentence production. It also explores differences among ratings of vowels in order to determine whether some second language (L2) vowels are more resistant to AOL effects than others. Finally, the study examines how accentedness in L2 vowels compares with vowel intelligibility.

THE PERCEPTION OF L2 VOWELS BY LATE LEARNERS

Although vowels have tended to receive less attention in L2 perceptual research than have consonants, some studies have revealed effects of AOL on vowel perception. Blankenship (1991), for instance, had native Spanish listeners with AOLs between 5 and 23 identify synthetic English vowels according to six English categories ([i ɪ ɛ æ ə ɑ]). She observed greater perceptual differentiation of the vowels (i.e., clearer evidence of separate vowel categories) in listeners who had begun learning English at relatively early ages than in those who had done so later in life.

One reason why learners with relatively late AOLs may perceive vowels in a "nonnative" way is that they may not use perceptual cues in the same way that native listeners do. For example, Bohn and Flege (1990), in a study of German learners of English, and Munro (1992), in a study of Arabic learners, observed that native and nonnative listeners tended to assign different "weights" to spectral and temporal cues when identifying tokens from a synthetic [i]-[ɪ] continuum (see also Gottfried & Beddor, 1988).

Another account of segmental perception by L2 learners comes from Best's Perceptual Assimilation Model (Best, 1994; Best, McRoberts, & Sithole, 1988), which proposes that L2 sounds tend to be perceived in terms of L1 categories. Nonnative phones are predicted to assimilate to various degrees to L1 categories (or not to assimilate at all), depending on the extent to which they resemble the L1 categories. The primary evidence supporting this model currently comes from studies of consonant perception, but the model is relevant to vowel perception as well.

L2 VOWEL PRODUCTION

The fact that vowels are often perceived differently by late L2 learners and native listeners may explain why some studies have revealed L1 effects on
L2 vowel production. Work by Rochet (1995), for example, indicated that speakers of Portuguese tended to perceive and produce French [y] more often as an [i]-like vowel than as [u], while speakers of English tended to do the reverse. In addition the finding by Munro (1992) of a high degree of sensitivity to vowel duration differences on the part of Arabic speakers may explain why they tend to exaggerate duration differences in productions of English tense-lax pairs (Mitleb, 1981; Munro, 1993).

Data from other studies indicate that L2 vowel productions from late learners sometimes show evidence of a phonetic “approximation” (see Flege, 1980) of L2 vowel targets. For instance, Basson (1988) reported that English [u] productions by native speakers of Japanese tended to exhibit first and second formant (F1 and F2) frequencies falling between the values typical of English [u] and Japanese [uu]. Munro (1993) and Mitleb (1981) also observed approximation in Arabic speakers' productions of some English vowels.

Research on Italian speakers' productions of English vowels is very limited. The “standard” Italian dialect has seven vowels [I e e a o u] (Ferrero, Magno-Caldognotto, Vagges, & Lavagnoli, 1978), but certain dialects lack [o] and/or [e]. Some English vowels have apparent analogs in Italian (e.g., [i u]), but some (e.g., [I u]) do not. Busà (1992) examined the English [u] and [u] productions of Italians living in the United States. An acoustic analysis indicated that nearly all speakers tended to produce English [u] with F2 values lower (i.e., more Italian-like) than those of native English speakers. In the case of [u], however, F1 and F2 values for the Italians tended to be lower than for native English talkers, but a larger number of Italian talkers produced [u] tokens with native-like spectral and temporal properties than was the case for [u]. The Italians, then, appeared to have more success with English [u] than [u].

The question of how to predict which segments in the L2 will be relatively easy or difficult to produce for particular groups of learners has received considerable attention from theorists and researchers. Traditional contrastive analysis, for instance, proposes that difficulties in segmental production arise because of differences between the L1 and L2 segmental inventories (see, e.g., Weinreich, 1957). In particular, L2 learners are expected to have difficulty producing phones that occur in the L2 but not in the L1. In contrast, Flege's Speech Learning Model (SLM) predicts that phones from the L2 that are “new” (i.e., relatively different from any L1 phone) will eventually be perceived and produced correctly by late L2 learners, while L2 phones that are “similar” (i.e., differ only slightly from L1 phones) will not be (see Flege, 1992a, 1992b). Differences in success with individual phones are hypothesized to depend on whether or not new perceptual categories can be established for them. In the case of similar sounds, the effects of equivalence classification are predicted to block new category formation. For new sounds, however, it is expected that new perceptual categories will eventually be established.

The findings of Busà (1992) appear to support the SLM, as do the results of a number of other studies of vowel production. Flege and Hillenbrand
(1984) and Flege (1987), for instance, observed that the productions of French [y] (which has no clear English analog) by native English learners of French showed more French-like F2 frequencies than did their productions of French [u] (but see also Rochet, 1995). Moreover, the findings of Bohn and Flege (1990) suggested that it was eventually possible for experienced German learners of English to establish a new phonetic category for [æ] (which has no German counterpart), but not for [i] or [ɪ].

In general, adherents to any account of L2 speech production that attempts to explain differential success with individual phones would like to be able to make predictions of the type, "Phone x, but not phone y, of the L2 will be produced correctly by late-learning native speakers of a particular L1." But, in order for any such prediction to be borne out, at least two conditions must be met. First, the learners from the L1 background in question must have more or less uniform success (or lack of success) in their productions of the phones of interest. It must be clear that all, or at least a large majority, of speakers produce the phones correctly or incorrectly. This type of prediction, then, is formulated in such a way that it precludes the possibility of extensive individual differences among talkers. Second, the means of evaluating the learners' success in producing particular phones must yield unambiguous results. It must be possible to state with a high degree of certainty whether any given production is "correct" or not.

Evidence from recent research suggests that neither of these conditions is likely to be met in studies of L2 speech. First, Munro's (1993) study casts doubt on the proposal that adult L2 learners are consistently able to produce some L2 vowels correctly, while producing others with an accent. Productions of 10 American English vowels by native speakers of Arabic who had been living in the United States for an average of 6 years were analyzed. Data on F1 and F2 frequencies and F1 and F2 movement patterns showed that, as a group, the Arabic speakers produced all 10 vowels differently from the vowels of the native English speakers. Furthermore, in an accent-rating task in which productions of the five front vowels [ɪ e ɪ eɪ æ] were rated by a panel of phonetically trained judges, only a minority of the vowel productions received scores comparable to those assigned to native English speakers' productions. Although some individual vowel tokens were rated more native-like than others, none of the vowels of English could be said to have been consistently produced in a native-like way by the Arabic-speaking learners as a group.

However, the Munro (1993) study did not provide firm evidence about the ultimate learnability of L2 vowels. For one thing, many of the Arabic speakers had been living in the United States for less than 5 years. It is likely, therefore, that some of the talkers had not yet reached their maximum level of proficiency in English pronunciation. Moreover, the vowel stimuli presented to the native English raters consisted of complete CVCs. It is possible that an Arabic accent was detected in the final consonant portions of these productions. In the present study, we remedied both of these methodological problems. First, our subject pool was restricted to talkers who had been living in an English-speaking environment for an
extended period of time (a minimum of 15 years). The talkers were selected so that AOL varied systematically. Second, the talkers' CVC productions were edited to minimize the effects of consonantal properties that might create an impression of a foreign accent.

Evidence also suggests that studies in which native listeners rate the productions of nonnative speakers are unlikely to meet the second condition (i.e., to make use of an unambiguous means of evaluating L2 speech). In a study by Munro and Derwing (1995), native English listeners transcribed orthographically Mandarin-accented utterances and rated the same utterances for accentedness. Although individual listeners tended to show agreement on which utterances were heavily accented, there was considerable interlistener variation in their ratings. Furthermore, utterances that were rated as heavily accented were sometimes observed to be highly intelligible to the listeners (as determined from transcription scores). The latter observation paralleled the finding by Flege, Takagi, and Mann (1995) that late-learning Japanese speakers' productions of English /i/ and /I/ may be accented yet identifiable. It is not known whether these findings might be replicated in a study of the production of individual vowels. In the present study we address this question by carefully examining data from individual raters and by comparing accent-rating data with vowel identification data.

In summary, the available evidence suggests that the degree of nativeness in L2 vowel production probably depends on AOL, and that the limitations of L2 vowel production may have a perceptual basis. However, the exact nature of the relationship between AOL and vowel production is unknown. It is not known, for example, at what ages of learning perceptible foreign accents begin to emerge in L2 vowel production. While between-vowel differences in “nativeness” of production have been observed, it is not known whether highly experienced late L2 learners consistently produce any vowels from the L2 in a native-like way. Moreover, there may be important individual talker differences in the accuracy with which L2 vowels are produced. The two experiments reported here were carried out in an attempt to explore these issues, rather than to evaluate any particular predictive model.

EXPERIMENT 1: FOREIGN ACCENT RATINGS

The purpose of Experiment 1 was to determine the relationship between AOL and the degree of perceived accentedness in 11 English vowels produced by native Italian learners of English. A second purpose was to determine which vowels, if any, from the Canadian English vowel system would be consistently produced without a detectable accent by the Italian learners.

METHOD

Talkers

The talkers were 240 native speakers of Italian (130 females, 110 males) and 24 native speakers of Canadian English (14 females, 10 males). The Italian speakers were members of a Catholic church in Ottawa, Ontario,
who had immigrated to Canada from Italy at ages ranging from 2 to 23 years. At the time of the experiment, they had a mean age of 44 years and had been living in Canada for 15 to 44 years ($M = 32$). A member of the parish screened and selected subjects so that the range of AOLs represented would be evenly distributed across the range of 2 to 23 years. The native English talkers were all speakers of General Canadian English who had grown up in Ontario. They had a mean age of 39 years, and most were members of the same church as the native Italian speakers. All participants had normal hearing, according to self-report.

The Italian-speaking talkers were rank-ordered according to AOL, and each talker was assigned to one of ten AOL groups consisting of 24 talkers each. The 24 talkers assigned to group 1 had a mean AOL of 3.1 years; the next 24 talkers in group 2 had a mean AOL of 5.2 years; and so on. Descriptive data for each group of talkers are given in Table 1. The talkers came from more than one dialect area in Italy, but the majority were from three regions: Calabria (33%), Abruzzi (23%), and Sicily (10%). Other regions that were represented included Puglia, Molise, and Lazio. An examination of the distribution of dialects across AOL groups revealed that the largest region represented in all 10 subgroups was Calabria. The representation of the other dialects was roughly proportional across groups.

### Table 1. Mean age of arrival in Canada of the ten groups of Italian talkers (n = 24 per group) who participated in the study

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean Age</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.1</td>
<td>1.9–4.1</td>
</tr>
<tr>
<td>2</td>
<td>5.2</td>
<td>4.2–6.4</td>
</tr>
<tr>
<td>3</td>
<td>7.5</td>
<td>6.5–8.6</td>
</tr>
<tr>
<td>4</td>
<td>9.6</td>
<td>8.7–10.6</td>
</tr>
<tr>
<td>5</td>
<td>11.6</td>
<td>10.6–12.6</td>
</tr>
<tr>
<td>6</td>
<td>13.6</td>
<td>12.7–14.8</td>
</tr>
<tr>
<td>7</td>
<td>15.8</td>
<td>15.0–16.8</td>
</tr>
<tr>
<td>8</td>
<td>17.5</td>
<td>16.8–18.5</td>
</tr>
<tr>
<td>9</td>
<td>19.3</td>
<td>18.5–20.2</td>
</tr>
<tr>
<td>10</td>
<td>21.5</td>
<td>20.2–23.2</td>
</tr>
</tbody>
</table>

Procedure

**Recordings.** Individual recording and interview sessions were held in a quiet room in the church. The experimenter (MJM) first verified that each participant met the selection criteria. During the recording phase, the talkers heard (through a loudspeaker) a recorded stimulus sentence beginning with the target word ("____ is the next word"). During the pause that followed, they produced a carrier sentence containing the target word in sentence-final position ("Now I say ____"). Each stimulus word was also
provided in written form on a sheet of paper. The stimulus items, which were part of a longer list of words and sentences used in experiments reported elsewhere (see Flege et al., 1995a, 1995b) were beat, bit, bait, bet, bat, but, Bert, bock, boat, book, and boot, representing, respectively, the Canadian English vowels [ɪ i ɛ e ʌ æ ɔ oʊ oʊ u]. Each word appeared twice on the randomized list. Recordings were made with a head-mounted microphone connected to a Sony Pro-II cassette recorder. When the recording phase was completed, the subjects answered a detailed questionnaire about their use of English (see Flege et al., 1995b). Each session lasted about 25 minutes.

**Stimulus preparation.** The second production of each word was low-pass filtered (10 kHz), digitized at 22.05 kHz with 16-bit resolution, and saved as an audio file. Signal editing was carried out visually and aurally with a Kay Computerized Speech Lab (CSL). In each case, the word of interest was edited from the carrier sentence. To minimize the effect of the final consonant on perceived accentedness in the listening task, the final [t] or [k] was removed, beginning at the point of complete constriction. Word-initial voiced Italian stop consonants differ from their English analogs in that Italian stops always show closure voicing prior to release (Vagges, Ferrero, Magno-Caldognetto, & Lavagnoli, 1979), whereas English stops do so only some of the time (Lisker & Abramson, 1964). For this reason it was hypothesized that the presence of strong prevoicing in the [bVt] words used here might give an impression of an Italian accent to some productions. Therefore, the closure voicing of all initial [b] tokens was removed to the point of the release burst. After editing, each signal was multiplied by a ramping function (20 ms on and 30 ms off) to eliminate potentially distracting transients at the beginning and the end of each signal. The edited tokens were then normalized for peak amplitude. The final stimulus set consisted of 2,904 vowel tokens (11 Vowels × 264 Talkers).

**Ratings.** Individual listening sessions were held in a sound-treated booth. The listeners were 10 native speakers of General Canadian English (2 females, 8 males), 9 of whom were phonetically untrained and 1 of whom was a trained phonetician (MJM). They ranged in age from 19 to 36 years at the time of the experiment ($M = 27$), and all of them passed a pure-tone hearing screen (250–4000 Hz). Most of the listeners were from Ontario, though some had grown up elsewhere in Canada. All of them had participated previously in a similar consonant-rating task involving the same talkers (see Flege et al., 1995a). The entire experiment was completed in three sessions of approximately 1 hour each, held on different days. Each listener rated each of the 2,904 vowels once. The stimuli were blocked by vowel, and during each session either three or four different blocks of vowels were rated. Each listener rated the 11 vowels in a different random order, and, within blocks, different stimulus randomizations were used for each listener. The listeners had no knowledge of which AOL group each talker belonged to.
The listeners were seated in front of a keyboard and heard the stimuli binaurally through headphones at a comfortable listening level. During each block, a card with the relevant word was attached to the keyboard. The listeners were advised that the stimuli had been modified so that the final consonant would not be audible, and that they were to direct their attention to the vowel portion of the word only. They were asked to judge how well each vowel was spoken by pressing a number from 1 (wrong vowel or very strong foreign accent) to 5 (correctly produced English vowel, no foreign accent).

RESULTS
A three-way analysis of variance was performed on the foreign accent ratings, with AOL group (11 levels) and vowel (11 levels) as between-subjects factors and listener (10 levels) as a repeated measure. All three main effects were found to be significant, as were all two-way interactions. The three-way interaction was also significant, \( F(900, 25,047) = 1.572, p < .05 \).

An examination of the data from individual listeners indicated that all of them showed a tendency to rate the vowels produced by the late-arriving Italian talkers as more accented than those produced by the native English talkers. Figure 1 gives the mean ratings (averaged over vowels) for each
listener according to AOL group. It can be readily seen that all listeners assigned progressively lower ratings as AOL increased. On average, the last-arriving AOL subgroup received ratings that were 26% lower than those assigned to the native English group. However, the fact that this difference ranged from 11 to 55% across listeners suggests that some listeners showed stronger AOL effects than others. The difference for the one trained listener was of an intermediate degree (23%): some of the untrained listeners were apparently more sensitive to an Italian accent than the trained listener, and some were less sensitive.

Because of the significant three-way interaction, separate one-way ANOVAs were carried out on each listener’s data for each of the 11 vowels. Because this resulted in a large number of tests, a very conservative alpha level of .0005 for each test was adopted to maintain an experiment-wise alpha level of roughly .05. Even when this stringent criterion was applied, significant effects of AOL group were observed on every vowel for at least some listeners. In other words, for every one of the 11 vowels, at least some listeners rated the productions of at least one of the 10 Italian AOL groups as more accented than the native English comparison group. Furthermore, every listener showed an AOL effect for at least some of the vowels.

To explore further the three-way interaction, Williams's (1971) test was used. This post-hoc procedure identifies the first of a logically ordered series of groups (in this case, Italian AOL subgroups) which differs significantly from a control group (here, the native English talkers); it was applied in those cases in which a significant AOL effect was observed in the ANOVA. Figure 2 summarizes the data by listener and vowel. The numerals at the top of each symbol represent the number of listeners who showed a significant effect of AOL for the vowel in question and in the Williams post-hoc test ($p < .005$). This total ranged from 2 listeners in the case of [e] to all 10 in the case of [æ]. The vertical segments represent the range of AOLs between which accents were first detected by those listeners who showed significant effects of AOL. For instance, in the case of [u], 9 listeners showed an effect of AOL. Of these, at least one rated the Italian group arriving at age 7.5 years as the first one whose vowel productions were more accented than those of the native English group, and at least one listener rated the group arriving at age 17.5 years as the first such group. Across vowels, the earliest age at which a foreign accent was detected ranged from 7.5 ([u]) to 15.8 ([n], [æ], [i]).

Once again it is clear that the 10 listeners showed varying degrees of sensitivity to accentedness in the vowels they evaluated. For instance, one listener showed significant effects for all 11 vowels, while one showed effects for only 2. Furthermore, even in those instances in which an AOL effect was observed, there was noticeable variability in the AOL at which the listeners first detected a foreign accent. While some listeners detected accents in [u], [æ], and [o] in talkers who had arrived in Canada before age 10, others first detected accentedness in these vowels in groups arriving 6 to 9 years later.
Figure 2. AOL effects for the 11 English vowels. Numbers at the top of each segment indicate how many listeners showed a significant effect of AOL for that vowel. The vertical segments indicate the range of ages at which accents were first detected. Means are indicated by darkened squares.

Cross-vowel comparisons

We also performed separate one-way ANOVAs to explore the effect of vowel in each listener's ratings. Because of the potential complexity of the effect, we restricted these tests to the data for the last-arriving group of talkers. We reasoned that this was the most likely group for whom clear between-vowel effects would be observed. In this case, an alpha level of .005 was adopted to maintain an experiment-wise level of .05. Three of the listeners showed no significant effect of the vowel factor by this criterion. For the remaining listeners, post-hoc Tukey HSD (honestly significant difference) tests were used to determine which pairs of vowel ratings differed significantly from each other. A relatively small number of these tests (24) yielded significant ($p < .005$) results. Nonetheless, a few patterns emerged. In 12 cases, [ɛ] received a rating that was significantly worse than that of some other vowel; in 6 cases, [ɪ] received such a rating; in 3 cases [ʊ] did; in 2 cases, [ʌ] did. On the other hand, in 5 cases, [æ] was rated as significantly better than some other vowel; in 4 cases [ɛ] received such a rating; in 3 cases, [ɔ] did; in 2 cases, [i] did.

We also estimated the number of talkers in each AOL group who had succeeded in producing each vowel in a “native-English-like” way. The
criterion we established was similar to one used in previous work (Flege et al., 1995b). In that study, foreign accent scores falling within two standard deviations of the native English mean score were deemed to be within the native English range. The individual differences among listeners in the present study made the use of an overall mean rating (pooled over listeners) an unsatisfactory approach. We therefore adopted a criterion which took into account individual listener differences. First, we computed individual listener means and standard deviations for the native English talker group. We then determined the number of productions (for each AOL group) that fell within two standard deviations of the native English mean for 9 out of 10 talkers. This approach resulted in a criterion which appeared to be fairly strict; in fact, some of the native English talkers’ vowel productions (about 9%) were ruled to be outside the “English” range. However, previous studies of English vowel identification by native listeners have shown that listeners often misidentify even some vowel productions of native speakers, perhaps because some vowels are not produced distinctly or because of listener errors (see, e.g., Assmann, Nearey, & Hogan, 1982; Flege, 1992a; Strange, 1989). Therefore, a criterion which resulted in native-like ratings for all native productions would be too lax.

Figure 3 shows the number of talkers at each AOL who produced each vowel in a native-like way, according to this criterion. This figure supports our finding of AOL effects on every vowel. Of particular interest here is the last AOL subgroup, that is, the talkers who had arrived in Canada at a mean age of 21.5 years. For four vowels ([i], [u], [ʌ], [ɔ]), 6 or fewer talkers (i.e., 25% or less) had native-like productions. The vowels produced well most often were [ɛ], [æ], [i], and [o], with 58%, 50%, 42%, and 42%, respectively, of the latest-arriving talkers producing them correctly. These findings clearly fail to support the hypothesis that at least some of the vowels from the English inventory would eventually be “mastered” by all late-arriving learners. The data indicate that, as a group, these learners did not produce any of the English vowels in a native-like way. However, some talkers produced some vowels without a noticeable accent. These results closely match the results of the cross-vowel ANOVAs, in that the same vowels emerged as the best- and worst-produced in both analyses.

When the data were considered from the perspective of individual talkers, it was found that none of the talkers in the last AOL group produced all of the vowels correctly. In fact, the number of vowels produced correctly varied from 0 for one of the talkers to 8 for another talker. The modal number of correct vowels was 3.

Although we cannot rule out the possibility that L1 dialect differences influenced the Italian talkers’ success in producing the English vowels, it is highly unlikely that such differences played a systematic role in the results. First, the distribution of dialects across the AOL groups was quite uniform. Second, a survey, by region, of the talkers who correctly produced or failed to produce individual vowels in a native-like way revealed no consistent patterns. For instance, the three talkers in the last-arriving group who pro-
Figure 3. The number of talkers in each AOL group who were rated by 10 Canadian English listeners as producing each vowel in a native-like way.
Acoustic comparisons

Mean F1 and F2 formant frequencies for 25 male productions of Standard Italian vowels have been reported by Ferrero, Magno-Caldognetto, Vagges, and Lavagnoli (1978). So that a comparison of Canadian English and Italian vowel inventories would be possible, F1 and F2 frequencies were measured for each of the 110 vowel tokens spoken by the 10 male native English speakers. LPC formant tracks were overlaid on standard spectrograms on the CSL, and measurements were made at the acoustic midpoint of each vowel. Figure 4 shows the mean values for the Canadian English speakers (plotted in an F1 × F2 mel space, commonly regarded as useful for examining acoustic differences) compared with the values for Italians obtained by Ferrero et al. (1978). A visual inspection of the two data sets indicates that some English vowels are more acoustically distant (in terms of F1 and F2) from any Italian vowels than others. For instances, English [ɔː] and [æ] have no apparent counterparts in Italian and are relatively distant from any 

Figure 4. F1 and F2 values (in mels) for Canadian English and Standard Italian vowels. The English data (dark circles) are based on measurements for the 10 male speakers in the present study. The Italian data (open squares) are based on data from male speakers reported by Ferrero et al. (1978).
Italian vowels.\textsuperscript{3} However, an examination of the foreign accent ratings for these two vowels indicates quite different patterns of results. If the number of native-like ratings assigned to these vowels is taken as an indication of how well the last AOL group produced them, then [ɛ] was the most poorly produced vowel, while [æ] was one of the best produced vowels. It appears that an accurate prediction of how well a vowel will be produced cannot be made here solely on the basis of acoustic distances (at least as measured here).

\section*{DISCUSSION}

The age of arrival in Canada of the Italian speakers examined in this study had an effect on the degree of perceived accent in all of the English vowels studied here. In many respects the findings are comparable to those of Flege et al. (1995b) for foreign accent in sentence-length utterances. In that study, accents were first detectable for AOLs as early as 3.1 years in some cases, but as late as 11.6 years in others. In the present study, the range was 7.5 years to 15.8 years. In this study, accents were first detected slightly later than in the previous study, perhaps because listeners had more opportunity to evaluate accentnedness in long (i.e., sentence-length) stimuli than in isolated CV utterances.

Some evidence was found here of differential success in the production of the vowels. In the case of [ɛ] and [æ], more than half of the talkers in the last-arriving AOL group produced vowels that were rated as good as the native speakers’ productions. In contrast, only a small minority of talkers could be said to have produced native-like tokens of other vowels such as [ʌ] and [ɜ]. These findings suggest that, for the late-arriving Italian learners, some English vowels were indeed more learnable than others. However, two important observations should be made here. First, the differences in learnability cannot be described in simple generalizations of the type, “Vowel x, but not vowel y, is readily learnable.” In no case did all, or even an overwhelming majority, of learners succeed in producing a particular vowel in a native-like way. Rather, individual differences in success among learners were observed in the productions of all 11 vowels. Second, even if it is accepted that some vowels are more learnable than others, no simple explanations can be offered for the patterns observed here. For example, the absence of a particular vowel from the L1 inventory does not appear to be a determining factor: neither [æ] nor [ɜ] occurs in Italian, yet in this study these vowels were the second-best and least well-produced, respectively. Nor do simple acoustic (F1 $\times$ F2) distances between the phones of English and Italian appear to account for the results. The hypothesis that English vowels that are acoustically “close” to Italian vowels should be produced less well than vowels that are more distant was not borne out by the data. From Figure 3 it can be seen that [ɜ] is quite distant from any Italian vowels, yet the latest-arriving speakers’ productions of this vowel rarely received ratings comparable to those of the native English speakers.
On the other hand, ratings of [i], which appears to be quite close to its Italian analog, were more often in the native-like range. These findings, then, run counter to the predictions of traditional contrastive analysis and to earlier versions of Flege's Speech Learning Model (Flege, 1992a, 1992b; but cf. Flege, 1995). Instead, they confirm Munro's (1993) finding that the mastery of any English vowel is unlikely for all, or even a large majority, of late-arriving learners of English. They also provide evidence of important individual differences in L2 vowel production that clearly require further study. For instance, given that some talkers produced some vowels correctly, one question to be asked is whether these talkers had, for some reason, established native-like perceptual representations for these vowels.

The finding of important interrater differences also raises issues that must be considered in further studies of L2 vowel production and perception. In particular, it appears that analyses of listener rating data do not lead to simple generalizations about the emergence of accent in vowel production. It was not possible here to identify a particular AOL at which accents were perceived to emerge on any given vowel because the AOL in question varied from listener to listener. Instead, it was necessary to identify a range of speaker AOLs within which listeners first began to recognize accentedness. It was beyond the scope of this study to attempt to account for these interrater differences, but this phenomenon requires further exploration.

EXPERIMENT 2: VOWEL IDENTIFICATION TASK

In Experiment 1, most of the vowels produced by the late-arriving learners of English were rated as foreign accented by the listeners. However, it was not determined to what extent these vowels were intelligible productions, despite their accentedness. Experiment 2 was undertaken with this question in mind. One phonetically trained listener (MJM) identified randomly presented vowels from all 264 talkers who participated in Experiment 1.

METHOD

The task was performed in a sound-treated booth during eight listening sessions lasting about 15 minutes each. The same 2,904 vowel tokens used in Experiment 1 were each presented once in random order through headphones at a comfortable listening level. In each case, the listener, a native speaker of General Canadian English, pressed a key on a computer keyboard marked with one of the phonetic symbols corresponding to the 11 vowels. During each session, an equal number of vowels from each category was presented. The distribution of talkers across sessions was random, except that the balance of native English (9%) and native Italian (91%) talkers was always the same. A practice set of five tokens was presented at the beginning of each session. Scores on these tokens were not used in the subsequent analyses.
RESULTS

Figure 5 shows the number of correct identifications (of a maximum of 24) by AOL group for each vowel. The results depicted here contrast strikingly with those presented in Figure 3. In general, the identification scores were quite high in comparison with the foreign accent scores. For the last-arriving AOL group (those who arrived at a mean age of 21.5), for instance, the total percentage of vowels rated as native-like in Experiment 1 was 33%, while the percentage of vowels identified correctly in the identification task was 70%.

For 10 of the vowels produced by this group, identification scores were well above 50%. Even in the cases of [i] and [u], two English vowels with no analogs in Italian, the identification rates for the last-arriving AOL group were only slightly lower (by 4% to 8%) than for the native English group. Furthermore, this group of Italian talkers produced identifiable [ɛ] tokens in 83% of instances. Only one vowel could be said to be poorly identified overall: in the case of [ʌ], only 25% of the tokens were identified correctly. Figure 6 illustrates the confusions for each target vowel for all 264 talkers. The majority of misidentified [ʌ] tokens were heard as [æ], though some were also heard as [o].

The identification data also strongly suggest that, in most cases, the late-arriving learners did not merely substitute the nearest Italian vowels for the target English vowels. Rather, these speakers appear to have learned to distinguish more than simply the seven Italian vowel categories. To estimate the number of vowel categories the speakers actually distinguished, we counted the number of different vowel labels that were assigned to the productions of each speaker. Table 2 gives the mean number of vowels distinguished by each AOL group, along with the minimum and maximum numbers for individual speakers in the group. It appears from this table that even the members of the latest-arriving group distinguished, on average, at least one more vowel than the total number in the typical Italian repertoire. The number of categories distinguished tended to decrease, however, with increasing AOL.

DISCUSSION

Taken together with the results of Experiment 1, the results of Experiment 2 indicate that the majority of the vowel production errors made by the Italian learners of English were probably within-category errors, rather than errors involving the substitution of a vowel from a category different from the intended one. Even those speakers who arrived in Canada the latest (i.e., at 21.5 years) were able to produce vowels that were correctly identified by the listener most of the time, though in Experiment 1 their productions were usually heard as foreign accented. Moreover, even the English vowels with no clear Italian analogs (viz., [u], [i], [æ], and [ɛ]) were produced at high rates of identification by the last-arriving group. Only one vowel ([ʌ]) was correctly identified less than 50% of the time.
Figure 5. The number of correct identifications of each vowel by one trained native English listener.
Figure 6. Confusions for each of the 10 vowels in Experiment 2. The numerals at the top of each bar indicate the total number of confusions for the vowel category indicated.
These findings indicate that the late-arriving learners did not merely substitute the nearest Italian vowels for the English targets. If they had done so, they would not have been able to distinguish more than the seven vowel categories of Italian. The data in Table 2, however, suggest that, on average, more than seven categories were recognized in their productions. It is possible that most of these learners had established new vowel categories in their second language, at least for those cases in which Italian has no analog for a particular English vowel. The fact that their vowel productions were heard as accented, however, suggests that such categories were not specified in the same way as those of the native English talkers.

CONCLUSIONS

This study was concerned with three major issues: whether accentedness in L2 vowel production would increase as a function of AOL, as was the case for accentedness in sentence-length utterances (Flege et al., 1995b); whether L2 vowels would show differing degrees of resistance to AOL effects; and whether L2 vowel intelligibility would show the same type of relationship to AOL as does vowel accentedness.

Pervasive AOL effects were found on the English vowel productions of the native Italian speakers examined here. In general, early-arriving learners tended to produce vowels that were rated as unaccented by the panel of 10 listeners, while the latest-arriving learners produced only a minority of the vowels without a perceptible accent. The AOL at which accents first became detectable varied considerably across both listeners and vowels.

The finding of between-vowel differences suggests that some vowels were easier than others for the Italian learners to acquire; however, not one of the 11 vowels could be said to have been “mastered” by the late arrivals as a group. No obvious explanations can be offered here for why some of the
vowels were produced in a more native-like way than others. In fact, the between-vowel differences were somewhat overshadowed by the differences between talkers in the extent to which native-like productions were achieved. This finding has important implications for the types of generalizations that may be possible in descriptions of L2 vowel acquisition. In particular, generalizations of the type, “Speakers of Italian produce vowel \(x\), but not vowel \(y\), in a native-like way,” cannot be made here because the between-vowel effects did not occur uniformly for all, or even for a large majority, of the learners. Instead, perhaps the strongest type of statement that can be made is that some vowels have a higher probability of being produced without an accent than others. Further research is essential for an understanding of the between-vowel and between-learner effects seen here.

It is possible, for instance, that the accurate production of L2 vowels is contingent on accurate perceptual representations, and that the effects seen here are the result of between-vowel differences in the difficulty with which perceptual accuracy can be achieved in combination with individual learner differences in ability to establish good perceptual representations.

The finding of relatively high correct identification scores (Experiment 2) requires further study with a larger group of listeners. However, the preliminary evidence obtained here suggests that many of the latest-arriving learners may have established partially accurate perceptual representations for those vowels without analogs in Italian, with the possible exception of \([\mathrm{A}]\). This perceptual awareness may have allowed them to produce these vowels in an identifiable, though accented, manner. If they had not had this awareness and had instead merely substituted a perceptually similar Italian vowel for each English vowel, they could not have distinguished more than the total number of vowels in the Italian inventory (viz., 7 vowels). The finding that the L2 vowel productions in this study were readily identifiable, despite being noticeably accented, parallels the findings of earlier studies demonstrating that accentedness and intelligibility are at least partially independent dimensions. It is evident that, in studies of L2 speech, predictions of “success” in productions of consonants, vowels, or larger units must be carefully formulated so that it is clear whether native-like production or intelligibility is at issue.

ACKNOWLEDGMENTS
This research was supported by the National Institute for Deafness and Other Communication Disorders Grant #DC00257 and by a fellowship from the Social Sciences and Humanities Research Council of Canada to the first author. The authors thank Mrs. J. Prosperine, the parishioners of St. Anthony’s Catholic Church, and V. Reddy for their assistance. We also acknowledge S. Russell, A. M. Schmidt, N. Takagi, and two anonymous reviewers for their useful comments and suggestions.

NOTES
1. One listener clearly misinterpreted the scale when rating the productions of \([\mathrm{l}]\), assigning low ratings to the native English talkers and high ratings to the late
AOL groups. For this reason, his data for [I] were excluded from the analyses reported here.

2. The Canadian English data for [u] show somewhat higher F2 frequencies than those believed typical of male speakers of American English. Peterson and Barney (1952), for instance, reported a value of 870 Hz, though Hillenbrand, Getty, Clark, and Wheeler (1995) observed a slightly higher mean frequency of 997 Hz. Our value of 1138 Hz is, however, in fairly good agreement with a value of 1171 Hz for two male speakers of Canadian English reported by Andruski (1990).

3. Because the role F3 plays in the identification of [3] distinguishes it from other vowels, the acoustic difference between [3] and other English vowels is presumably even greater than what would be evident in a two-dimensional (F1 x F2) plot.

4. The observed difference between rating and identification scores was not simply an artifact of comparing the identification scores from one listener against the rating scores of several listeners. An examination of the rating and identification scores (on all vowels except [A]) assigned by the one trained listener revealed that the number of vowels rated as native-like was 32% lower for the last-arriving group than for the native English talkers, while the number of vowels identified correctly was only 19% lower.

REFERENCES


Munro et al.: Age-of-learning effects on English vowels


Williams, D. (1971). A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27, 103–117.