Journal of Phonetics (1997) **25,** 437–470

Effects of experience on non-native speakers' production and perception of English vowels

James Emil Flege*

University of Alabama at Birmingham, U.S.A.

Ocke-Schwen Bohn

Aarhus University, Denmark

and

Sunyoung Jang

University of Pennsylvania, U.S.A.

Received 4th November 1996, and in revised form 6th May 1997

This study assessed the effect of English-language experience on non-native speakers' production and perception of English vowels. Twenty speakers each of German, Spanish, Mandarin, and Korean, as well as a control group of 10 native English (NE) speakers, participated. The non-native subjects, who were first exposed intensively to English when they arrived in the United States (mean age = 25 years), were assigned to relatively experienced or inexperienced subgroups based on their length of residence in the US ($M = 7.3$ *vs.* 0.7 years). The 90 subjects' accuracy in producing English /i ɪ ɛ æ/ was assessed by having native English-speaking listeners attempt to identify which vowels had been spoken, and through acoustic measurements. The same subjects also identified the vowels in synthetic *beat-bit* (/i/-/ɪ/) and *bat-bet* (/æ/-/ɛ/) continua. The experienced non-native subjects produced and perceived English vowels more accurately than did the relatively inexperienced non-native subjects. The non-native subjects' degrees of accuracy in producing and perceiving English vowels were related. Finally, both production and perception accuracy varied as a function of native language (L1) background in a way that appeared to depend on the perceived relation between English vowels and vowels in the L1 inventory.

1. Introduction

Most if not all individuals who learn a second language (L2) after the age of 15 years will speak it with a detectable foreign accent (Flege, Munro & MacKay, 1995). When taken together with the observation that foreign accent is cued in part by the incorrect

*Author to whom all correspondence should be addressed at: Department of Rehabilitation Sciences, VH 503, Birmingham, Alabama 35294, U.S.A.

0095-4470/97/040437 + 34 $25.00/0/jp 970052

production of vowels and consonants, and the hypothesis that a critical period exists for speech learning (Scovel, 1969; Patkowski, 1989), one might reasonably ask if adults are capable of learning to accurately produce and perceive the vowels in an L2.

If the capacity to learn L2 vowels is reduced or lost after a critical period, it may be impossible to develop truly effective vowel training techniques (Leather, 1990). However, some researchers believe that the perception of vowels and consonants remains malleable to some extent, even in adulthood (e.g., Best & Strange, 1992). And others have proposed that although L2 production accuracy may be limited by perceptual factors, the *capacity* to learn new forms of speech remains intact over the life span (Flege, 1981, 1988, 1995; see also Ioup, 1995). These views imply that, given sufficient native speaker input and the absence of perceptually-based limitations, adults can eventually learn to produce certain L2 vowels with native-like accuracy.

1.1. *Previous research*

A number of previous studies have examined the role of experience on adults' production and perception of L2 sounds. The research on consonants has suggested that learning is possible. For example, the results of studies with native Japanese speakers of English show that, over time, Japanese adults may learn to accurately produce English /r/ and /l/, neither of which occurs in Japanese (e.g., Yamada, Strange, Magnuson, Pruitt & Clark, 1994; Flege, Takagi & Mann, 1995). Similarly, their perception of these sounds improves with English-language experience (e.g., Yamada & Tohkura, 1991; Best & Strange, 1992; Flege, Takagi & Mann, 1996). It appears that the learning of consonants may take place over the course of years, or perhaps even decades. Non-native speakers' rate of learning, and perhaps their ultimate degree of success, may be influenced by the perceived similarity of English consonants to consonants in the L1. However, too little is known at present to ascertain if these generalizations apply across the entire L2 inventory of consonants in all possible phonetic contexts and syllable positions, or whether they will apply equally to L2 vowels (see, e.g., Cutler & Otake, 1994).

Non-native speakers' accuracy in producing English vowels is related inversely to their age of first extensive exposure to native-produced English. Flege (1992*b*) found that native Spanish subjects who arrived in the US as children produced English vowels that were identified as intended more often than were the vowels spoken by native Spanish subjects who arrived in the US in adulthood. Acoustic analyses revealed less spectral overlap between adjacent vowels spoken by the early than the late bilinguals. Munro, Flege & MacKay (1996) had native English-speaking listeners rate English vowels spoken by Italian/English bilinguals who arrived in Canada between the ages of 2 and 22 years. They observed a strong inverse correlation between vowel production accuracy and age of arrival.

Vowel production accuracy may vary among individuals who began learning their L2 in adulthood. Major (1987) examined vowels spoken by students in Brazil whose overall pronunciation of English was relatively good or poor. The /æ/ tokens spoken by students with mild foreign accents were more intelligible than the /æ/s spoken by students with stronger foreign accents. In a study examining the production of English vowels by Dutch students in the Netherlands, Flege (1992*a*) also found that English /æ/s spoken by students with mild accents were more intelligible than were the /æ/s spoken by students with stronger foreign accents. The basis for these foreign accent and vowel production differences is unknown. They do raise the question, however, of whether improvements in vowel production by adults occur simply as the result of naturalistic exposure to an L2, or whether some special talent is needed.

Jun & Cowie (1994) examined groups of Korean adults who had lived in the US for 1–5 or 26–31 years. The experienced Koreans produced English /ɪ/ more accurately than did the less experienced Koreans. This might lead one to conclude that an improved production of L2 vowels is inevitable given a sufficient amount of native-speaker input. However, Munro (1993) examined English vowels spoken by native Arabic adults who had lived in the US for 1–27 years ($M = 5.7$ years). Goodness ratings and acoustic measurements revealed that most of the native Arabic subjects produced many English vowels, even those without a direct phonetic counterpart in Arabic, differently than did native English speakers.

The study by Munro (1993) also yielded a null finding that is of interest here: the 23 Arabic subjects' length of residence in the US and their accuracy in producing English vowels were not correlated significantly. This finding cannot be interpreted, however, without knowing the rate at which L2 vowels are learned by adults. Just two subjects had lived in the US for less than 2.5 years. The native Arabic subjects' productions of English vowels might have improved as they gained experience in English. However, if most of the improvement in L2 vowel production takes place fairly rapidly (say, within a 2-year period), then a correlation between length of residence and vowel production accuracy could not have been observed in the Munro study because many subjects would have reached an asymptote in their vowel learning.

Finally, a study by Bohn & Flege (1992) was designed specifically to assess the effect of English-language experience on L2 vowel production accuracy. These authors examined native German subjects who had lived in the US either for less than 1 year ($M = 0.6$ years) or more than five years ($M = 7.4$ years). One vowel examined, /æ/, does not have a close counterpart in German whereas the other three vowels examined, /i ɪ ɛ/, do have German counterparts (see Section 1.4.). Acoustic measurements revealed that the relatively experienced Germans produced English /æ/ more accurately than did the relatively inexperienced subjects. An effect of English-language experience was not observed for the other vowels, however (see also Flege, 1987).

1.2. *The present study*

The aim of this study was to extend the study by Bohn & Flege (1992) to native speakers of three typologically diverse languages (Spanish, Mandarin, and Korean) in order to better understand the role of experience on adults' production of L2 vowels. As described in Section 1.4., the vowel systems of these additional languages differ from English in interesting ways. Each L1 group, in turn, was comprised of subgroups of subjects differing in English-language experience (indexed by length of residence in the US, as in the original study). The accuracy with which relatively experienced and inexperienced subjects produced English vowels was assessed through acoustic measurements, and through an intelligibility test.

We also extended the Bohn & Flege (1990) vowel perception study to the three new groups of subjects, which permitted an assessment of the relation between vowel production and perception accuracy. After producing English /i ɪ ɛ æ/, our subjects identified the members of two synthetic continua. The vowels in one continuum ranged from *beat* (/i/) to *bit* (/ɪ/); those in the other ranged from *bat* (/æ/) to *bet* (/ɛ/). In both continua, vowel quality was varied in 11 steps (by changing the frequencies of the first and second formants) and vowel duration was varied in three steps, yielding 33 stimuli per continuum. The NE subjects examined by Bohn & Flege (1990) based their identification of vowels (/i/ *vs.* /ɪ/, /ɛ/ *vs.* /æ/) primarily on spectral quality but, as expected (Gottfried, Miller & Payton, 1990), they also showed a small but significant effect of vowel duration.

If the consonant results mentioned earlier apply to vowels, then one would expect the identification responses of the experienced non-native subjects to resemble NE subjects' more closely than those of the inexperienced non-native subjects. Two vowel studies indirectly support this inference. Gottfried (1984) found that native English speakers of French discriminated vowels in French/tVt/ words more accurately than did English monolinguals. Gottfried & Beddor (1988) showed that native English speakers of French resembled French native speakers more closely than did English monolinguals in identifying the members of a French /o/-/ɔ/ continuum.[1]

One might infer from the consonant results summarized above that non-native subjects' accuracy in producing and perceiving English vowels will be related. Indirect support for this inference comes from a study by Rochet (1995), who presented synthetic vowels to several groups of subjects. Native Portuguese subjects tended to hear and to imitate French /y/s as /i/ whereas native English subjects tended to hear, and also to imitate, the same vowels as /u/. From this, Rochet concluded (1995, p. 404) that some vowel production errors may "be the consequence of the target phones having been assigned to an L1 category" rather than being motoric in nature. If this conclusion applies to adult learners of an L2 (Flege, 1988), then one might expect individuals who do not distinguish English /i/-/ɪ/ or /ɛ/-/æ/ perceptually on the basis of spectral (or temporal) cues to not *produce* spectral (or temporal) differences between these English vowels.

1.3. *The relation of* L1 *and* L2 *vowels*

Any two languages being compared might differ in terms of the number of contrastive vowels they possess (Maddieson, 1984), in how the vowels they can be said to "share" are realized phonetically, and/or in contextually induced variation (i.e., allophony). Adult beginners typically interpret L2 vowels as instances of the closest L1 vowel, and produce them accordingly. Knowing how, or if, L1 and L2 vowels are related perceptually to one another thus provides an important determinant of how inexperienced adults learners will produce L2 vowels. However, if speech perception processes remain malleable (Best & Strange, 1992) or if L2 learners establish new phonetic categories for certain L2 vowels (Flege, 1992*b*; 1995), then the perceived relation between L1 and L2 vowels may change during L2 acquisition. Such changes in perception may, in turn, engender changes in vowel production.

Researchers have used a variety of techniques to infer how the vowels in two languages are related perceptually. L2 vowels represented by phonetic symbols not used to transcribe any L1 vowel have been classified as "new." L2 vowels represented by the same symbol as that used for some L1 vowel have been classified as "identical" or "similar" (Flege, 1988). Symbol-based comparisons have been supplemented by finer-grained auditory and articulatory descriptions provided by phoneticians familiar with both members of an L1-L2 pair (e.g., Collins & Mees, 1984). Although valuable, this

[1] They did so by reducing their use of temporal cues to vowel identity, which appear to be more important in English than in French.

approach may not yield valid conclusions concerning the perceived relation between L1 and L2 vowels (Rochet, 1995). One problem with this approach is that transcription practices and symbolization may vary across languages, and vowels transcribed using the same IPA symbol (e.g., the /u/s of Korean and English; Yang, 1996) may differ systematically. One might also question the assumption that phoneticians' perceptual judgments typify those of untrained adult L2 learners, or that L1 and L2 vowels will be related perceptually in a constant manner as learners gain experience in their L2.

Another technique used to infer how L1 and L2 vowels are related perceptually is to plot first and second formant (F_1, F_2) measurements in a two-dimensional space (e.g., Flege, 1992*b*). One problem with doing so is that potential artifacts abound, including uncontrolled differences in the phonetic context in which the L1 and L2 vowels being compared were produced, and cross-language differences in degree of stress, speaking rate, speaking clarity or formality, or vocal tract size. Another problem is that two-dimensional plots, by their very nature, fail to display other dimensions (e.g., f_0 and higher formants, duration, formant movement patterns) that might influence the perceived similarity of vowels drawn from two languages (e.g., Di Benedetto, 1989). Also, the acoustic comparison approach makes the implicit assumptions, which are probably incorrect,[2] that L2 learners will use the same acoustic dimensions and weights as native speakers of the target L2 and that their use of features will not change during L2 acquisition.

At present, the only reliable way to determine the perceived relation, if any, between vowels in two languages is empirically, through a cross-language mapping study (Ingram & Park, 1997). Two procedures are needed (see, e.g., Schmidt, 1996). First, phonetically untrained individuals with little or no L2 experience must identify multiple natural tokens of all L2 vowel categories in terms of L1 vowel categories (using carefully chosen keywords, or unambiguous orthographic representations). Later, they must rate multiple natural tokens of each L2 vowel category for goodness as instances of each relevant L1 category (determined from the foregoing identification experiment). Possible experience-based changes in perceived vowel relatedness can be determined by comparing the results for highly experienced L2 learners to those obtained from beginners.

The procedures just outlined are likely to show that instances of certain L2 vowel categories are identified consistently as "good" instances of one L1 category. Other L2 vowels may also be identified consistently in terms of a single L1 category, but not be judged as good instances of that category (see, e.g., Polka & Bohn, 1996; Ingram & Park, 1997). Still other L2 vowels may be identified as instances of multiple L1 categories, but not be rated as good instances of any of those categories.[3] L2 vowels showing these response patterns might be classified as "identical" "similar", and "new", respectively (Flege, 1988). It has been hypothesized (Flege, 1988; 1992*a*, *b*) that adult learners will ultimately produce new L2 vowels more accurately than similar L2 vowels because they are more likely to establish additional phonetic categories for new vowels (but see Flege,

[2] Munro (1992) found that when asked to make perceptual judgments pertaining to French vowels, native Arabic subjects gave greater weight to vowel duration than did native English speakers. The difference appears to have arisen because Arabic but not English possesses vowel distinctions (phonemic length contrasts) based primarily on vowel duration. See also Fox, Flege & Munro (1995) regarding the use of features in perceiving vowels.

[3] It is also possible, of course, that the perceived relation of an L2 vowel will be found to vary importantly as a function of the phonetic context in which the L2 vowel tokens under examination have been produced (see, e.g. Steensland, 1981).

1995, for an update of the hypothesis regarding the influence of degree of perceived cross-language vowel similarity on L2 vowel production accuracy).

1.4. *Cross-language differences*

A secondary aim of this study was to assess the effect of L1 background on the production of English vowels by native speakers of German, Spanish, Mandarin, and Korean. The vowel systems for these four L1s differ from that of English in unique ways. Cross-language phonetic interference plays an important role in how non-natives produce and perceive English vowels, so the possibility existed that subjects from the four L1s might learn English vowels at different rates, or to different extents. We will therefore begin by summarizing how the four L1s differ from English.

Spanish. The five vowels of Spanish, /i e a o u/, exhibit less formant movement than do the closest English vowels and are apparently not distinguished from one another by vowel duration differences, as is the case for English vowels such as /i/-/ɪ/ (Harris, 1969). Acoustic analyses indicate that realizations of Spanish /i/ may occupy a portion of vowel space that is occupied by English /i/, and to a lesser extent English /ɪ/. Not surprisingly, auditory evaluation suggests that Spanish /i/ is closer to English /i/ than /ɪ/ (Stockwell & Bowen, 1965). Acoustic analyses reveal that Spanish /e/ realizations occupy a portion of vowel space occupied by English /ɪ/ and, to a lesser extent English /ɛ/ (see Flege, 1991, Fig. 1). Spanish /e/ is realized as a monophthongal [e] in most contexts, but it may be realized as an [ɛ]-quality vowel (much like that in English *bet*) when found in Spanish syllables closed by a consonant other than /s/ or /z/ (Dalbor, 1980).

L2 error analyses (see Flege, 1991, for a summary) indicate the following pattern of misidentification of English vowels spoken by native speakers of Spanish: intended /ɪ/ → [i], /ɛ/ → [e], /æ/ → [a]. Flege (1991) had Spanish monolinguals use the letters ⟨i e a o u⟩ to label the vowels in naturally produced English words. Realizations of English /i/ and /ɪ/ were classified most often as instances of Spanish /i/, realizations of /ɛ/ as instances of Spanish /e/, and realizations of English /æ/ as instances of Spanish /a/. Spanish/English bilinguals, on the other hand, showed greater response variability and were more likely to label English /ɪ/ tokens as "not a Spanish vowel".

Mandarin. About a dozen different vowel segments appear on the surface in Mandarin (not counting diphthongs). The number of underlying segments varies according to analysis, ranging from three to as many as 12, with five or six being most common (Cheng, 1973; Light, 1976; Li & Thompson, 1981; Maddieson, 1984). Howie (1976) analyzes Mandarin as having six vowels /i y ə ɑ r u/. According to his analysis, Mandarin /i/ may be realized as an [ɪ]-quality vowel in certain contexts (e.g., following /ə ɑ/, preceding a sequence of /ə/ + nasal consonant), /ə/ may be realized as an [ɛ]-quality vowel (in open syllables preceded by /i y/) or as an [eᶦ]-quality vowel (in /u_i/), and /ɑ/ as an [æ]-quality vowel (in the context of /i_n/ and /y_n/). Howie's acoustic measurements indicate that Mandarin /i/ but not /ɑ/ shows a great deal of formant movement when produced with certain tones. According to Norman (1988, p. 142), Mandarin /ɑ/ may have [ɛ] and [æ]-quality variants preceding /n/, depending on whether or not the preceding consonant is palatalized. Although vowel duration may be used as a cue to tone in Mandarin (see Lin & Repp, 1989, for Taiwanese), it is apparently not used to distinguish vowels in Mandarin.

Korean. Standard Korean has 10 vowels, /i e ɛ a o ø u y ʌ ɨ/. Acoustic measurements (Yang, 1996) suggest that the Korean /i/ is closer to English /i/ than /ɪ/, and that Korean

/ɛ/ is closer to English /ɛ/ than /æ/. However, the Korean vowel /ɛ/ is sometimes symbolized as /æ/ because it is often produced with [æ]-like variants, especially by individuals for whom a mid front vowel, /e/, is still phonemic (Hong, 1988; David Silva, personal communication).

Kim (1972) asked Koreans with little or no knowledge of English to write the vowels and consonants in 60 aurally presented English words using the Korean writing system. English /i/ was usually transcribed using the symbol for Korean /i/. English /ɪ/ was identified most often with Korean /ɨ/, English /ɛ/ with Korean /ɛ/ (or /e/), and English /æ/ also with Korean /ɛ/. Colhoun & Kim (1976) claimed that, from the standpoint of Korean, English /ɪ/ and /æ/ are "totally new" vowels. Robson (1982) noted that Koreans living in the US may realize both English /i/ and /ɪ/ as [i]-quality vowels, and they may realize English /æ/ as an [ɛ]-quality vowel (see also Ingram & Park, 1997).

Korean is traditionally analyzed as having a distinction between phonemically long and short vowels (e.g., [kil] 'road' *vs.* [kiːl] 'long'), with long vowels generally more peripheral in the vowel space than short ones (Lee & Zhi, 1987). However, phonemic length distinctions are subject to idiosyncratic, lexical and dialectal variation in Korean (e.g., Martin, 1951; Colhoun & Kim, 1976), and are not maintained by many young Koreans from Seoul (see, e.g., Magen & Blumstein, 1993).

German. Moulton (1962) noted that English vowels show more formant movement than do corresponding German vowels, and that the English distinction between /i/-/ɪ/ is based more on formant movement than duration whereas the reverse holds true for German. Bohn & Flege (1992) reviewed literature suggesting that the German vowels /i ɪ ɛ/ may be somewhat higher in vowel space than are English /i ɪ ɛ/, and that a larger duration difference exists between German /i/-/ɪ/ than between English /i/-/ɪ/. Their acoustic comparison of vowels produced by native speakers of English and German confirmed these observations, and showed that realizations of English /æ/ occupy a portion of vowel space not occupied by vowels spoken by the German subjects they examined. In a listening test using keywords as response alternatives, Bohn & Flege (1990) found that subjects from northern Germany usually identified tokens of English /i/, /ɪ/, and /ɛ/ as instances of German /i/, /ɪ/, and /ɛ/, respectively. The same subjects identified English /æ/ in terms of German /ɛː/, but with low confidence ratings.

1.5. *Expected findings*

The information just summarized, although rich, is nonetheless insufficient to allow us to generate specific predictions from the hypotheses of either the Perceptual Assimilation Model (Best & Strange, 1992) or the Speech Learning Model (Flege, 1992*b*).[4] Still, the information reviewed does lead to some general expectations, which will be stated here.

[4] Both PAM and the SLM take the view that many L2 production errors have a perceptual origin. PAM can generate predictions concerning the discriminability of vowels from an unknown foreign language or vowels encountered in early stages of L2 acquisition. The SLM can generate predictions concerning highly experienced learners' accuracy in producing L2 vowels. According to PAM, learners perceive the gestures used to form sounds in the L2. For the SLM, the objects of cross-language perception are vowel and consonant segments as perceived via a set of phonetically relevant features. Both models require empirical evidence bearing on the perceived relation, if any, of phonic elements in the L2/foreign language and those in the L1 before they can generate specific predictions. Unfortunately, the kind of data these models would require to generate predictions concerning how English /i ɪ ɛ æ/ will be perceived, and eventually produced, by native speakers of German, Spanish, Mandarin, Korean, is unavailable.

Of the four L1s considered here, only German possesses a phonemic contrast between /i/ and /ɪ/. The foregoing analyses suggest, on balance, that the Germans will be more successful in producing these vowels than will the native speakers of Spanish, Mandarin, and Korean. It is not possible to predict the relative degree of difficulty for speakers of the three L1s because the perceived relation between English /i/ and /ɪ/ (in a /b_t/ context) to vowels in the three L1s is unknown. However, we suspect that the Koreans might have the greatest difficulty producing (and perhaps perceiving) the distinction between English /i/-/ɪ/. Koreans are taught in school that these English vowels are distinguished by duration, not spectral quality. (Dictionaries mark the distinction in the same way.) The young Koreans who participated in our study were probably exposed to high front Korean vowels differing in duration ([i], [iː]) in the speech of older Koreans, but they themselves were unlikely to maintain such a phonemic distinction. In an unpublished study, Yang and Flege had some of the Koreans who participated in the present study produce minimally paired Korean words differing in vowel length. When these words were later played back to the subjects for identification, they identified words with /i/ and /iː/ at rates that did not significantly exceed a chance level.

None of the four L1s considered here has /æ/. The German subjects were not expected to have the same advantage for English /ɛ/-/æ/ as for /i/-/ɪ/ but, once again, it is not possible to predict relative degree of difficulty. However, we speculate that the native Korean subjects will have the greatest difficulty producing (and perhaps also perceiving) a distinction between English /ɛ/ and /æ/. This is because Korean has a vowel phoneme realized with both [ɛ] and [æ]-quality variants (and symbolized as either /ɛ/ or /æ/). Further, the German subjects might be aided in learning to distinguish /ɛ/ from /æ/ because German /ɛ/ contrasts with /ɛː/ in German, which might sensitize Germans to the duration difference between English /ɛ/ and /æ/, and because certain German dialects have an [æ]-quality vowel.

Finally, the native Spanish subjects may in turn be more successful in distinguishing /ɛ/-/æ/ than the native German subjects. If the Spanish subjects identified realizations of English /ɛ/ as instances of Spanish /e/ and realization of English /æ/ as instances of Spanish /a/, then they might produce even larger spectral differences between English /ɛ/-/æ/ than do NE subjects. The difficulty of English /ɛ/-/æ/ for the Mandarin subjects was uncertain because it was not known whether they would interpret these English vowels as instances of one Mandarin vowel (*viz.* /ɑ/) or two (/ə/, /ɑ/). This uncertainty was compounded by the fact that the context in which the English vowels were examined, *viz.* /b_t/, does not exist in Mandarin (nor in Spanish).

The results will be presented in three parts. In Section 2, vowel production accuracy is evaluated in an intelligibility test. In Section 3, vowel production accuracy is evaluated through acoustic measurements. Section 4 presents the results of a vowel perception experiment examining the subjects' forced-choice identification of the vowels in two synthetic continua.

2. Intelligibility test

Native and non-native subjects' accuracy in producing English /i ɪ ɛ æ/ was assessed by presenting their productions of these vowels to native English-speaking listeners in a forced-choice identification test. The subjects and speech materials will be described before presenting the procedures used and the results obtained.

2.1. *Method*

2.1.1. *Subjects*

Ninety subjects recruited through a newspaper ad or personal contact were tested in Birmingham, Alabama, after having passed a pure-tone hearing screening. The 10 subjects (5 males, 5 females) in the native English (NE) control group were native speakers of American English having a mean age of 27 years (SD $= 7$). As summarized in Table I, eight non-native groups (each with 5 males and 5 females) were established. Each consisted of individuals who had arrived in the US at the age of 14 years or older ($M = 25$ years) and spoke no language other than their L1 (German, Spanish, Korean, or Mandarin) and English. The non-natives were assigned to "experienced" or "inexperienced" subgroups based on their length of residence (LOR) in the US ($M = 7.3$ for experienced *vs.* 0.7 years for inexperienced) [$F(1, 1.72) = 113.9$, $p < 0.01$].

The LOR differences between the experienced and inexperienced native speakers of German (7.4 *vs.* 0.6 years), Spanish (9.0 *vs.* 0.4), Mandarin (5.4 *vs.* 0.9) and Korean (7.3 *vs.* 0.8 years) were roughly the same, so the L1 $\times$ Experience interaction was non-significant [$F(3, 72) = 1.72$]. The age difference between the experienced and inexperienced subjects (30.2 *vs.* 37.6 years) and the number of years of formal education in English they had received before coming to the US (7.2 *vs.* 6.7 years) were non-significant [$F(1, 72) = 5.71, 1.37$], but the experienced subjects arrived in the US at a significantly earlier age (22.9 *vs.* 26.9 years) [$F(1, 72) = 17.0$, $p < 0.01$] and estimated using English more (68% *vs.* 52%) than did the inexperienced subjects [$F(1, 72) = 14.3$, $p < 0.01$].

These last two differences represent a confound with LOR, for both might be expected to promote more a native-like performance in English on the part of the experienced non-native subjects. A series of ANOVAs was carried out to assess the effect of L1 background (4 levels) and English-language experience (2 levels) on the non-native subjects' age of arrival (AOA), in the US their percentage use of English (PER), amount of formal education in English (EDUC), and chronological age (AGE). The two-way interactions

TABLE I. Characteristics of the four relatively experienced and inexperienced groups of non-native speakers, each with five male and five female subjects. Standard deviations are in parentheses

Native language	L2 experience	Chron. age	Res. in US[a]	Arrival age[b]	Educ. in English[c]	% Use[d]
German	Experienced	33(6)	7.4(2.5)	25(5)	7(2)	87(15)
	Inexperienced	28(6)	0.6(0.4)	28(6)	8(2)	67(26)
Spanish	Experienced	28(6)	9.0(5.2)	20(5)	7(4)	75(11)
	Inexperienced	26(3)	0.4(0.1)	26(4)	6(3)	68(24)
Mandarin	Experienced	28(3)	5.4(2.4)	23(4)	6(2)	51(22)
	Inexperienced	28(3)	0.9(0.6)	27(3)	7(2)	32(15)
Korean	Experienced	31(7)	7.3(4.5)	24(5)	9(2)	59(9)
	Inexperienced	27(2)	0.8(0.3)	27(2)	10(3)	43(20)
	Mean (SD) Range	29(5)	4.0(4.3)	25(5)	8(3)	60(24)
		20–45	0.2–23	14–38	2–16	20–100

[a] Res. in US; Length of residence in the US, in years.
[b] Arrival Age, Age of arrival in the US, in years.
[c] Educ. in English, Years of formal education in English prior to arrival in the US.
[d] % Use, Self-estimated percentage daily use of English.

in these analyses never reached significance ($p > 0.10$). However, we deemed it prudent to use AOA, PER, and EDUC as covariates in analyses of covariance (ANCOVAs) carried out to examine differences between non-native groups defined on the basis of length of residence in the US.

The subjects representing the four L1s did not differ in terms of how long they had lived in the US [$F(3, 72) = 1.04$], how old they were upon their arrival in the US (German—26.5, Spanish—22.8, Mandarin—24.9, Korean—25.3 years) [$F(3, 72)$ 2.63], or their age when tested (German—30.6, Spanish—27.3, Mandarin—28.3, Korean—29.4 years) [$F(3, 72) = 1.82$]. However, significant main effects of L1 were found for formal education in English [$F(3, 72) = 4.69, p < 0.01$] and self-estimated percentage daily use of English [$F(3, 72) = 16.1, p < 0.01$]. Tukey's post-hoc tests revealed that both the native German and Spanish subjects reported using English more often ($M = 77\%$, 72%) than did the native Korean and Mandarin subjects ($M = 51\%, 42\%; p < 0.01$), and that the native Korean subjects had more formal education in English than did the native Spanish subjects (9.2 *vs*. 6.6 years; $p < 0.01$).

We would like to have tested non-natives drawn from a single, well-defined variety of their L1 who had been exposed to a single, well-defined variety of English L2. However, resource limitations made this unfeasible. Six native German subjects were from northern Germany, nine were from central Germany, and five were from southern Germany. None were judged to speak German with a strong regional accent as judged by the second author (a native speaker of German). The native Spanish subjects came from Central and South American countries [Colombia (8), Puerto Rico (2), Nicaragua (2), Chile (2), Bolivia (1), Honduras (1), Venezuela (1), Guatemala (1), Argentina (1)]. The Chinese subjects came from various places [Taipei (5), Beijing (2), and most of the remaining 13 from northern China). However, a native Mandarin experimenter, Chipin Wang, determined that Mandarin was the native language of each. All 20 Korean subjects came from the Seoul region and spoke standard Korean (as judged by the third author). Our NE subjects came from a variety of places in the US [Alabama (5), South Dakota (1), Tennessee (1), Ohio (1), Washington, DC (1), Missouri (1)] but none spoke English with a marked regional dialect (as judged by the first author, a native speaker of Midwestern American English).

2.1.2. *Procedure*

Each subject was tape recorded (Marantz Model PMD 420) in a sound booth after responding to a language background questionnaire. The subjects read four randomized lists of words in a carrier phrase (*I will say*_). One list contained five tokens of the test word *beat* and two tokens of seven other words with /i/. The other three lists contained a test word with the vowel /ɪ/, /ɛ/, or /æ/ (i.e., five tokens of *bit*, *bet*, or *bat*) and other words containing the vowel in the test word (e.g., *sit*, *pick*, and *did* on the /ɪ/ list). We had the subjects read separate lists for each vowel to reduce production errors due to orthographic confusions but, as discussed below, this procedure was not entirely successful. Finally, after completing the production task, the native Spanish and German subjects used a 9-point scale to rate the words for degree of familiarity.

The last five of six available repetitions of each test word were digitized at 10 kHz. Some groups were more likely to prevoice the initial /b/ in the test words than others so, to avoid creating a bias, any voicing prior to the /b/ release was edited out. The 1800 words (90 subjects $\times$ 4 vowels $\times$ 5 tokens) were then normalized for peak amplitude.

The edited words were presented to three native English-speaking listeners (one male, two females). The listeners were young adult residents of Birmingham, Alabama with normal hearing. One was from Birmingham, one was from a Birmingham suburb, and the third was from Wisconsin. The listeners were chosen primarily on the basis of their availability for participation in sessions spread over a 1-year period. None of the listeners had training in speech or language, or had participated previously in speech research.

The listeners were tested one at a time in a sound booth in 15 one-hour sessions each. The 20 words spoken by each of the 90 subjects were randomly presented over headphones six times each in a separate block. The listeners were told to identify each vowel using one of seven keywords and to guess if unsure. They indicated their response by pushing one of seven buttons on a response box, each marked with a keyword. The interval between each response and the presentation of the next word was 1.0 sec. The keywords used to identify vowels consisted of the four test words (*beat*, *bit*, *bet*, *bat*) and three other words containing a potentially confusable vowel (*viz. bait*, *but*, *bottle*)[5].

Vowels spoken by subjects from the four L1s were presented in the order in which the groups were recorded: first Spanish, then German, Mandarin, and Korean. The order of presentation of the subjects within each L1 background was counterbalanced across the three listeners, however. Vowels spoken by the NE subjects were interspersed among those of the native Spanish subjects in the first five sessions. The vowels spoken by two NE subjects were later presented a second time during the subsequent 10 sessions to determine that the listeners continued to perform the identification task in the same way. Analyses revealed that they continued to identify the NE subjects' vowels at near perfect rates. (Just the first set of data for the NE subjects will be presented here.)

The "intelligibility" score obtained for each subject's production of /i/, /ɪ/, /ɛ/ and /æ/ was the percentage of times that each vowel was identified correctly out of 75 judgments (3 listeners $\times$ 5 tokens $\times$ 5 repetitions). When a vowel was misidentified, the keyword that was used provided an indication of how the vowel was mispronounced. Therefore, the percentage of times that incorrect responses were given for each target vowel was also calculated.

2.2. *Results*

2.2.1. *Gender*

To determine if gender influenced the non-native subjects' accuracy in producing English vowels, as seems to be the case for degree of global foreign accent (see Flege, Munro & MacKay, 1995), the intelligibility scores were submitted to an ANCOVA in which Gender (2 levels) and L1 (4 levels) served as between-subjects factors and Vowel (4 levels) served as a within-subjects factor. Three language background questionnaire items mentioned earlier were used as covariates (AOA, age of arrival in the US; EDUC, amount of formal education in English; PER, self-estimated use of English). In this analysis, LOR (length of residence in the US) was also used as a covariate. The difference in the males' and females' scores (75.5 *vs.* 71.6% correct[6]) was non-significant [$F(1, 68) = 1.03$,

[5] Preliminary analyses suggested that these keywords adequately represented what the listeners might hear regardless of whether the four target English vowels were produced correctly or incorrectly.

[6] The scores just presented were adjusted for variation in the four covariates mentioned earlier. They differed little from the raw scores for the males (76.8%) and females (70.4%). The scores presented in subsequent reports of ANCOVAs will also be adjusted scores.

$p > 0.10$] and Gender did not interact significantly with any other factor, so the data for the males and females were pooled in all subsequent analyses.

2.2.2. *Mean intelligibility scores*

The mean intelligibility scores obtained for vowels spoken by the nine groups of subjects are presented in Table II. The NE subjects' vowels were identified in virtually every instance (100% correct for /i ɪ æ/, 99.3% for /ɛ/). The non-native subjects' vowels were often misidentified but, due to a lack of variance in the NE subjects' scores, native *vs.* non-native differences could not be tested statistically.

The intelligibility scores were submitted to an ANCOVA in which Experience (2 levels) and L1 (4 levels) served as between-subjects factors, Vowel served as a within-subjects factor, and AOA, EDUC, and PER (but not LOR) served as covariates. The small difference between the experienced and inexperienced subjects (75.9 *vs.* 71.3% correct) was non-significant [$F(1, 69) = 1.26$, $p > 0.10$], but a marginally significant Experience × Vowel interaction [$F(3, 216) = 3.30$, $p > 0.02$] was obtained. Tests of simple main effects revealed that the experienced and inexperienced subjects did not differ significantly for /i/

TABLE II. **Boldface** values: the mean percentage of times that English vowels spoken by the subjects in nine groups were identified as intended. *Italicized* values: the percentage of times that other vowels were heard instead of the intended target vowels. (Only responses accounting for more than 2% of the responses are shown.)

		English Target Vowels			
Native language	L2 experience	/i/ in *beat*	/ɪ/ in *bit*	/ɛ/ in *bet*	/æ/ in *bat*
English	—	**/i/-100**	**/ɪ/-100**	**/ɛ/-99**	**/æ/-100**
German	Experienced	*/i/-100*	*/ɪ/-100*	*/ɛ/-77* */ɪ/-22*	*/æ/-66* */ɛ/-34*
	Inexperienced	**/i/-100**	**/ɪ/-99**	**/ɛ/-88** */æ/-12*	**/æ/-53** */ɛ/-47*
Spanish	Experienced	**/i/-57** */ɪ/-42*	**/ɪ/-61** */i/-39*	**/ɛ/-99**	**/æ/-73** */ɑ/-13* */ʌ/-10* */ɛ/-4*
	Inexperienced	**/i/-69** */ɪ/-24* */ɛ/-7*	**/ɪ/-51** */i/-49*	**/ɛ/-91** */ɪ/-9*	**/æ/-70** */ʌ/-30*
Mandarin	Experienced	**/i/-84** */ɪ/-16*	**/ɪ/-90** */i/-6* */ɛ/-4*	**/ɛ/-63** */æ/-27* */eˡ/-29*	**/æ/-77** */ɛ/-23*
	Inexperienced	**/i/-80** */ɪ/-19*	**/ɪ/-83** */i/-16*	**/ɛ/-60** */æ/-13* */eˡ/-225*	**/æ/-58** */ɛ/-31* */eˡ/-10*
Korean	Experienced	**/i/-60** */ɪ/-40*	**/ɪ/-92** */i/-8*	**/ɛ/-81** */æ/-19*	**/æ/-43** */ɛ/-57*
	Inexperienced	**/i/-75** */ɪ/-25*	**/ɪ/-61** */i/-39*	**/ɛ/-81** */æ/-19*	**/æ/-18** */ɛ/-82*

(76 *vs.* 80%), /ɛ/ (79 *vs.* 81%) or /æ/ (61 *vs.* 53%; $p > 0.10$), but the difference for /ɪ/ (88 *vs.* 71%) was marginally significant [$F(1, 75) = 6.08, p = 0.02$].

Differences between speakers of the four L1s varied as a function of vowel. The significant L1 × Vowel interaction that was obtained [$F(9, 216) = 9.13, p < 0.01$] was explored through tests of simple main effects. The L1 factor was significant for /i/ [$F(3, 73) = 5.26$], /ɪ/ [$F(3, 73) = 10.8$] and /ɛ/ [$F(3, 73) = 5.71, p > 0.01$] and marginally significant for /æ/ [$F(3, 73) = 3.75, p = 0.014$]. Tukey's post-hoc tests revealed that the /i/ scores were significantly higher for the German than Spanish subjects (97% *vs.* 63%); the /ɪ/ scores were significantly higher for the native Korean, Mandarin, and German subjects than for the native Spanish subjects (81%, 90%, and 95% *vs.* 53%); the /ɛ/ scores were significantly higher for the native Spanish than Mandarin subjects (95% *vs.* 61%); and the /æ/ scores were higher for the native Mandarin than Korean subjects (72% *vs.* 40%) (all p's < 0.01).

The differences between the four vowels varied as a function of L1. The simple effect of Vowel was significant for each L1 group [F-values ranging from 4.6–15.9, df $= 3, 57$, $p = 0.01$]. Tukey's post-hoc tests ($\alpha = 0.01$) revealed that the Koreans' productions of /i/, /ɪ/, and /ɛ/ were identified correctly more often than were their /æ/ productions (67%, 77%, and 81% *vs.* 30% correct). The Germans' productions of /i/, /ɪ/, and /ɛ/ were also identified more often than their /æ/s (99.9%, 99.2%, and 82% *vs.* 58% correct). The Spanish subjects' /ɛ/s were identified correctly significantly more often than their /i/s or /ɪ/s (95% *vs.* 63% and 56% correct), whereas the Mandarin subjects' /ɪ/s were correctly identified more often than were their /ɛ/s (86% *vs.* 61%).

2.2.3. *Patterns of misidentification*

The way in which the target English vowels were misidentified is also summarized in Table II. The native German subjects' /i/s and /ɪ/s were almost always identified as intended. However, bi-directional confusions (i.e., /i/s heard as /ɪ/ and /ɪ/s heard as /i/) were evident for 11 native speakers of Spanish, six native speakers of Korean, and five native speakers of Mandarin. Bi-directional confusions for /ɛ/-/æ/ were evident for 12 native speakers of Korean (5 experienced, 7 inexperienced), five native speakers of Mandarin (3 experienced, 2 inexperienced) and three native speakers of German (all inexperienced). However, the range of mis-identifications was wider for /ɛ/ and /æ/ that it was for /i/ and /ɪ/.

The experienced German subjects' productions of English /ɛ/ and /æ/ tended to be heard as distinct vowels. Their apparent substitutes for these vowels (/ɪ/ and /ɛ/, respectively) both appeared to be higher in vowel space than the targets. Four native Mandarin subjects' productions of /ɛ/ were heard as /eˈ/ (i.e., as a vowel higher in vowel space than the target), but the remaining Mandarin subjects' /ɛ/s were heard as a lower vowel, /æ/. Finally, the native Spanish subjects differed from the other non-native subjects in that their /æ/s were often heard as /ɑ/. This pattern of mid-identification suggests that the native Spanish subjects produced a vowel (perhaps Spanish /a/) that was posterior in vowel space to the English target vowel.

2.3. *Discussion*

The NE subjects' vowels were identified as intended more often than were the non-natives' vowels (100% *vs.* 74% correct). Although the experienced non-natives' vowels were identified correctly slightly more often than were the relatively inexperienced non-natives', a (marginally) significant difference between these subgroups was observed for just one vowel, /ɪ/. This may have been due in part to our method of eliciting vowel production and/or to the method used here for assessing vowel intelligibility. Recall that the subjects read four lists, each containing English words with just one vowel. Later, the digitized words spoken by each subject were presented to listeners in a separate block. This may have allowed the listeners to "calibrate" on a non-native subject's system of (non-English) vowel contrasts. For example, a subject's [ɛ]-quality rendition of /æ/ might have been identified as /æ/ if she realized /ɛ/ as a vowel that was higher still in vowel space.

Some important differences were observed between subjects from different L1 backgrounds. We anticipated that the German subjects would produce English /i/ and /ɪ/ more accurately than subjects from the other L1s, and this was indeed the case (Germans—100% *vs.* Spanish—60%, Mandarin—84%, Korean—72%). The native speakers of Spanish, Mandarin, and Korean, whose L1s do not possess a contrast between /i/ and /ɪ/, showed bi-directional confusions (/i/s heard as /ɪ/, and vice versa) whereas the Germans did not.

In the Introduction, we speculated that the Koreans might produce a less effective contrast between English /i/ and /ɪ/ than speakers of the other L1s if they over-relied on vowel duration (that is, if they treated English /i/ and /ɪ/ as a phonemic length distinction). However, this either did not happen (see Ingram & Park, 1997), or else some of the Koreans produced spectral differences in addition to length differences, as is apparently the case for the Korean /i/ *vs.* /iː/ distinction (Lee & Zhi, 1987).

Also as expected, the Germans did not enjoy a clear advantage over subjects from the other L1s in producing /ɛ/ and /æ/ (Germans—71%, Spanish—84%, Mandarin—65%, Koreans—56%). The Koreans—and some inexperienced German subjects—showed bi-directional confusions for /ɛ/-/æ/. Some Mandarin subjects produced vowels that were higher in vowel space than /ɛ/, but others produced vowels that were lower in vowel space, thereby showing bi-directional confusions. The experienced Germans produced a difference between the target vowels /ɛ/ and /æ/, but their productions were often identified as a vowel higher in vowel space than the targets. The native Spanish subjects' /ɛ/s were nearly always identified correctly, perhaps because the [ɛ] allophone of Spanish /e/ (Dalbor, 1980) transferred directly into English. When the Spanish subjects misproduced /æ/, it was as a vowel posterior in vowel space to target /æ/.

The Koreans and German subjects' /æ/s were correctly identified significantly less often than all three of the other vowels examined. This suggests that the absence of a vowel from the L1 phonemic inventory may represent, in itself, a source of learning difficulty, as claimed by the Contrastive Analysis Hypothesis (Lado, 1957). However, a contrastive analysis does not explain why the same finding was not obtained for the native speakers of Spanish. Their productions of /ɛ/ were correctly identified significantly more often than were their productions of English /i/ (a vowel found in Spanish) and /ɪ/ (a vowel *not* found in Spanish). Nor can a contrastive analysis explain why the native Mandarin subjects' /ɪ/s were identified significantly more often than their /ɛ/s. (Neither vowel is phonemic in Mandarin.) As mentioned in the Introduction, understanding empirical data such as these will probably require detailed knowledge of the perceived relation of vowels in the L1 and L2, and perhaps also knowledge of the properties (e.g., duration, formant movement) used to contrast vowels in the L1 (see Flege, 1995).

Finally, it is important to note that the intelligibility results obtained here from three native English-speaking listeners may not generalize to other listeners. As mentioned in the Method section, the three listeners were highly accurate in identifying vowels spoken by the NE subjects at two different times. However, they sometimes differed among themselves in identifying non-native subjects' vowels.[7]

3. Acoustic analysis

The intelligibility analysis just presented was limited in two important ways. First, the responses of particular listeners may vary idiosyncratically or dialectally. Thus one can never be certain that the listeners chosen for an intelligibility test adequately represent the variety (or varieties) of the target language that one's non-native subjects have heard and presumably learned to some degree. Second, the listeners may have overlooked small but systematic differences spoken by the non-native speakers. Therefore, the 1800 vowels examined in Section 2. were measured acoustically to provide a more fine-grained measure of vowel production accuracy.

3.1. *Method*

Vowel duration was measured from the first positive peak in the periodic portion of each digitized waveform to the constriction of the post-vocalic consonant, which was signalled by a decrease in overall amplitude and a decrease in waveform complexity. Fundamental frequency (f_0) was measured as the inverse of the average duration of three glottal periods at the acoustic midpoint of each vowel token.

The frequencies of the first three formants (F_1–F_3) were estimated using linear predictive coding (LPC) analysis. A 25.6 ms Hamming window was centered at the acoustic midpoint of each vowel. Twelve LPC coefficients were calculated for male talkers, and 14 for females. A token was re-measured if its measured value diverged substantially from other tokens of the same vowel spoken by the same talker. All of the productions of a vowel spoken by a subject were re-measured if the subject's mean for that vowel diverged substantially from the means of the other subjects in her group or was associated with an unusually large SD (see Labov, 1986). The technique used to re-measure selected tokens was to increase the number of LPC coefficients and/or to move the Hamming window slightly off center. This usually yielded values that conformed better to the expected value.

3.2. *Results*

3.2.1. *Duration measurements*

Table III presents mean vowel durations. An ANOVA examining the NE subjects' vowel durations yielded a significant effect of Vowel [$F(3, 27) = 59.8$, $p < 0.01$]. A Tukey's test revealed that their /i/-/ɪ/ duration differences (175 *vs.* 144 ms) and /ɛ/-/æ/ duration differences (181 *vs.* 238 ms) were significant ($p < 0.01$), as has been reported previously

[7] For example, Listener 1 correctly identified fewer /i/s spoken by the native Spanish subjects (39%) than did Listener 2 (76%) or Listener 3 (74%), but he correctly identified more of the native Spanish subjects' /ɪ/s than did the other two listeners (82% *vs.* 42%, 44%). The results of a follow-up study suggested that his internal phonetic category representation for English /i/ differed from the other two listeners'. When asked to identify the members of a synthetic /i/-to-/ɪ/ continuum (see Section 4), Listener 1 identified fewer vowels as /i/ than did the other two listeners.

for English (e.g., Lehiste & Peterson, 1959; Hillenbrand, Getty, Clark & Wheeler, 1995). One-way ANOVAs carried out for the eight non-native groups were all significant as well [F-values ranging from 7.3 to 24.5, df $= 3, 27$, $p < 0.01$], although the pattern of significant pairwise between-vowel differences varied across groups. Of the eight non-native groups, only the experienced Germans produced significant duration differences between both /i/-/ɪ/ and /ɛ/-/æ/. This is probably because their L1 was the only one of four considered here to afford good temporal models for both distinctions (*viz*., German /i/-/ɪ/ and /ɛ/-/ɛː/). The remaining significant between-group differences are listed in Table III and will be summarized in the Discussion.

3.2.2. *Frequency measurements*

Given that gender did not affect the intelligibility scores, and that the small number of male and female subjects per group (*viz*. 5) precluded a meaningful analysis of gender, we combined data for the male and female subjects in analyses of formant frequency. A technique described by Syrdal & Gopal (1986) was used to minimize the differences between male and female subjects. Frequency values for f_0, F_1, and F_2 were converted from Hertz to Barks. The mean B0, B1, and B2 values were then calculated for each vowel based on the five available tokens for each subject. Vowel height and frontness-backness were estimated by subtracting the mean B0 from the mean B1 values (B1-B0) and B1 from B2 (B2-B1). When the data for the 10 NE subjects were plotted in a high/low (B1-B0) by front/back (B2-B1) space, a small amount of spectral overlap was evident between /i/-/ɪ/ and /ɛ/-/æ/. The duration differences that the NE subjects produced

TABLE III. The mean duration of English vowels spoken by the subjects in nine groups, in ms. The /i/ *vs*. /ɪ/ and /ɛ/ *vs*. /æ/ differences that were significant at the 0.01 level according to a Tukey's test are marked by an asterisk

		beat vs. bit			*bet vs. bat*		
Native language	L2 experience	/i/	/ɪ/	Δ /i/-/ɪ/	/ɛ/	/æ/	Δ /ɛ/-/æ/
English	—	175 (45)	144 (46)	31*	181 (48)	238 (52)	51*
German	Experienced	155 (34)	91 (29)	64*	120 (33)	162 (45)	42*
	Inexperienced	175 (38)	102 (43)	73*	159 (41)	188 (52)	29
Spanish	Experienced	173 (60)	143 (35)	30	161 (31)	206 (41)	45*
	Inexperienced	159 (49)	131 (35)	28	157 (39)	181 (34)	24
Mandarin	Experienced	183 (69)	112 (41)	71*	162 (44)	174 (52)	12
	Inexperienced	166 (60)	99 (35)	67*	161 (39)	188 (45)	27
Korean	Experienced	162 (37)	119 (30)	43	181 (41)	187 (39)	6
	Inexperienced	200 (47)	129 (32)	71*	180 (35)	174 (31)	− 6

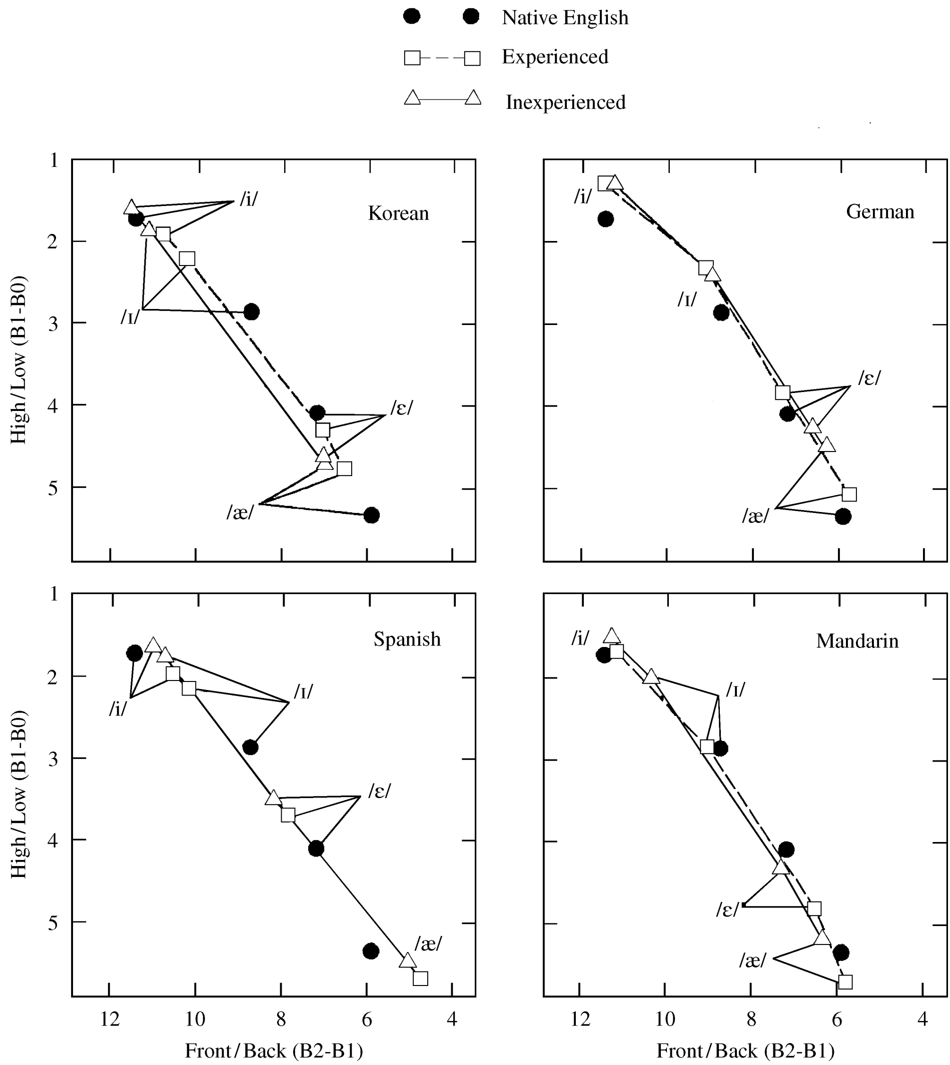

Figure 1. The mean vowel height (B1-B0) and front/back (B2-B1) values obtained for productions of English /i ɪ ɛ æ/ by 10 adult native English speakers (●) and by relatively experienced (□) and inexperienced (△) native speakers of Korean, German, Spanish, and Mandarin. For clarity, the mean values obtained for the experienced and inexperienced subjects are connected by dashed and solid lines, respectively.

between these pairs of vowels probably helped the native-English speaking listeners correctly identify their vowels in virtually every instance (see Bennett, 1968; Ainsworth, 1972; Hillenbrand *et al.*, 1995). Similar plots prepared for non-native subjects were readily interpretable in light of the intelligibility data presented earlier and showed extensive overlap in some instances (e.g., the native Spanish subjects' productions of /i/ and /ɪ/).

The four panels in Fig. 1 show the mean B1-B0 (high/low) and B2-B1 (front/back) values obtained for the native German, Spanish, Mandarin, and Korean subjects' productions of the four target English vowels. The means for the experienced and inexperienced non-native subjects that are shown in each panel have been juxtaposed to the mean values obtained for the NE subjects, who produced systematic differences between /i/, /ɪ/, /ɛ/, and /æ/ in both the high/low (B1-B0) and front/back (B2-B1) dimensions.

The extent to which each non-native subject's productions of a vowel differed from the phonetic norm of English were estimated by computing the Euclidean distance of the individual subjects' B1-B0 and B2-B1 values from the NE groups' mean B1-B0 and B2-B1 values. The 320 "spectral distance" scores (80 non-native subjects $\times$ 4 vowels) computed in this way were submitted to an ANCOVA in which Experience (2 levels) and L1 (4 levels) served as between-subjects factors, Vowel served as a within-subjects factor, and AOA, EDUC, and PER (see Section 2.1.1.) served as covariates. As before, the mean values reported in the text have been adjusted according to the values of the covariates (but differ little from the raw values).

The spectral distance scores were somewhat smaller for the experienced than the inexperienced subjects (1.13 *vs.* 1.27 Barks), indicating that the experienced subjects diverged somewhat less from the phonetic norms of English. The Experience factor did not reach significance [$F(1, 69) = 1.62, p > 0.10$], but it entered into a marginally significant interaction with Vowel [$F(3, 216) = 3.51, p = 0.016$]. This was explored through tests of simple main effects. The spectral distance scores obtained for the experienced and inexperienced subjects' productions of /i/ (0.82 *vs.* 0.72), /ɛ/ (1.10 *vs.* 1.15) and /æ/ (1.21 *vs.* 1.24) did not differ significantly ($p > 0.10$) , but the difference for /ɪ/ (experienced—1.95, inexperienced—1.38 Barks) was marginally significant [$F(1, 75) = 5.17, p = 0.026$]. Recall that this same finding was obtained earlier for the intelligibility scores.[8]

The vowel height and frontness/backness scores obtained for all nine groups of subjects including the NE subjects, were examined next. Both B1-B0 and B2-B1 values were submitted to ANOVAs examining the effect of Vowel (4 levels). All 18 ANOVAs were significant ($p > 0.01$). Pairwise differences between vowels were then examined by Tukey's post-hoc tests using an alpha level of 0.01.

Three groups (the NE subjects, the experienced Mandarin subjects, and the experienced Germans) produced significant B1-B0 differences between both /i/-/ɪ/ and /ɛ/-/æ/. Three groups did not produce a significant difference between either /i/-/ɪ/ or /ɛ/-/æ/ (the inexperienced Mandarin subjects and both Korean groups). The remaining three groups produced a significant difference between the members of just one pair of vowels (the inexperienced Germans for /i/-/ɪ/, both native Spanish groups for /ɛ/-/æ/). Most importantly, in three instances we found that experienced non-native subjects produced a significant difference that was not produced by inexperienced subjects from the same L1 (*viz*., the Germans for /ɛ/-/æ/, and the native Mandarin subjects for /ɛ/-/æ/ and /i/-/ɪ/). In no instance, however, did inexperienced subjects produce a significant difference not produced by experienced subjects from the same L1.

The post-hoc tests yielded the same results for the B2-B1 scores, with just one exception: the experienced Mandarin subjects did not produce a significant B2-B1 difference between /ɛ/-/æ/. Thus, there were just two instances in which experienced subjects produced a difference not produced by inexperienced subjects from the same L1

[8] The similarity of these two analyses is not surprising. Significant simple correlations were found to exist between the 80 non-native subjects' intelligibility scores and the spectral distance scores computed for all four English target vowels ($p < 0.01$ in each instance).

(the Germans for /ɛ/-/æ/, the native Mandarin subjects for /i/-/ɪ/), but no instances of the opposite relation.

It was also important to assess the magnitude of spectral differences that subjects produced between /i/-/ɪ/ and /ɛ/-/æ/. To do so, we computed the Euclidean distance between each subject's mean B1-B0 and B2-B1 values for both pairs. The mean "spectral difference" scores computed for the eight groups of non-native subjects are shown in Fig. 2. The dashed reference line indicates the average spectral difference produced by the NE subjects for /i/-/ɪ/ (top panel) and /ɛ/-/æ/ (bottom panel). Certain groups produced smaller spectral differences than did the NE subjects for /i/-/ɪ/, whereas the

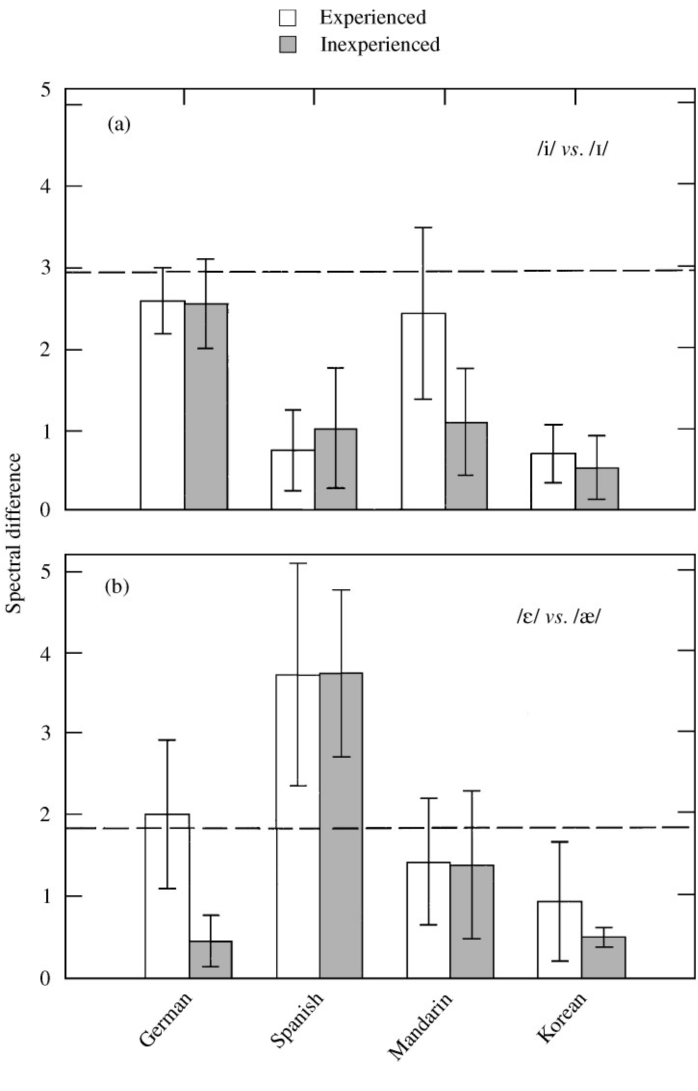

Figure 2. The mean spectral differences between (a) /i/ *vs.* /ɪ/ and (b) /ɛ/ vs. /æ/ that were produced by eight groups of non-native subjects, in Barks. The dashed reference line in each panel represents the mean value obtained for the native English subjects. The error bars enclose $\pm$ 1.0 SE.

two Spanish groups produced larger spectral differences for /ɛ/-/æ/ than did the NE subjects.

The spectral difference scores were submitted to a (9) Group × (2) Vowel Contrast ANOVA, which yielded a two-way interaction [$F(8, 81) = 9.72, p < 0.01$]. The interaction was explored through tests of the simple main effects of Group, which were significant for both /i/-/ɪ/ and /ɛ/-/æ/ ($p < 0.01$) . A series of *t*-tests was carried out to determine which non-native groups differed significantly from the NE group. The NE subjects were found to have produced significantly larger spectral differences between /i/-/ɪ/ than did both Korean groups, both Spanish groups, and the inexperienced Mandarin group (Bonferroni $p < 0.01$). The NE subjects did not produce significantly larger differences between /ɛ/-/æ/ than any nonnative group ($p > 0.10$). However, they produced a difference that was marginally smaller than that of both native Spanish groups (Bonferroni $p = 0.02$ for both NE *vs.* Spanish comparisons).

3.2.3. *Relation to intelligibility*

Stepwise multiple regression analyses were carried out to determine how the acoustic measures related to the intelligibility scores presented in Section 2.2. The criterion variable in each of four analyses (one for each vowel) was the percentage of correct identifications. The predictor variables were B0, B1, B2, B3, B1-B0 ("vowel height"), B2-B1 ("frontness/backness"), and vowel duration.

As expected, the listeners' identifications of vowels spoken by the native and non-native speakers were influenced by both spectral and temporal dimensions. The acoustic variables accounted for a significant 47% of variance in the intelligibility scores for /i/ (B2-B1 = 34% at Step 1, duration = 7% at Step 2; B1 = 6% at Step 3); 37% of the variance for /ɪ/ (B2-B1 = 34% at Step 1, vowel duration = 3% at Step 2); and 43% of the variance for /æ/ (B1-B0 = 27% at Step 1, vowel duration = 13% at Step 2; B3 = 3% at Step 3). However, a significant regression model could not be developed for /ɛ/, probably because there was somewhat less variance in the intelligibility scores for /ɛ/ than for the other three vowels.

3.3. *Discussion*

Our primary aim in this section was to determine if the experienced non-native subjects produced English vowels more accurately than did inexperienced subjects from the same L1 background. This question was answered affirmatively. Spectral distance scores were calculated to estimate the extent to which the non-native subjects' vowels differed from the NE subjects' in front/back (B2-B1) *vs.* high/low (B1-B0) space. The experienced non-native subjects' productions of /ɪ/ diverged marginally less ($p < 0.05$) from the English "norm" than did the inexperienced non-native subjects'. Given that the same result was obtained in the intelligibility test, it seems reasonable to conclude that additional experience in English may lead to a significant improvement in non-natives' productions of English /ɪ/.

Similar differences were not observed for the other three vowels that were examined, however. The null finding for /i/ is not surprising, for all four L1s considered in this study have an /i/. Also, all four L1s have an /ɛ/ phoneme or allophone that might continue to interfere phonetically with, or simply serve as a satisfactory substitute for, the English target vowel /ɛ/. The finding for /æ/ might be seen as surprising if one considers it to be a learnable "new" vowel (see Flege, 1988). However, it appears that [æ]-quality vowels occur in certain contexts in Mandarin, and as a frequent realization of Korean /ɛ/. One might speculate that L1 allophony hindered the Mandarin and Korean subjects. The Germans were somewhat more successful in learning English /æ/, perhaps due to the presence of [æ] in German dialects other than their own. Further research will be needed to test both hypotheses.

We also carried out tests to determine if the experienced non-native subjects produced significant acoustic differences between English vowels that were not produced by inexperienced subjects from the same L1 background. This question was also answered affirmatively. For B1-B0, this held true in three instances (the Germans for /ɛ/-/æ/ and the native Mandarin subjects for both /ɛ/-/æ/ and /i/-/ɪ/). For B2-B1, this held true in two instances (the Germans for /ɛ/-/æ/ and the native Mandarin subjects for /i/-/ɪ/). In no instances did *inexperienced* subjects produce a difference not produced by experienced subjects.

For duration differences, it was twice the case that experienced subjects produced a significant difference not produced by inexperienced subjects (the German and Spanish groups for /ɛ/-/æ/). We might attribute the lack of significant duration differences between /æ/ – /ɛ/ for the inexperienced German subjects, the inexperienced Spanish subjects, and both Mandarin groups to the absence of similar vowel duration differences in the L1 and a failure to learn this aspect of English. As expected, both German groups produced significant, and somewhat larger /i/-/ɪ/ differences than did the NE subjects. This was probably the result of transfer from German (see Bohn & Flege, 1992; Bohn, 1995).

There was one instance in which inexperienced subjects *did* produce a significant duration difference that was not produced by experienced subjects. The inexperienced but not the experienced Koreans produced a significant duration difference between /i/-/ɪ/. We anticipated that the Koreans would over-rely on vowel duration to distinguish English /i/-/ɪ/. Indeed, the inexperienced Koreans' /i/-/ɪ/ difference was much larger than the NE subjects' (71 *vs.* 31 ms). Thus, the fact that the experienced Koreans produced a smaller (43 ms) and non-significant /i/-/ɪ/ difference might be taken as evidence of learning. They may have learned to stop treating English /i/-/ɪ/ as a phonemic length distinction or, alternatively, made less use of duration as they became more sensitive to spectral differences between English /i/-/ɪ/.

The acoustic analyses corroborated the intelligibility test results presented earlier. For example, the intelligibility results suggested that the Spanish subjects produced English /ɛ/ accurately but realized English /æ/ as a more posterior vowel than did the NE subjects. Here we found that the native Spanish subjects produced an even larger spectral difference between /ɛ/-/æ/ than did the NE subjects. To take another example, the inference that the experienced Germans produced both English /ɛ/ and /æ/ with variants that were too high for English was confirmed by the spectral (B1-B0) values obtained here, and are consistent with the previous findings mentioned in Section 1.4.

Finally, the acoustic results generally conformed to the tentative expectations set forth in the Introduction. Our expectation that the Germans would be more successful than the Koreans in contrasting /i/-/ɪ/ was confirmed by the finding that both German groups, but neither Korean group, produced a significant spectral difference between /i/-/ɪ/. Our expectation that the Spanish subjects would be more successful than the Koreans in producing /ɛ/-/æ/ contrasts was also confirmed. Both Spanish groups, but neither Korean group, produced a significant spectral difference between /ɛ/-/æ/. However, the finding that neither experienced nor inexperienced Spanish subjects produced significant spectral or temporal differences between /i/-/ɪ/ was unexpected.

4. Vowel perception

We saw earlier that the non-native subjects often did not produce the same temporal or spectral differences between /i/-/ɪ/ and /ɛ/-/æ/ as did the NE subjects. This might have been due to differences in vowel perception. It has been claimed (e.g. Rochet, 1995) that non-natives' production errors may reflect errors in speech perception. If this is so, then the duration and frequency data presented in Section 3 can be used to generate two predictions concerning the effect of English-language experience:

1) The experienced German subjects will make greater use of spectral cues than the inexperienced Germans to identify vowels in the *bat*-*bet* continuum;
2) The experienced Mandarin subjects will make greater use of spectral cues than the inexperienced Mandarin subjects to identify vowels in both the *beat*-*bit* and *bat*-*bet* continua.

The acoustic data can also be used to generate additional predictions concerning differences between subjects from the four L1 backgrounds:

3) Both Korean groups will make more use of temporal cues than will the NE subjects to identify vowels in the *beat*-*bit* continuum;
4) Both Korean groups will make less use of spectral cues to identify vowels in the *bat*-*bet* continuum than will the NE subjects;
5) Both Spanish groups will use spectral cues for *beat*-*bit* less than the NE subjects;
6) Both Spanish groups, however, will use spectral cues in *bat*-*bet* much like the NE subjects.

4.1. *Method*

The *beat*-*bit* (/i/-/ɪ/) and *bat*-*bet* (/ɛ/-/æ/) continua used here were described in detail by Bohn & Flege (1990). Briefly, the frequencies of F_1–F_3 in the endpoints were based on the mean frequencies reported for American English males by Peterson & Barney (1952). Eleven spectral steps per continuum were created by linear interpolation between the endpoints. Each spectral step had a variant with relatively short (138 ms), mid (190 ms), and long (233 ms) vowel duration. The 33 stimuli (11 spectral steps $\times$ 3 temporal steps) were identifiable as *beat*, *bit*, *bet*, or *bat* because the formant transitions into and out of the "steady state" portion defined /b/ and /t/ consonants, respectively.

The 90 subjects were tested one at a time in a sound booth after producing the English words described earlier. The order in which the continua were tested was counterbalanced over the subjects in each group. The stimuli were randomly presented via headphones 11 times each. The subjects identified each vowel by pushing a button marked "beat" or "bit" (for the /i/-/ɪ/ continuum) or "bet" or "bat" (for the /ɛ/-/æ/ continuum). They were told to guess if uncertain. The interval between each response and the presentation of the next stimulus was 1.0 s. The dependent variable was the percentage of /i/ (*beat*-*bit* continuum) or /æ/ (*bat*-*bet* continuum) judgments given in response to the final 10 presentations of each stimulus.

4.2. *Results*

The NE subjects' mean identification responses are presented in Fig. 3. Their percentage of "beat" responses decreased systematically as spectral quality changed from /i/ to /ɪ/, and their percentage of "bat" responses increased systematically as spectral quality changed from /ɛ/ to /æ/. Vowel duration influenced the NE subjects' identifications, primarily for vowels at the middle of the continua where spectral cues were insufficient to define a vowel's identity unambiguously. As expected, long vowels that were spectrally ambiguous were labeled "beat" (/i/) more often than were the stimuli with short vowels. So, too, spectrally ambiguous stimuli with long vowels were labeled "bat" (/æ/) more often than were the stimuli with short vowels.

The averaged data in Fig. 3 are representative of all 10 NE subjects. This kind of data is typically subjected to probit analysis in order to obtain phoneme boundary estimates, but this could not be done here because many non-native subjects did not show a well-defined cross-over from one response category to the other. Also, the effects of the spectral and/or temporal manipulations observed for many non-native subjects were opposite in direction to the effects shown here for the NE subjects. (For example, some non-native subjects gave more /i/ responses to stimuli with short than long vowels.) Reversals were noted in 27 (34%) instances for /i/-/ɪ/ spectral cues, in five (6%) instances for /i/-/ɪ/ temporal cues, in ten (13%) instances for /ɛ/-/æ/ spectral cues, and in 17 (21%) instances for /ɛ/-/æ/ temporal cues. There were slightly more reversals for inexperienced than experienced non-native subjects (34 *vs.* 25), and more for the native speakers of Korean and Spanish (26, 19) than for the native speakers of Mandarin and German (13, 1).

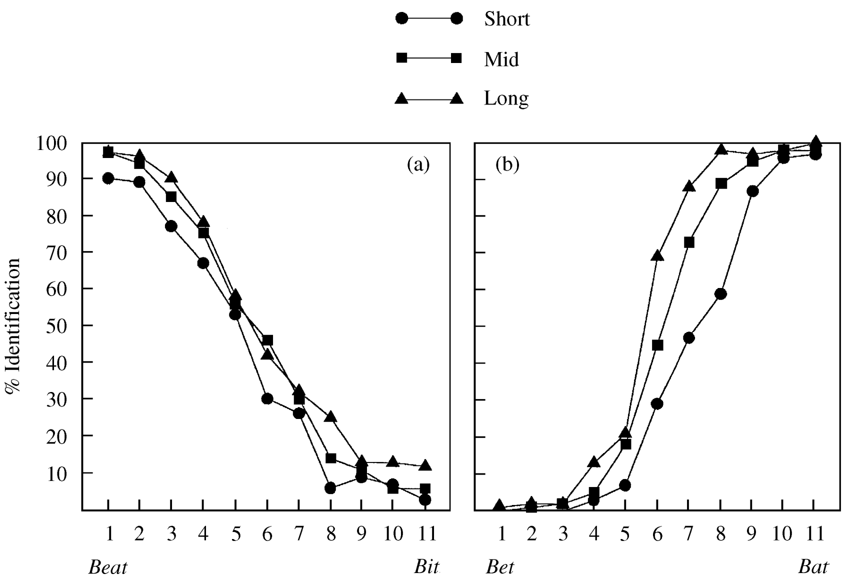

Figure 3. (a) The mean percentage of "beat" responses given by the ten native English subjects to the vowels in a synthetic *beat-bit* (/i/-/ɪ/) continuum; (b) the mean percentage of "bat" responses to a *bet-bat* (/ɛ/-/æ/) continuum. The vowels in both continua varied in 11 spectral steps (x-axis) and three vowel duration steps (represented by separate curves).

The following procedure was adopted to make comparisons between the subject groups possible. For each subject we derived /i/-/ɪ/ and /ɛ/-/æ/ "spectral effect" scores. This was done by subtracting the percentage of "beat" (or "bat") responses given to the /ɪ/ (or /ɛ/) endpoint from the percentage of "beat" (or "bat") responses given to the /i/ (/æ/) endpoint. In calculating the spectral effect scores, we averaged over the three vowel durations for the endpoint stimuli. "Temporal effect" scores were calculated for the /i/-/ɪ/ and /ɛ/-/æ/ continua by subtracting the percentage of "beat" ("bat") responses given to stimuli with short vowels from the percentage of "beat" ("bat") responses given to stimuli with long vowels, averaging over the 11 spectral steps. If a subject showed an effect that was opposite in direction to that shown by the NE subjects (e.g., identifying stimuli with short vowels as /i/ more often than stimuli with long vowels) then the calculated score had a negative value.

The mean spectral and temporal effect scores shown on the right side of the slashes in Table IV are based on both positive and negative values. However, our analyses focused on the unsigned value of the spectral and temporal effect scores. (The means based on unsigned values are shown to the left of the slashes in Table IV.) We reasoned, for example, that if a subject identified 80% of the short-vowel stimuli in the *beat-bit* continuum as /i/ it meant that she was relying on vowel duration to identify vowels even though her use of vowel duration was opposite that of the NE subjects.

As can be seen in Fig. 4, the unsigned spectral and temporal effect scores for *beat-bit* were inversely correlated ($r = -0.943$). The more the 90 subjects based their identifications on duration, the less they used formant frequency (i.e., spectral quality), and vice versa. The NE subjects cluster at the upper left-hand corner because they had large spectral effects and relatively small temporal effects. The spectral and temporal scores obtained for the *bat-bet* continuum, which are shown in Fig. 5, were also inversely correlated ($r = -0.822$). The NE subjects again clustered together in the upper left-hand corner.

TABLE IV. The mean temporal and spectral effect scores obtained for the nine groups of subjects. Means to the left of the slashes, which were used in the analyses reported in the text, were calculated after removing the signs of the difference scores. Means to the right of the slashes are based on the signed values, and thus show the effect of "reversals" (see text). Underlined means differed significantly from the NE subjects' mean values (boldfaced) according to a Tukey's post-hoc test ($p < 0.01$)

		beat vs. bit		*bet vs. bat*	
Native language	L2 experience	Temporal effect	Spectral effect	Temporal effect	Spectral effect
English	—	**11/9**	**88/88**	**15/15**	**98/98**
German	Experienced	30/32	64/64	45/45	70/70
	Inexperienced	36/35	63/63	59/59	43/38
Spanish	Experienced	46/16	50/-6	19/19	88/88
	Inexperienced	50/46	44/13	12/12	94/94
Mandarin	Experienced	29/29	60/59	35/22	42/40
	Inexperienced	85/85	11/0	72/54	17/11
Korean	Experienced	66/66	29/-24	54/4	31/16
	Inexperienced	81/81	15/-4	66/-49	24/9

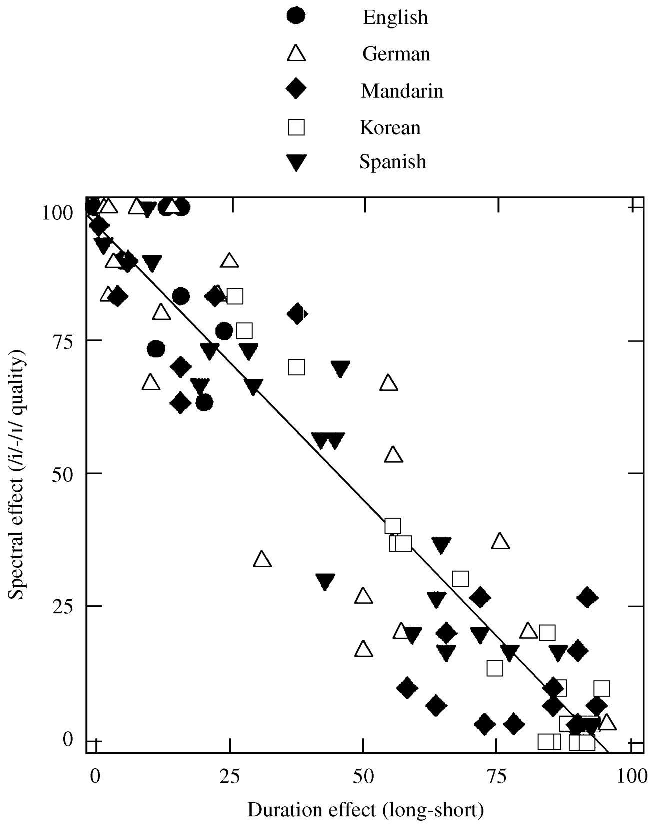

Figure 4. The relation between (unsigned) spectral and temporal effect scores obtained for 10 native English and 80 non-native subjects for the *beat-bit* continuum.

In analyzing the vowel perception data, our first aim was to determine which non-native groups differed significantly from the NE subjects. The unsigned temporal and spectral effect scores for the *beat-bit* continuum were submitted to one-way ANOVAs examining the effect of Group (9 levels). Both were significant ($p < 0.001$). Tukey's post-hoc tests revealed that three non-native groups—the inexperienced Koreans, the experienced Koreans, and the inexperienced native Mandarin subjects—had significantly smaller spectral effect scores than did the NE subjects ($p < 0.01$). These same three groups had significantly *larger* temporal effect scores than did the NE subjects ($p < 0.01$). For the *bat-bet* continuum, the same pattern was observed for the inexperienced native speakers of three L1s: German, Mandarin, and Korean. No experienced non-native group showed the inverse pattern seen earlier (*viz.* significantly smaller spectral effects and significantly larger temporal effects) for the *bat-bet* continuum. However, the experienced Mandarin and Korean subjects had significantly smaller temporal effect scores than did the NE subjects ($p < 0.01$).

The effect of English-language experience was assessed by submitting the temporal and spectral effect scores to a series of ANCOVAs in which L1 (4 levels) and Experience (2 levels) served as between-subjects factors, and AOA, PER, EDUC (see above) served as

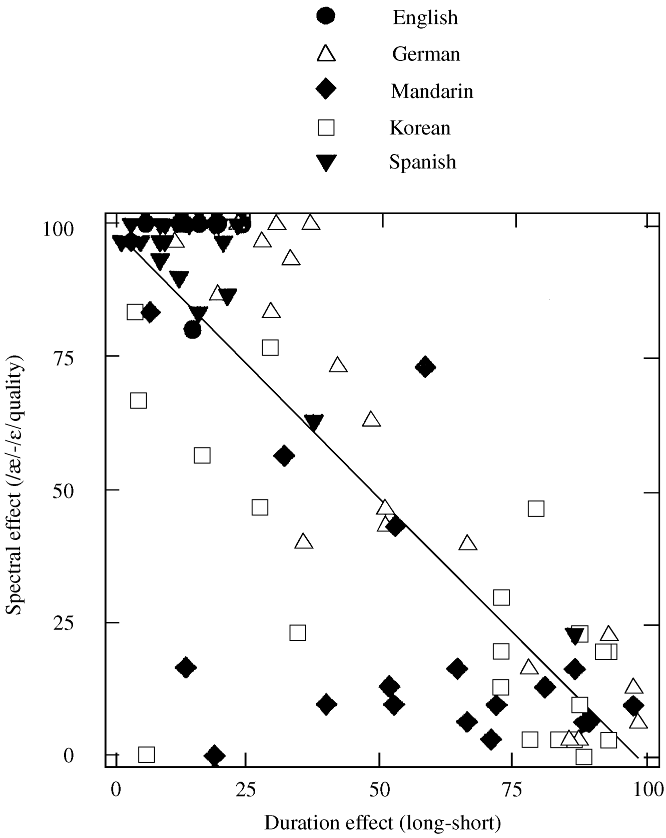

Figure 5. The relation between (unsigned) spectral and temporal effect scores obtained for 10 native English and 80 non-native subjects for the *bet-bat* continuum.

covariates. The spectral effect scores for both continua were greater for the experienced than the inexperienced subjects (*beat-bit*: 59% *vs.* 29%, *bat-bet*: 61% *vs.* 41%). Conversely, the temporal effect scores were *smaller* for the experienced than the inexperienced subjects (*beat-bit*: 40% *vs.* 66%, *bat-bet*: 35% *vs.* 55%).[9] Each of these differences was significant [F-values for the main effect of Experience ranging from 7.04 to 14.5, with 1 and 69 df's, $p < 0.01$].

Of the four L1 groups, the Germans resembled the NE subjects most closely for the *beat-bit* continuum, and the Spanish subjects resembled the NE subjects most closely for the *bat-bet* continuum. The main effect of Language was significant in all four analyses [F-values ranging from 7.9 to 19.8, with 3 and 69 df's, $p < 0.01$]. Tukey's post-hoc tests revealed that the German subjects' *beat-bit* spectral effect scores were larger than were the Korean and Mandarin subjects' (68% *vs.* 22%, 31%; $p < 0.01$) and, conversely, the German's temporal effect scores were *smaller* than the Korean and Mandarin subjects' (30% *vs.* 71%, 60%; $p < 0.01$). The Spanish subjects' *bat-bet* spectral effect scores were

[9] Once again, these values and those to follow have been adjusted according to variations in the covariates. They differed little from the raw values.

TABLE V. The number of subjects per group (maximum = 10) who showed a predominant use of spectral or temporal cues when identifying vowels (see text)

Native language	L2 experience	*beat-bit*		*bat-bet*	
		Temporal	Spectral	Temporal	Spectral
English	—	0	10	0	10
German	Experienced	2	6	3	7
	Inexperienced	4	6	6	3
Spanish	Experienced	3	6	1	9
	Inexperienced	6	4	0	10
Mandarin	Experienced	3	7	3	4
	Inexperienced	10	0	9	1
Korean	Experienced	8	2	6	2
	Inexperienced	9	1	7	2

significantly larger than were the German, Korean, and Mandarin subjects' (91% *vs.* 58%, 29%, 26%; $p < 0.01$). Conversely, the Spanish subjects' temporal effect scores were significantly smaller than those of the German, Korean, and Mandarin subjects (16% *vs.* 50%, 59%, 56%; $p < 0.01$).[10]

The Experience by Language interaction was non-significant in the analyses of the *bat-bet* spectral effect and temporal effect scores [$F(3, 69) = 1.62, 2.52, p > 0.05$]. The interaction reached marginal significance for the *beat-bit* spectral effect scores [$F(3, 69) = 3.05, p = 0.03$] and significance for the *beat-bit* temporal effect scores [$F(3, 69) = 4.72, p < 0.01$], however. Tests of the simple effect of Experience revealed that the interactions were due to the fact that the experienced Mandarin subjects had significantly larger spectral effect scores than did the inexperienced Mandarin subjects, and significantly smaller temporal effect scores ($p < 0.01$). The differences between the experienced and inexperienced Koreans were marginally significant for the spectral effect ($p = 0.032$) and the temporal effect scores ($p = 0.013$).

To assess individual performance, each subject was classified as being a "predominant" user of spectral or temporal cues depending on whether her (unsigned) spectral score or temporal effect score was the larger. An assignment was made, however, only if the larger of the two scores exceeded 50%. (Three subjects were not classified as predominant users of either spectral or temporal cues for *beat-bit*, and seven for *bat-bet*, because neither score exceeded 50%.) As summarized in Table V, all 10 NE subjects were predominantly spectral cue users when identifying vowels in both the *beat-bit* and *bat-bet* continua. Fewer inexperienced than experienced subjects resembled the NE subjects in being predominant spectral cue users for *beat-bit* (21 *vs.* 11) and *bat-bet* (22 *vs.* 16). Conversely, more inexperienced than experienced subjects were predominant temporal cue users for *beat-bit* (29 *vs.* 16) and *bat-bet* (22 *vs.* 13). This supports the view that vowel perception becomes more native-like as a function of L2 experience. In the Introduction we speculated that the Koreans would over-rely on vowel duration. In fact, more

[10] The Germans' spectral effect scores were also larger than those of the Korean and Mandarin subjects ($p < 0.01$).

Koreans were predominant temporal than spectral cue users for both *beat*-*bit* (17 *vs.* 3) and *bat*-*bet* (13 *vs.* 4).

Finally, the individual subject data are consistent with the view that vowel production and perception are related. The acoustic analyses presented earlier indicated that the native Spanish subjects did not produce significant duration or spectral differences between English /i/-/ɪ/. However, they produced an even larger spectral difference between English /ɛ/ and /æ/ than did the NE subjects. These findings suggested that the Spanish subjects would not favor temporal or spectral cues when identifying members of the *beat*-*bit* continuum, but would rely on spectral cues to identify vowels in the *bat*-*bet* continuum. In fact, the Spanish subjects were equally likely to be predominant temporal or spectral cue users for *beat*-*bit* (9 *vs.* 10) whereas they were far more likely to be predominant spectral than temporal cues users for *bat*-*bet* (19 *vs.* 1).

4.2.1. *The production-perception relation*

Stepwise multiple regression analyses were also carried out to examine the relation between the non-natives' vowel production and perception. We reasoned that if non-natives' accuracy in producing and perceiving English vowels are related, then more variance in the /i/-/ɪ/ production data should be accounted for by variables related to the perception of /i/ and /ɪ/ (*viz*., the *beat*-*bit* spectral and temporal effect scores) than by variables related to the perception of /ɛ/ and /æ/ (*viz*., the *bat*-*bet* scores), and vice versa.[11] Four analyses were undertaken. In each, the four predictor variables were the spectral and temporal effect scores obtained in the vowel perception experiment for the *beat*-*bit* and *bat*-*bet* continua. These predictor variables were regressed onto four criterion variables, all measures of the non-native subjects' vowel production accuracy: (1) listeners' perceptual confusions for the vowel pair /i/-/ɪ/, i.e., the number of times one of these vowels was heard for the other; (2) listeners' perceptual confusions of the vowel pair /ɛ/-/æ/; (3) the spectral distance between /i/-/ɪ/ in a B1-B0 *vs.* B2-B1 space; and (4) the spectral distance between /ɛ/-/æ/.

The results of the four regression analyses summarized in Table VI support the view that vowel production and perception accuracy are related. When averaged over the listener-based and acoustic estimates of vowel production accuracy, the /i/-/ɪ/ perceptual data (i.e., the /i/-/ɪ/ spectral and temporal effect scores) accounted for more variance in the /i/-/ɪ/ production scores than did the /ɛ/-/æ/ perception data (29.8% *vs.* 10.3%). Conversely, when averaged over the two production variables, the /ɛ/-/æ/ perception data accounted for more variance in the /ɛ/-/æ/ production scores (33.5%) than did the /i/-/ɪ/ spectral and temporal effects scores (3.6%).

4.3. *Discussion*

This study demonstrated that the amount of English-language experience that non-natives had exerted an important influence on how they perceived English vowels. When identifying the members of the *bat*-*bet* continuum, the experienced non-native subjects made significantly more use of spectral cues than did the inexperienced non-native

[11] We used the signed spectral and temporal effect scores in these regression analyses. Preliminary analyses revealed that most non-native subjects who showed reversals in the perception experiment produced very small or no acoustic difference between the two English vowels contrasted in the perceptual continuum.

TABLE VI. Summary of multiple regression analyses examining the relation between four vowel perception scores (the spectral and temporal effect scores obtained for two synthetic continua) and two measures each of the accuracy with which nonnative subjects produced /i-/ɪ/ and /ɛ/-/æ/ (see text). All variance values represent a significant increase from the previous step ($p < 0.01$)

Production of /i/-/ɪ/		Production of /ɛ/-/æ/	
A. Perceptual confusions		*A. Perceptual confusions*	
Step 1. /i/-/ɪ/ spectral effect:	21.5%	Step 1. /ɛ/-/æ/ spectral effect:	29.6%
Step 2. /i/-/ɪ/ duration effect:	10.0%	Step 2. /ɛ/-/æ/ duration effect:	10.0%
Step 3. /ɛ/-/æ/ duration effect:	8.5%		
Total:	40.0%	Total:	39.6%
B. Spectral distances		*B. Spectral distances*	
Step 1. /i/-/ɪ/ spectral effect	28.0%	Step 1. /ɛ/-/æ/ spectral effect	27.4%
Step 2. /ɛ/-/æ/ duration effect	12.0%	Step 2. /i/-/ɪ/ spectral effect	7.1%
Total:	40.0%	Total:	34.5%

subjects. Conversely, they made significantly less use of temporal cues. As a result, the experienced subjects resembled the NE subjects more closely than did the inexperienced subjects. For the *beat*-*bit* continuum, the same pattern of results was obtained for experienced *vs.* inexperienced Mandarin subjects ($p < 0.1$) and, at a marginal level of significance ($p < 0.05$), for the experienced *vs.* inexperienced Korean subjects.

Six predictions were presented in the introduction to this section as to how the non-native subjects would use temporal and spectral cues to identify vowels in the two continua. The predictions were based on the acoustic data obtained earlier. Five predictions (1, 2, 3, 4, and 6) were confirmed, thereby supporting the view (e.g., Rochet, 1995; Flege, 1995) that L2 vowel production is influenced importantly by how L2 vowels are perceived. Also, multiple regression analyses showed that the non-native subjects' accuracy in producing English vowels was related to their accuracy in *perceiving* the same English vowels.

One other prediction (number 5), *viz.* that both the experienced and inexperienced native Spanish subjects would make less use of spectral cues than would the NE subjects, was not supported. The native Spanish subjects identified the /i/ endpoint of the *beat*-*bit* continuum 47% more often than the /ɪ/ endpoint on average. Although their difference scores were smaller than those of the NE subjects ($M = 88\%$), the spectral effect scores obtained for neither the experienced nor the inexperienced Spanish subjects differed significantly from the NE subjects' scores. Earlier we found that neither Spanish group *produced* a significant spectral difference between English /i/-/ɪ/. Taken together, these findings suggest that non-natives' production and perception of L2 vowels do not always match perfectly. At times, their perception may be somewhat more native-like than is their production.

Our nonnative subjects often showed "reversals", that is, had spectral and temporal effects that were opposite in direction to those of the NE subjects. A possible explanation for this phenomenon is that one of the two keywords used to identify the vowels in a continuum was more familiar to nonnative subjects than the other keyword, creating a response bias. To test this, we examined the word familiarity ratings obtained from the native German and Spanish subjects (see Method section). These subjects rated the four response labels used in the vowel perception experiment (*viz. beat*, *bit*, *bet*, *bat*) using a scale that ranged from "very often seen or heard" (1) to "never seen or heard" (9). The native Spanish subjects showed more reversals than did the German subjects (19 *vs.* 1). However, an ANOVA examining the word familiarity ratings revealed that both the native Spanish and German subjects showed similar, and non-significant, differences between the two keywords used as response labels in both continua (see also Miranda & Strange, 1989).

Orthographic confusions may have led to some reversals, at least for the native Spanish subjects. Just one Spanish subject showed a reversal for the *bat*-*bet* continuum. The keywords used to identify the vowels in this continuum were not likely to be confused orthographically (see Flege, 1991). However, the Spanish subjects showed 19 reversals for the *beat*-*bit* continuum, where orthography probably played a role. Of the reversals for *beat*-*bit*, 18 involved the labeling of /ɪ/-quality stimuli as /i/. Although the English keyword "bit" contains /ɪ/, this word is spelled with "i", a letter that always represents /i/ in Spanish. The fact that two native Spanish subjects produced *beat* with [e] or [ɛ] quality vowels is also consistent with the orthographic confusion hypothesis.

Earlier we reported that the Koreans, especially those who were inexperienced in English, produced larger /i/ *vs.* /ɪ/ duration differences than did the NE subjects. They also made greater use of vowel duration than did the NE subjects when *identifying* the vowels in both the *beat*-*bit* and *bat*-*bet* continua. Korean is traditionally analyzed as having a distinction between long *vs.* short vowels. Perhaps the Koreans made greater perceptual use of vowel duration than did the NE subjects because duration is a more important perceptual cue to vowel identity in Korean than in English (see, e.g., Gottfried & Beddor, 1988). If so, then our findings for the Koreans might be taken as supporting the view that L1 features transfer from the L1 to an L2 and govern how non-natives' produce and perceive L2 vowels (see Flege, 1995).

There are two reasons, however, to question this conclusion. First, as discussed in the Introduction, many young speakers of Seoul Korean do not maintain these distinctions. Second, a general finding of the vowel perception experiment was that the use of temporal and spectral cues were correlated inversely. The Koreans did not produce significant spectral contrasts between English /i/-/ɪ/ or /ɛ/-/æ/. They may have relied on vowel duration in the perception experiment not because of its cue value in Korean, but because they were unable to make effective use of the spectral differences distinguishing vowels in the two continua. This issue could be resolved by training Korean subjects to identify the endpoint stimuli in the *beat*-*bet* and *bat*-*bet* continua with 100% accuracy, then assessing their use of spectral and temporal cues on the full continua (presented without feedback).

5. General discussion

This study provided evidence that adults who learn a second language (L2) will come to produce and perceive certain vowels in their L2 more accurately as they gain experience in the L2. Significant differences between relatively experienced and inexperienced adult non-native speakers of English were found to exist when vowel production accuracy was assessed in an intelligibility study and through acoustic analyses. Significant differences were also noted in a vowel perception experiment using synthetic stimuli.

The findings just summarized are consistent with the view that speech learning abilities remain intact across the life span. However, it is important to note that evidence of phonetic learning was not obtained *in all possible instances*. For example, the experienced Korean and Spanish subjects were no more accurate in producing a contrast between English /i/-/ɪ/ than were less experienced subjects from the same native language (L1) background. Also, both experienced and inexperienced German subjects, and both experienced and inexperienced Mandarin subjects, produced much larger duration differences between English /i/-/ɪ/ than did the NE subjects. (The Germans' overuse of duration may have been the result of transfer from the L1, whereas the Mandarin subjects' overuse may have been a strategy learned at school; see Introduction).

It is also important to note that the performance of the relatively "experienced" non-native subjects examined here was seldom completely native-like. Perhaps it takes more than an average of 7 years using English regularly in a predominantly L2-speaking environment before a native-like level of performance is attained (see Flege, Takagi & Mann, 1995). Another possibility is that adult L2 learners' performance may be constrained by a variety of psychosocial factors, or by phonetic factors that continue to operate even when the L2 has been spoken for many years (e.g., the dissimilatory or assimilatory effects predicted to occur when new vowel categories are, or are not, established during L2 acquisition; see Flege, 1995). Still another possibility is that hearing English spoken with a foreign accent by other non-natives may have influenced our subjects' performance. On the other hand, it may be that adults' ability to learn aspects of the L2 sound system is limited in an absolute sense (Takagi & Mann, 1995).

This study also demonstrated that the accuracy with which non-natives perceive and produce English vowels are related. Acoustic-based predictions concerning how the various groups of non-native subjects would perform in the vowel perception experiment were largely supported. Also, regression analyses examining variables derived from the vowel perception experiments accounted for a significant amount of variance in measures of the non-native subjects' vowel production accuracy. However, a substantial amount of variance in the production data remained unaccounted for. A reason for this might be that certain aspects of perception change before corresponding changes are implemented in production. Another potential explanation is that age-based limitations on learning are firmer for productive than perceptual aspects of L2 acquisition. Of course, we cannot rule out the possibility that changes in production occur first, or that they occur in the absence of corresponding changes in perception.

Finally, the results obtained here suggest that the nature of the L1 vowel inventory and its perceived relation to vowels in an L2 influence the extent to which L2 vowel production and perception will improve as non-natives gain experience in their L2. To take an example, the inexperienced German subjects were far less successful in producing and perceiving a distinction between English /ɛ/-/æ/ than were the inexperienced Spanish subjects. It is likely that many of the German subjects identified realizations of English /ɛ/ and /æ/ as instance of a single L1 vowel (German /ɛ/ or perhaps /ɛː/) whereas the Spanish subjects probably identified the same English vowels as instances of *two different* L1 vowels (*viz*. Spanish /e/ and /a/; see Flege, 1991), which led them to produce *even larger* spectral differences between English /ɛ/-/æ/ than did the NE subjects.

In summary, the results of this study support the view that adults who learn an L2 become able to produce and perceive certain L2 vowels more accurately as they gain experience in an L2. The non-natives speakers' accuracy in producing English vowels was related to their accuracy in perceiving the same English vowels, although the exact nature of the production-perception relation was uncertain. Finally, the non-natives' degree of accuracy in producing and perceiving the English vowels, as well as the extent to which their performance improved with experience in English, varied as a function of L1 background. This was likely due to cross-language differences in the perceived relation between vowels found in the L1 and English. As discussed in the Introduction, additional research is needed to define the degree of perceived relatedness, if any, between vowels found in English and in each of the L1s examined here.

This research was supported by grant DC00257 from the National Institute of Deafness and Other Communicative Disorders. The Mandarin data are from a M.A. thesis by Chipin Wang, which was carried out under the supervision of the first author. The authors thank P. Beddor, T. Gottfried, S. Guion, and an anonymous reviewer for their comments on an earlier version.

References

Ainsworth, W. (1972) Duration as a cue in the recognition of synthetic vowels, *Journal of the Acoustical Society of America*, **51,** 648–651

Bennett, D. (1968) Spectral form and duration as cues in the recognition of English and German vowels, *Language and Speech*, **11,** 64–85

Best, C. & Strange, W. (1992) Effects of phonological and phonetic factors on cross-language perception of approximants, *Journal of Phonetics*, **20,** 305–331

Bohn, O.-S. (1995) Cross-language speech perception in adults: first language transfer doesn't tell it all. In *Speech perception and linguistic experience*: *theoretical and methodological issues* (W. Strange, editor) Timonium, MD: York Press.

Bohn, O.-S. & Flege, J. E. (1990) Interlingual identification and the role of foreign language experience in L2 vowel perception, *Applied Psycholinguistics*, **11,** 303–328

Bohn, O.-S. & Flege, J. E. (1992) The production of new and similar vowels by adult German learners of English, *Studies in Second Language Acquisition*, **14,** 131–158

Cheng, C.-C. (1973) *A synchronic phonology of Mandarin Chinese*, The Hague: Mouton

Cheng, R. (1966) Mandarin phonological structure, *Journal of Linguistics*, **2,** 135–158

Colhoun, E. & Kim, T. W. (1976) English loanwords in Korean. Lexical and phonological problems, *Applied Linguistics*, **8,** 237–250

Collins, B. & Mees, I. (1984) *The Sounds of English and Dutch*. Leiden: Leiden University Press

Cutler, A. & Otake, T. (1994) Mora or phoneme? Further evidence for language-specific listening, *Journal of Memory and Language*, **33,** 824–844

Dalbor, J. (1980) *Spanish pronunciation*, *theory and practice*. New York: Holt, Rinehart, and Winston

Di Benedetto, M. (1989) Vowel representation: Some observations on temporal and spectral properties of the first formant frequency, *Journal of the Acoustical Society of America*, **86,** 55–66

Flege, J. (1981) The phonological basis of foreign accent, *TESOL Quarterly*, **15,** 443–455

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

Flege, J. (1988) The production and perception of speech sounds in a foreign languages. In *Human communication and Its Disorders*, *A Review 1988* (H. Winitz, editor) pp. 224–240. Norwood, NJ: Ablex

Flege, J. (1991) Orthographic evidence for the perceptual identification of vowels in Spanish and English, *Quarterly Journal of Experimental Psychology*, **43**(3), 701–731

Flege, J. (1992*a*) The Intelligibility of English vowels spoken by British and Dutch talkers. In *Intelligibility in speech disorders*: *theory*, *measurement*, *and management* (R. Kent, editor) Amsterdam: John Benjamins

Flege, J. (1992*b*) Speech learning in a second language, In *Phonological development*: *models*, *research*, *and application* (C. Ferguson, L. Menn & C. Stoel-Gammon, editors). Timonium, MD: York Press

Flege, J. (1995) Second-language speech learning: theory, findings, and problems, In *Speech perception and linguistic experience*: *theoretical and methodological issues* (W. Strange, editor), Timonium, MD: York Press

Flege, J., Munro, M. & MacKay, I. (1995) Factors affecting degree of perceived foreign accent in a second language, *Journal of the Acoustical Society of America*, **97,** 3125–3134

Flege, J., Takagi, N. & Mann, V. (1995) Japanese adults can learn to produce English /ɹ/ and /l/ and accurately, *Language and Speech*, **38,** 25–55

Flege, J., Takagi, N. & Mann, V. (1996) Lexical familiarity and English-language experience affect Japanese adults' perception of /ɹ/ and /l/, *Journal of the Acoustical Society of America*, **99,** 1161–1173

Fox, R., Flege, J. & Munro, M. (1995) The perception of English and Spanish vowels by native English and Spanish listeners: a multidimensional scaling analysis, *Journal of the Acoustical Society of America*, **97,** 2540–2551

Gottfried, T. (1984) Effect of consonant context on the perception of French vowels, *Journal of Phonetics*, **12,** 91–114

Gottfried, T. & Beddor, P. (1988) Perception of temporal and spectral information in French vowels, *Language and Speech*, **31,** 57–75

Gottfried, T., Miller, J. & Payton, P. (1990) Effect of speaking rate on the perception of vowels, *Phonetica*, **4,** 155–172

Harris, J. (1969) *Spanish phonology*, Cambridge, Mass.: MIT Press

Hillenbrand, J., Getty, L., Clark, M. & Wheeler, K. (1995) Acoustic characteristics of American English vowels, *Journal of the Acoustical Society of America*, **97,** 3099–3111

Hong, Y. (1988) *A Sociolinguistic Study of Seoul Korean.* Ph.D. dissertation, University of Pennsylvania

Howie, J. (1976) *Acoustical studies of Mandarin vowels and tones*, Cambridge: Cambridge University Press

Ingram, J. & Park, S.-G. (1997) Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics*, **25,** 343–370

Ioup, G. (1995) Evaluating the need for input enhancement in post-critical period language acquisition. In *The age factor in second language acquisition* (D. Singleton & Z. Lengyel, editors), pp. 95–123. Clevedon: Multilingual Matters

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

Kim, D. J. (1972) A contrastive study of English and Korean phonology (on the basis of aural perception), *Language Teaching*, **5,** 1–36

Labov, W. (1986) Sources of inherent variation in the speech process. In *Invariance and variability in speech processes* (J. S. Perkell & D. H. Klatt, editors), pp. 402–425, Hillsdale, N.J.: Erlbaum

Lado, R. (1957) *Linguistics across cultures.* Ann Arbor: University of Michigan Press

Leather, J. (1990) Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In *New Sounds 90*: *proceedings of the Amsterdam symposium on the acquisition of second-language speech* (J. Leather & A. James, editors), pp. 72–89. Amsterdam: University of Amsterdam

Lee, H. B. & Zhi, M-Je. (1987) A spectrographic study of Korean Vowels, *Korea Journal*, February, 37–41

Lehiste, I. & Peterson, G. E. (1959) Vowel amplitude and phonemic stress in American English, *Journal of the Acoustical Society of America*, **31,** 428–435

Li, C. & Thompson, S. (1981) *Mandarin Chinese*, *a functional reference grammar*. Los Angeles: University of California Press

Light, T. (1976) *The Chinese syllabic final.* Ithaca: Cornell University Press

Lin, H.-B. & Repp, B. (1989) Cues to the perception of Taiwanese tones, *Language and Speech*, **32,** 25–44

Maddieson, I. (1984) *Patterns of sounds.* Cambridge: Cambridge University Press

Magen, H. & Blumstein, S. (1993) Effects of speaking rate on the vowel length distinction in Korean, *Journal of Phonetics*, **21,** 387–410

Major, R. (1987) Phonological similarity, markedness, and rate of L2 acquisition, *Studies in Second Language Acquisition*, **9,** 63–82

Martin, S. E. (1951) Korean phonemics, *Language*, **27,** 519–533

Miranda, S. & Strange, W. (1989) The role of spectral, temporal, and dynamic cues in the perception of English vowels by native and non-native speakers, *Journal of the Acoustical Society of America*, **85,** S53(A)

Moulton, W. (1962) *The Sounds of English and German.* Chicago: The University of Chicago Press

Munro, M. (1992) *Perception and Production of English Vowels by Native Speakers of Arabic.* Unpublished University of Alberta Ph.D. dissertation

Munro, M. (1993) Production of English vowels by native speakers of Arabic: acoustic measurements and accentedness ratings, *Language and Speech*, **36,** 39–66

Munro, M., Flege, J. & MacKay, I. (1996) The effects of age of second-language learning on the production of English vowels, *Applied Psycholinguistics*, **17,** 313–334

Patkowski, M. (1989) Age and accent in a second language: a reply to James Emil Flege, *Applied Linguistics*, **11,** 73–89

Peterson, G. & Barney, H. (1952) Control methods used in a study of the vowels, *Journal of the Acoustical Society of America*, **24,** 175–184

Peterson, G. & Lehiste, I. (1956) Duration of syllable nuclei in English, *Journal of the Acoustical Society of America*, **32,** 693–703

Polka, L. & Bohn, O.-S. (1996) A cross-language comparison of vowel perception in English-learning and German-learning infants, *Journal of the Acoustical Society of America*, **100,** 577–592

Robson, B. (1982) Teaching English pronunciation to speakers of Korean, *Focus*, **11,** 1–8 (National Clearinghouse for Bilingual Education, Washington, D.C.)

Rochet, B. (1995) Perception and production of second-language speech sounds by adults. In *Speech perception and linguistic experience* (W. Strange, editor). Timonium, MD: York Press

Scovel, T. (1969) Foreign accents, language acquisition, and cerebral dominance, *Language Learning*, **19,** 245–253

Schmidt, A. (1996) Cross-language identification of consonants. Part 1. Korean perception of English, *Journal of the Acoustical Society of America*, **99,** 3201–3211

Steensland, L. (1981) A method for measuring perceptual distances between different vowel qualities. *Acta Universitatis Upsaliensis*, **21**. Uppsala. Stockholm: Almqvist and Wiksell

Stockwell, R. & Bowen, D. (1965) *The sounds of English and Spanish*. Chicago: University of Chicago Press

Syrdal, A. & Gopal, H. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels, *Journal of the Acoustical Society of America*, **79,** 1086–1100

Takagi, N. & Mann, V. (1995) The limits of extended naturalistic exposure on the perceptual mastery of English /r/ and /l/ by adult Japanese learners of English, *Applied Psycholinguistics*, **16,** 379–405

Wang, C. (1988) The production and Perception of English vowels by native speakers of Mandarin. Unpublished M.A. thesis, University of Alabama at Birmingham.

Yamada, R. & Tohkura, Y. (1991) Age effect on acquisition of non-native phonemes: perception of English /r/ and /l/ for native speakers of Japanese. In *Proceedings of the XIIth international congress of phonetic sciences, August 19–24, 1991*, Aix-en-Provence, France

Yamada, R., Strange, W., Magnuson, J., Pruitt, J. & Clark, W. III. (1994) The intelligibility of Japanese speakers' production of American English /r/, /l/, and /w/, as evaluated by native speakers of American English. *International Conference on Spoken Language Processing*, **94,** 2023–2026

Yang, B. (1996) A comparative study of American English and Korean vowels produced by male and female speakers, *Journal of Phonetics*, **24,** 245–262