Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure[a)]

G a J)
Department of Psychology, Lehman College, City University of New York, 250 Bedford Park Blvd. West, Bronx, New York 10468

W St a
Graduate Center, City University of New York, 365 5th Avenue, New York, New York 10016

Ya W)
Department of Psychology, Peking University, Beijing, China, 100871

J a C a
Department of Psychology, Lehman College, City University of New York, 250 Bedford Park Blvd. West, Bronx, New York 10468

QG a
Department of Psychology, Peking University, Beijing, China, 100871

(Received 5 May 2005; revised 18 November 2005; accepted 22 November 2005)

In this study we assessed age-related differences in the perception and production of American English (AE) vowels by native Mandarin speakers as a function of the amount of exposure to the target language. Participants included three groups of native Mandarin speakers: 87 children, adolescents and young adults living in China, 77 recent arrivals who had lived in the U.S. for two years or less, and 54 past arrivals who had lived in the U.S. between three and five years. The latter two groups arrived in the U.S. between the ages of 7 and 44 years. Discrimination of six AE vowel pairs /i-ɪ/, /i-eɪ/, /ɛ-æ/, /æ-ɑ/, /ɑ-ʌ/, and /u-ɑ/ was assessed with a categorial AXB task. Production of the eight vowels /i, ɪ, eɪ, ɛ, æ, ʌ, ɑ, u/ was assessed with an immediate imitation task. Age-related differences in performance accuracy changed from an older-learner advantage among participants in China, to no age differences among recent arrivals, and to a younger-learner advantage among past arrivals. Performance on individual vowels and vowel contrasts indicated the influence of the Mandarin phonetic/phonological system. These findings support a combined environmental and L1 interference/transfer theory as an explanation of the long-term younger-learner advantage in mastering L2 phonology. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2151806]

PACS number(s): 43.71.Hw, 43.71.Ft, 43.70.Ep [ALF] Pages: 1118–1130

I. INTRODUCTION

In cross-language developmental studies of non-native speech learning, two primary research goals are to accurately document and explain developmental changes in the ability to learn new speech sounds. To address these goals, the current study investigated how age-related differences vary along one important dimension of learning, the amount of exposure to the target sounds.

Past research on age-related differences in non-native speech learning can be classified into two main categories: laboratory studies and immersion studies. In laboratory studies, participants live in their native country and have no immersion experience[1] with the target language. Participants are exposed to the target speech sounds of a foreign language only in the study setting, usually a research laboratory. This approach allows the assessment of age differences at the initial encounter with the new sounds, and offers good control over the amount of exposure to these sounds. Laboratory studies, which have focused primarily on production, have yielded inconsistent findings. Findings from some studies support the notion of "the younger the better." When imitating Spanish words, monolingual English-speaking 7 year olds were slightly but significantly more accurate than young adults (Cochrane and Sachs, 1979). Similar findings were obtained among a group of native English-speaking 5–15 year olds when imitating French and Armenian words and phrases (Tatha, Wood, and Loewenthal, 1981a). In contrast, findings from other studies support the notion of "the older

[a)]A portion of this work was published in "Age differences in perceptual sensitivity to new speech sounds: The younger the better?" Proceedings of the 29th Boston University Conference on Language Development, Boston, November, 2004, and was presented in "Age differences in the perception and production of American English vowels by native Mandarin speakers." Poster presentation at the 1st Acoustical Society of America Workshop on Second Language Speech Learning, Vancouver, Canada, May, 2005.

[b)]Corresponding author. Electronic mail: giselajia@yahoo.com
[c)]Corresponding author. Electronic mail: wuyh@pku.edu.cn

the better." The accuracy in imitating Dutch words increased linearly with age among native English speakers ranging in age from 7 years to young adulthood (Snow and Hoefnagel-Höhle, 1977). The ability to imitate French words and discriminate French-sound pairs also increased with age among English-speaking first to ninth graders (Politzer and Weiss, 1969).

Taking advantage of the immigration phenomenon, immersion studies have examined age-related differences exhibited by immigrants as they acquire a new language in the second language (L2)-speaking country. These studies investigate the relation between age of exposure to L2, usually indexed by age of arrival (AoA) in the L2-speaking country, and learners' L2 speech perception and production abilities. Immersion studies can be conducted at different points in time along a wide spectrum of length of L2 immersion. Long-term immersion studies include L2 learners who have resided in the L2-speaking country for many years when their L2 proficiency supposedly has reached relative stability following massive L2 exposure. Short-term immersion studies include L2 learners at a more recent stage of L2 immersion. There is no clear cut division between short and long terms, as some studies adopt a five-year (e.g., Jia, Aaronson, and Wu, 2002) and others a ten-year criterion (e.g., Flege, Munro, and MacKay, 1995a).

Findings from long-term immersion studies have consistently shown that, when the length of residence in the L2 country being equal, younger arrivals obtain better L2 speech perception and production skills than older arrivals. The benefit of early arrival existed for the overall degree of perceived foreign accent by English L2 learners speaking various native languages (Asher and Garcia, 1969; Flege *et al.*, 1995a; Oyama, 1976; Yeni-Komshian, Flege, and Liu, 2000), and for the accuracy in the perception and production of American English (AE) vowels and consonants (Flege, MacKay, and Meador, 1999; Flege, Munro, and MacKay, 1995b; MacKay, Flege, Piske, and Schirru, 2001; Munro, Flege, and MacKay, 1996) by native Italian speakers. Such age-related differences in production exist at an even earlier point of L2 immersion, i.e., after about two to three years of L2 immersion (Fathman, 1975; Tatha, Wood, and Loewenthal, 1981b).

Different from long-term attainment and laboratory studies that examine the performance at the one-time point, longitudinal immersion studies track performance over time. Snow and Hoefnagel-Höhle (1977) studied native English speakers living in Holland ranging in age from three years to adulthood. At three time points with a four to five month interval, participants distinguished Dutch minimal pairs, as well as imitated and spontaneously produced Dutch words. Although there were no significant age-related differences for perception at any point, age-related differences in production changed with increasing immersion experience. At the first testing session, older children and adults did significantly better than younger children in pronouncing many vowels and consonants. At the second session, age differences in pronouncing most of the segments disappeared. At the third session, age differences became reversed, with younger children outperforming older children and adults.

More recently, Flege *et al.* (in press) studied 155 native Korean-speakers living in the U.S. and Canada. The child arrivals (AoA between 6–12 years) and adult arrivals (AoA between 21–35 years) were tested after 3–4 years and then at 5 years of residence in these countries. The adult arrivals were judged to speak English with a significantly stronger foreign accent than the child arrivals. In a subgroup of these participants ($n=108$), the ability to discriminate and imitate English vowels was examined. Child arrivals outperformed adult arrivals after both 3 and 5 years of residence on both perception and imitation (Tsukada *et al.*, 2005). Similar to Snow and Hoefnagel-Höhle (1977), these two studies demonstrated a period of younger-learner advantage. Different from Snow and Hoefnagel-Höhle, these two studies did not observe a period of older-learner advantage. This is likely due to the fact that the time 1 of the Flege *et al.* and Tsukada *et al.* studies was already after 2 years of L2 immersion, when the adult advantage could well have already disappeared. Indeed, in a longitudinal study of native Japanese speakers' perception and production of English consonants /l/, /r/, and /w/, adult arrivals performed significantly better than child arrivals after six months of L2 immersion. However, after a year, the trend was reversed (Aoyama *et al.*, 2004).

In sum, the few existing longitudinal studies suggest that age-related differences may change with increasing L2 exposure. The extent to which a study can demonstrate the crossover pattern depends on the time point(s) selected for the study. In the beginning of L2 immersion, older learners may have an advantage in discriminating and producing nonnative speech sounds. This advantage may persist in early years of L2 immersion, but disappear and become reversed later. The exact timetable of this crossover pattern may depend on which aspects of phonological skills are assessed and the difficulty level of the assessment tasks.

Three general theoretical accounts have been put forth to explain the robust early learner advantage found in long-term immersion studies: the Critical/Sensitive Period account, the L1 Transfer/Interference account, and the Environmental account. The Critical/Sensitive Period account postulates a genetically guided maturation of a domain specific language learning mechanism as the reason for the declining abilities to learn new speech sounds. That is, there is a "limit to the programming of new phonological rules at the cortical level" (Cochrane, 1980; pp. 332–333) due to factors outside the learning process (e.g., Patkowski, 1990; Scovel, 2000).

In contrast, the L1 Transfer/Interference account posits that "new processing activities are hampered by the strength of heavily utilized patterns or processing routines" (Cochrane, 1980; pp. 332–333) used in native language processing. This account has been represented by three models of cross-language speech perception, the Perceptual Assimilation Model (PAM) (Best, 1995), the Native Language Magnet Model (NLM) (Kuhl, 2000; Werker and Tees, 1999), and the Speech Learning Model (SLM) (Flege, 1995). The three models, with some nonoverlapping foci, converge on an emphasis on the role of prior L1 learning and use in L2 speech perception and production. According to PAM, non-native segments "tend to be perceived according to their similarities

to, and discrepancies with, the native segmental constellations that are in the closest proximity to them in native phonological space." (Best, 1995, p. 193). As will be referred to in detail later, this model makes specific predictions about the relative difficulty with which particular non-native segments are perceived or produced, based on their relation to the native phonological system. NLM delineates the details as to how, as early as in their first year of life, infants form a complex perceptual network through which new speech sounds are perceived, or "filtered" (Kuhl, 2000; Werker and Tees, 1999). SLM focuses in part on explaining age-related differences in learning new speech sounds. According to SLM, the greater difficulties experienced by older individuals arise from the increasingly strong influence of L1 (Flege, 1995). More specifically, with increasing age, L1 phonetic categories exert stronger assimilation power on non-native speech sounds, making the establishment of new speech categories more difficult (Baker, Trofimovich, Mack, and Flege, 2002; Flege, 2003).

The Environmental account (Jia and Aaronson, 2003; Snow, 1983) has been developed to explain the younger-learner advantage in various aspects of L2 proficiency found in long-term attainment studies. According to this account, in the immigration setting, L2 learners of various ages are inherently at different levels of cognitive, social, and cultural maturation. Such variations expose early arrivals to a significantly richer L2 environment than late arrivals, and such environmental differences accumulate and lead to L2 proficiency differences.

The validity of these three accounts relies heavily on a more accurate description of age-related differences in L2 speech learning. If the age difference crossover pattern discussed earlier proves robust, all three theories need to address these related questions. Why are early arrivals better in the long run? Why do late arrivals initially have an advantage? Why does it take time for early arrivals to catch up with and eventually surpass late arrivals? Most previous studies have focused on a limited period of L2 exposure. This has prevented firm conclusions about the interaction between age-related differences and the amount of L2 exposure. Although we can summarize trends from different studies, their sampling of different language populations and use of different methods and designs weaken the conclusion. Therefore, to shed light on the validity of the theoretical accounts, it is important to conduct further research to capture the dynamic changes of age-related differences in L2 speech learning at different points of L2 immersion with diverse populations.

To examine the interaction of age and amount of exposure to the target language, the current study included three participant groups with different amounts of exposure to native-sounding AE. The first was a group of native Mandarin speakers (chronological age at time of study 7 to 20 years) living in the People's Republic of China (hereinafter referred to as China) with no English immersion experiences. Age differences found in this group are not confounded by age-related language environment differences existing in immigrant populations

glish speakers, and few had attended supplementary English classes outside of school. The number of years of English language instruction ranged from 0 to 11 years (M=4.41 year; SD=2.81), mostly beginning in the fourth (36.4%), third (22.70%), and first (28.40%) grades.

2. Native Mandarin speakers in the US

Participants in the U.S. were 131 native Mandarin speakers who immigrated to New York City (NYC) between 7 and 44 years of age and had lived in the U.S. for fewer than 5 years. They were divided into two subgroups according to their length of U.S. residence: 54 past arrivals who had lived in the U.S. for between 3 and 5 years, and 77 recent arrivals who had lived in the U.S. for two years or less (Table I). These two groups did not differ significantly in their age, AoA, and age of onset of English instruction. They were set apart by years of residence in the U.S., and consequently, years of education in the U.S. Participants were recruited from the Chinese communities in NYC through an advertisement in a Chinese newspaper. The majority (n=99) spoke Mandarin (75.57%) as their native dialect, 14 (10.69%) spoke Min dialect, 13 (9.92%) spoke Wu dialect, and 5 (3.82%) spoke Cantonese. All non-native Mandarin speakers were exposed to Mandarin from birth, and all started speaking Mandarin regularly in school before 9 years of age. Similar to participants in China, their exposure to native-sounding English before their arrival in the U.S. was minimal.

No hearing screening was conducted for participants in China or the U.S. However, all participants reported having normal hearing in a background questionnaire described later.

B. Stimulus materials

The AE vowel inventory can be described as including 11 nonrhotic monophthongal vowels differing in height (5 levels: high, mid-high, mid, mid-low, low) and position (front versus back). The front vowels are /i, ɪ, e, ɛ, æ/ and the back vowels are /u, ʊ, o, ɔ, ʌ, ɑ/. The mid vowels /e, o/ are usually phonetically realized as diphthongal [$e^{ɪ}$, $o^{ʊ}$] in stressed syllables, mid-low /ʌ/ is unrounded and centralized relative to mid-low rounded /ɔ/, and other vowels show some diphthongization in some dialects (Peterson and Barrey, 1952; Hillenbrand, Getty, Clark, and Wheeler, 1995). The duration of AE vowels also varies phonetically, with the four short vowels [ɪ, ɛ, ʊ, ʌ] alternating with the seven long vowels [i:, e:, æ:, u:, o:, ɔ:, ɑ:] (Peterson and Lehiste, 1960). Mandarin has a smaller vowel inventory than AE. The description of the Mandarin vowel inventory has been controversial, due to different classification criteria and methods of analysis, but researchers generally adopt a six-vowel system (Howie, 1976; Lin, 1989; Wan, 1999): three high vowels—front unrounded /i/, front rounded /y/, and back rounded /u/; two mid-vowels—central /ɘ/ and back /ɤ/; and one low vowel /a/. Allophonic variations of /i/ include high and mid-high variants [i, ɪ]; central /ɘ/ varies allophonically from mid-central to mid-front [ɘ, e]; low /a/ varies allophonically from central to back variants [a, ɑ]; mid-back /ɤ/ includes both unrounded and rounded allophones [ɤ, o]. Mandarin vowels appear in open syllables /CV, V, VV/, and /Vn/ and /VVn/ syllables. Vowel duration does not distinguish vowels in Mandarin. Stimulus materials for the current study included eight AE vowels /i, ɪ, $e^{ɪ}$, ɛ, æ, ʌ, ɑ, u/, of which /i, u/ have phonetically similar counterparts in Mandarin. The other six vary in their relation to Mandarin vowels. The mid-low and low front vowels /ɛ, æ/ and the mid-low back /ʌ/ have no phonetically similar counterparts, even when allophonic variation is taken into consideration. The AE vowels /ɪ, $e^{ɪ}$, ɑ/ are phonetically similar to contextual variants of Mandarin phonemes.

For the perception task, six contrasts were formed with these eight vowels, including /i-ɪ/, /i-$e^{ɪ}$/, /ɛ-æ/, /æ-ɑ/, /ɑ-ʌ/, and /u-ɑ/. The vowel contrasts were selected to present a wide range of difficulty for native Mandarin speakers, according to data from the few studies of native Mandarin speakers (Rogers, 1997; Flege, Bohn, and Jang, 1997), and the predictions of the relevant theoretical models, such as PAM (Best, 1995), and SLM (Flege, 1995). The vowels in the /ɛ-æ/ and /ɑ-ʌ/ pairs are close in articulatory and acoustic vowel space, and neither occurs in Mandarin (though /ɑ/ occurs in Mandarin as an allophonic variant of /a/). Vowels in the /i-ɪ/ pair are close in vowel space, but /i/ occurs in Mandarin. The pairs /i-$e^{ɪ}$/ and /æ-ɑ/ are more distant in vowel space than the preceding pairs, and /i/ occurs in Mandarin. Finally, the /u-ɑ/ vowels are distant in vowel space and have distinctive counterparts in Mandarin. In terms of Best's PAM, we speculated that the first two pairs fall into a Single Category Assimilation pattern, the next three pairs into a Category Goodness pattern, or for /i-$e^{ɪ}$/, possibly a Two-Category pattern, and the final contrast is a clear Two-Category pattern. PAM predicts the order of difficulty for these contrasts as (from most to least difficult): /ɛ-æ/, /ɑ-ʌ/, /i-ɪ/, /æ-ɑ/, /i-$e^{ɪ}$/, and /u-ɑ/.

The selected vowels were situated in /dV-pə/ disyllables spoken in citation form. The use of nonsense disyllables rather than real yreala

darin. The first two sets were used for practice, and the other eight sets were analyzed for five acoustic parameters of the target vowel (VOT, length, pitch, and Fl and F2 values), and two acoustic parameters of the nontarget vowel /ə/ (VOT and length). For each target vowel, three tokens were selected out of the eight tokens (see the Appendix). In order for a token to be selected, the target vowel had to have a minimum of four acoustic parameter values within the 95% confidence interval of the mean, and the nontarget vowel had to have as many as possible (ranging from 0–2) acoustic parameter values within the 95% confidence interval.

C. Design and procedure

1. Perception

Perception accuracy was assessed using a categorial (name identity) AXB discrimination task. This task was chosen among several discrimination tasks because it avoids the possibility of an age-related criterion shift found in same-different judgment tasks (Beving and Eblen, 1973) and possible difficulties that young children may have in understanding the concepts of "same" and "different." Further, an AXB task poses less memory and processing demands than the other two triplet formats (Oddity, ABX) because the middle target stimulus is next to both comparison stimuli (MacKain, Best, and Strange, 1981).

Each vowel pair was tested with 12 trials, 3 trials for each of the 4 possible position combinations (AAB, ABB, BAA, BBA). This resulted in 72 trials for the whole test (6 pairs $\times$ 4 position combinations $\times$ 3 trials). The 72 trials were presented in 6 blocks of 12 trials. Each vowel pair appeared twice in each block. The order of blocks and trials within each block were randomized across participants. Each of the three selected tokens of a vowel was used the same number of times. Vowel positions were also balanced within and across blocks. The two same vowels in each AXB triplet were always two physically different stimulus tokens. This allowed us to test categorial perception at the minimum level, though not to the full extent as no differences in speakers or consonantal context were included.

A block of 12 trials with five Mandarin vowels /i, y, ɘ, a, u/ designed in exactly the same format was presented before the test to familiarize participants with the task as well as to screen participants. Participants who made three errors or more were allowed to proceed with and complete the entire study, but their data were not included in analyses. According to the above criterion, four participants in China (one 8-year-old, two 9-year-olds, and one 15-year-old) were excluded from data analyses, leaving 87 participants for this group.

The AXB task was conducted using specialized computer software (written by Bruno Tagliaferri) available in the Speech Acoustics and Phonetics Laboratory (SAPL) at the CUNY Graduate Center. Each stimulus triad was preceded by a tone presented 300 ms prior to the first stimulus. After listeners heard the three disyllables (ISI=500 ms), two boxes appeared on the screen. The left one read "1," and the right box read "3." Participants were instructed to click "1" if they decided that the middle disyllable sounded like the first one, and click "3" if the middle one sounded like the third one. Once the participants clicked "1" or "3," the next trial was triggered, with a 1000 ms intertrial interval. The trial and test sessions together took between 10 and 15 min. After each block of 12 trials, participants were offered the choice to take a break, although no participant chose to do so. All participants were tested individually, listening to the stimuli through earphones with volume adjusted to a comfortable level for the individual.

Participants in China were tested in a quiet office in their schools in Beijing, on a 15-in. screen portable PC. Participants in U.S. were tested in a soundproof room in the CUNY laboratory, using a 19-in. screen desktop PC.

2. Production

Prior to the discrimination task, participants imitated each of the eight /dV-pə/ stimuli (/dæpə/, /dɛpə/, /dʌpə/, /dɑpə/, /dɪpə/, /dipə/, /deɪpə/, and /dupə/) three times consecutively, each time immediately after hearing the target disyllable. The production tokens were directly recorded as digitized sound files (22.05 kHz, 16-bit resolution) and then normalized for peak amplitude using Sound Forge. The files were further processed for an identification task by native English speakers. The files were first sliced into separate sound files each with one disyllable. Then, the nontarget vowel in each disyllable was removed by deleting all portions of the signal following the beginning of the /p/ stop closure defined as the cessation of upper formant energy. The aim of the editing was to eliminate the potential distraction of the nontarget vowel from the focus on the target vowel. Finally, each file was duplicated so listeners heard each stimulus twice. The time interval between the repetitions was 1000 ms.

For the purposes of token and response choice selections, a pilot identification task was conducted. Three native English-speaking listeners with IPA knowledge heard all three tokens of each vowel produced by the Mandarin speakers in China. A total of 16 AE monophthongs and diphthongs were used as response choices. Among the three tokens produced for each vowel, the second token elicited the highest agreement rate among the judges, and also yielded the most consistent identification results with both the first and the third token. Therefore, to reduce the amount of testing time, only the second repetition of each vowel was selected for the final task. Further, four of the 16 response choices that were never chosen by any listener were eliminated from the final identification choices.

There were a large number of clipped sound files for the participants in China. To counter their tendency to speak softly during the recording, they were instructed to speak loud, risking some signals being clipped. The productions of 42 participants in China and 127 participants in the U.S. who had at least one good token of each vowel were used. This yielded 168(42+126) $\times$ 8 utterances. These utterances were blocked by speakers, with 8 trials in each block. The productions were divided into four sessions with an approximately equal number of blocks. Each session had similar proportions of tokens produced by speakers from each group (speakers in China, recent arrivals and past arrivals), age (for

speakers in China), AoA (for speakers in the U.S.), and gender. When presented to the listeners, the order of the blocks and trials within a block were all randomized separately for each listener.

The 1344 utterances (168 participants × 8 vowels) were presented to five native speakers of English with a mean age of 39.4 years. Three listeners grew up in NYC and spoke English with the local accent. The other two were raised in Chicago or New Jersey, but both were familiar with New York City accent. All had IPA knowledge but were not experienced phoneticians. All listeners reported normal hearing. They listened to the tokens individually in an IAC acoustic chamber using customized software (written by Bruno Tagliaferri) that controlled stimulus presentation and recorded responses to an Excel data form. They completed two sessions on each of two separate days with a brief break between sessions. Listeners heard the stimuli through headphones at a comfortable level. They were instructed to pay attention to the vowel in the syllable, and identify, among the 12 orthographic labels and IPA symbols ("deep /dip/," "dip /dɪp/," "dape (date) /deᵗp/," "dep (debt)/dɛp/," "dap (dash) /dæp/," "dop (dock) /dɑp/," "dup (duck) /dʌp/," "dawp (dawn) /dɔp/," "dope (doze) /dop/," "doop (food) /dup/)," "dUp (could) /dʊp/," "dype (diaper) /daɪp/"), the one that sounded closest (though maybe not identical) to the token just heard. Before the test, listeners completed five practice blocks of 40 trials (5 speakers × 8 tokens) to familiarize themselves with the task. For the five speakers whose productions were used for the practice blocks, one was a monolingual English speaker who produced the stimuli for the current study, four were native Mandarin speakers (one adult male, one adult female, one child male, and one child female). Their imitation of the nonsense disyllables were elicited in exactly the same condition as the formal participants. Responses to these practice trials were not included in the data analyses. All five native listeners identified all the tokens of the monolingual English speaker correctly. Due to dialect influences, two additional native English listeners each made one or two errors identifying the monolingual tokens. These two listeners did not proceed with the identification study.

3. Background questionnaire

After the production and perception sessions, all participants filled out a background questionnaire. The questionnaires for participants in China and NYC were not identical but had overlapping items. The common items included gender, birth date, birth place, places where participants had lived, and any known hearing and health problems. Participants in China, in addition, listed their current school grade, the grade that English instruction began, and the number of hours of English classes in each week. Participants in NYC provided information about their age when English language instruction began, and their age of arrival in the U.S. Children and adolescents living with their parents rated their mothers' and fathers' English speaking ability along a 1–7 point scale (1=cannot speak English at all; 7=speak English as fluently as a native English speaker). They also reported the percentage of time that their father, mother, and siblings spoke to them in English, and the percentage of time that they watched TV and videos in English. All the above items regarding parents' English proficiency and language use were rated for every year that participants were in the US. The average English use in a situation across all the years of U.S. residence was calculated for use in the statistical analyses.

III. RESULTS

The results are organized into three sections: perception, production, and the relation between perception and production. For both perception and production, performance accuracy was compared among the groups and across the vowel pairs (perception) or vowels (production) using mixed Analyses of Variance. Age-related differences were examined within each group using bivariate correlations, and other predictive variables were also investigated using regression analyses. Correlation and regression analyses were chosen over age group analyses because the former treats age as a continuous variable and maximizes its variance. The relation between perception and production was examined at the individual level as indicated by correlations between performance on perception and production, and at the group level by the extent to which the rank order of vowel pair (or vowel) difficulties matched in perception and production. In all results of the Analyses of Variance (ANOVA) reported below, the effect size (ES) is indicated by eta-squared values (η^2).

A. Perception

1. Accuracy across groups and vowel pairs

Perception accuracy was indicated by the percentage of correct responses out of the total 72 trials (for total accuracy), or the 12 trials (for each vowel contrast). Performance accuracy for the total task and for each vowel pair was compared across three participant groups. All three groups performed well above chance level with over 70% accuracy for all contrasts (Table II). A mixed two-way 6×3 ANOVA was conducted, with the 6 vowel pairs as the within-subjects variable, and the three participant groups as the between-subjects variable. The results revealed a main effect of group, $F(2,215)=53.18$ ($\eta^2=0.33$), a main effect of pairs, $F(4,862)=101.98$($\eta^2=0.32$; with Greenhouse–Geisser correction of degrees of freedom), and an interaction between group and pairs, $F(8,862)=10.87$ ($\eta^2=0.09$; with Greenhouse–Geisser correction of degrees of freedom) (all $p<0.001$).

The main group effect reflects differential performance across the three groups. Pairwise comparisons (with Bonferroni corrections) indicate that participants in China had significantly lower accuracy than the recent and past arrivals. To further examine the group effect for each vowel pair, separate one-way ANOVA was performed for the performance on individual vowel pairs. There was a significant effect of group for all vowel pairs, including /i-eᵗ/, $F(2,215)=26.97$, /æ-ɑ/, $F(2,215)=11.66$, /ɑ-ʌ/, $F(2,215)=32.46$, /i-ɪ/, $F(2,215)=43.25$, /ɛ-æ/, $F(2,215)=32.77$, and /u-ɑ/, $F(2,215)=8.29$ (all $p<0.001$). Bonferroni post-hoc tests re-

TABLE II. Performance on the six contrasts by native Mandarin speakers in China (monolinguals) (n=87), recent arrivals (n=77), and past arrivals (n=54).

Vowel pairs	Monolinguals (n=87) % correct (SD; range)	Recent arrivals (n=77) % correct (SD; range)	Past arrivals (n=54) % correct (SD; range)
/i-ɪ/	82.6 (16.9; 33.3–100)	97.4 (6.4; 66.7–100)	97.8 (5.4; 66.7–100)
/i-eɪ/	90.2 (14.3; 41.7–100)	99.5 (2.1; 91.7–100)	99.7 (1.6; 91.7–100)
/ɛ-æ/	76.3 (14.2; 41.6–100)	89.4 (12.9; 33.3–100)	91.8 (9.5; 58.3–100)
/æ-ɑ/	88.9 (14.4; 41.7–100)	96.1 (8.7; 50.0–100)	96.6 (7.7; 58.3–100)
/ɑ-ʌ/	71.7 (16.5; 33.3–100)	85.2 (16.5; 25.0–100)	91.4 (9.4; 58.3–100)
/u-ɑ/	97.9 (4.6; 75.0–100)	99.7 (1.6; 91.7–100)	99.7 (1.6; 91.7–100)
Overall	84.6 (10.4; 55.6–98.6)	94.5 (5.5; 69.4–100)	96.2 (3.3; 87.5–100)

vealed that, for all vowel pairs, both the recent and past arrival groups scored significantly higher than the China group, and there were no significant differences between the two immigrant groups, probably due to ceiling effects.

Regarding the main effect of pairs, pairwise comparisons (with Bonferroni corrections) revealed that scores (averaged across the three participant groups) on most pairs of vowel contrasts (except for /æ-ɑ/ and /i-ɪ/) were significantly different (all $p<0.001$). The interaction effect of group and pair indicates that the performance difference in pairs varied among the three groups. To further examine this effect, paired-sample T tests were conducted separately for each participant group to compare the performance on each pair of vowel contrast. For the participants in China, only one difference between vowel contrasts (/i-eɪ/-/æ-ɑ/) were not significant, and all other 14 pairs were significant (all $p<0.01$). For the recent arrivals, two difference scores (/i-eɪ/-/u-ɑ/; /i-ɪ/-/æ-ɑ/) were not significant. For the past arrivals, three difference scores (/i-eɪ/-/u-ɑ/; /i-ɪ/-/æ-ɑ/; /ɛ-æ/-/ɑ-ʌ/) were not significant. In terms of the rank order of performance, the two most difficult pairs (/ɛ-æ/ and /ɑ-ʌ/) and the two easiest pairs (/i-eɪ/ and /u-ɑ/) were the same for all three groups. The difficulty order for two medium-level performance pairs (/i-ɪ/ and /æ-ɑ/) was the opposite for participants in China and the U.S.

2. Age differences

The age variable of interest is the age of L2 exposure. For recent and past arrivals, it was indicated by AoA in the L2-speaking country. For participants in China, it was indicated by chronological age at the time of the study, which coincides with AoA, as they could be regarded as a group of immigrants on their first day of arrival in the U.S. Participants in China showed significant positive correlations between age and performance on the total task (r=0.51, $p<0.001$) and on all the individual vowel contrasts (r=0.39, $p<0.001$ for /i-ɪ/, r=0.38, $p<0.001$ for /i-eɪ/, r=0.37, $p<0.001$ for /ɛ-æ/, r=0.43, $p<0.001$ for /æ-ɑ/, r=0.41, $p<0.001$ for /ɑ-ʌ/, and r=0.28, $p<0.01$ for /u-ɑ/), indicating that older participants, in general, achieved a higher level of accuracy (Fig. 1). However, recent arrivals showed no significant correlations between AoA and overall performance or individual vowel pairs. In contrast, past arrivals showed negative correlations between AoA and overall performance (r=−0.41, $p<0.01$), and two of the more difficult vowel contrasts (r=−0.36, $p<0.01$ for /æ-ɑ/, and r=−0.40, $p<0.01$ for /ɑ-ʌ/), a trend opposite that of the participants in China. That is, a younger AoA predicted significantly better performance on the task in general, and on the difficult vowel contrasts.

3. Other predictors

To pinpoint the unique predictive power of AoA, several other potential predictor variables of performance were also examined. For the participants in China with little variance in the English environment, the major variable was length of English instruction. Older participants had significantly more years of English instruction, r=0.89, $p<0.001$, and more years of English instruction predicted a better task performance, r=0.48, $p<0.001$. A partial correlation analysis was conducted to examine the unique contribution of chronological age (AoA in our definition) and length of English instruction. When the length of English instruction was partialed out, there was still a marginally significant relation between age and the total percentage correct, r=0.19, p=0.07. When age was partialed out, the correlation between years of English instruction and the total percent correct became nonsignificant (r=0.07, p=0.50).

The predictor variables for the two immigrant groups included the age that English instruction began, the length of U.S. residence, the length of U.S. education, parents' English speaking abilities, and the percentage of English use in various situations. Bivariate correlations between total accuracy and all of these predictive variables were obtained for each

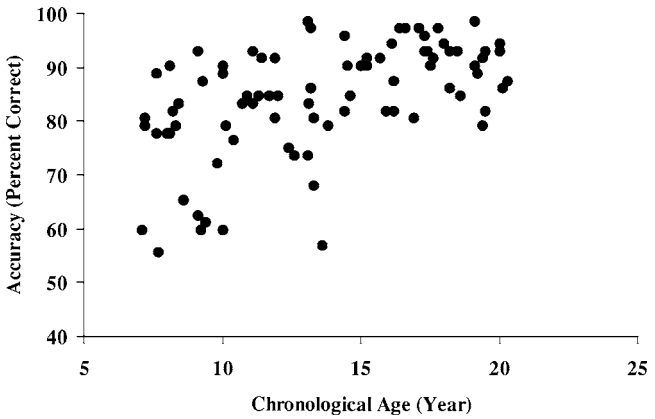

FIG. 1. Scatter plot of age and total accuracy (percentage correct) for native Mandarin speakers in China (n=87; r=0.51, $p<0.001$).

group. For recent arrivals, only one significant correlation emerged: those who had spoken more English with their friends tended to perform better on the task, $r=0.31$, $p<0.01$. For past arrivals, better performance on the task was associated with a younger age at which English instruction began, $r=-0.55$, $p<0.001$, more years of U.S. education, $r=0.40$, $p<0.01$, and better English speaking ability of mothers, $r=0.42$, $p<0.05$. To further detect the unique predictive power of the four significant predictors for past arrivals, a hierarchical regression analysis was conducted. AoA and the age of English instruction were entered in the first step, followed by years of education in the U.S., and then the mother's English speaking ability. The two age variables accounted for 20% of the variance, $p<0.05$. Adding U.S. education did not change the amount of variance explained, but adding the mother's English speaking ability significantly increased it to 33%, $p<0.01$.

B. Production

The listeners showed high agreement rates on the produced vowel identity. Of the 1344 vowel tokens (168 participants $\times$ 8 vowels), five listeners agreed on 617 (45.90%) of the tokens. Another 331

TABLE IV. Confusion matrix for the vowel productions by participants in China (first row; $n=42$), recent arrivals (second row; $n=76$), and past arrivals (third row; $n=50$).

Stimulus vowels (vowel said)	Response vowels (vowel heard)									
	i	ɪ	eᴵ	ɛ	æ	ɑ	ʌ	u	ɔ	o/ʊ/ai
i	**88**	10	2.0							
	88	10	2.0							
	89	6	3					2		
ɪ	23	**52**	5	9	<1	<1	<1	7	<1	1
	16	**76**	4	3	<1		<1			<1
	12	**77**	4	6						1
eᴵ	26	18	**54**							1
	2	1	**89**	1					<1	6
	<1	4	**88**	2						6
ɛ	1	8	<1	**65**	12	2	10			1
		3	2	**66**	20	2	5		<1	1
	<1	2	1	**72**	18	<1	5			2
æ				24	**52**	12	11			
		<1	2	23	**65**	7	1			
			1	22	**69**	5	2			1
ɑ					7	**51**	28		11	3
				<1	1	**77**	13		6	3
	<1			2	<1	**74**	17		6	
ʌ		<1		3	5	25	**54**		6	7
	<1			2	2	43	**45**		5	3
				<1	<1	40	**49**		6	4
u								**97**		3
						<1		**93**		7
	2							**95**		3

$=0.34$, $p<0.05$] and /ɛ/ [$r=0.34$, $p<0.05$]. No significant correlation was found for the recent arrivals. For the past arrivals, performance on two vowels, /ɪ/ [$r=-0.24$, $p<0.05$] and /eᴵ/ [$r=-0.33$, $p<0.01$] showed significant negative correlation with AoA, a trend opposite that of the participants in China.

3. Error patterns

The overall error patterns were analyzed by creating confusion matrices for the three groups (Table IV). Responses were classified by the 8 target (intended) vowels contained in each of the /dVp/ utterances produced by participants, and by the 12 vowels given as the response alternatives. The numbers on a row indicate the percentage of instances an intended vowel (produced by all participants) was identified as one of the 12 vowels by the native listeners. The proportion of target and response matches (diagonal bold numbers on Table IV) was regarded as the accuracy score for each vowel.

The four vowels with the lowest accuracy rates (/ɛ, æ, ɑ, ʌ/) showed bidirectional confusion patterns, with the two vowels tested as discrimination pairs (/ɛ, æ/ and /ɑ, ʌ/) being highly confused with each other. However, although /ɛ/ or /æ/ were misidentified as each other in approximately equal proportions of the instances (17.4% and 22.8%, respectively), /ʌ/ was more often misheard as /ɑ/ (38%) than the opposite (18%). Vowels /u/ and /i/ had the highest accuracy scores. In between, /ɪ/ showed a concentrated confusion pattern, being most often heard as /i/. In contrast, /eᴵ/ showed a more diffuse confusion pattern, heard as /i/, /ɪ/, or even /aᴵ/. For both /ɪ/and /eᴵ/, the immigrant groups showed considerable improvement in production accuracy.

TABLE V. Ranking of the bidirectional production error rate and discrimination accuracy for the six vowel pairs.

Vowel pairs	Production Bidirectional error rate/Difficulty rank		Perception Accuracy/Difficulty rank	
/i-ɪ/	8.5+16.3=24.8	3	94.93%	3
/i-eᴵ/	2.2+7.8=10	5	98.20%	5
/ɛ-æ/	22.8+17.4=40.2	2	95.37%	4
/æ-ɑ/	7.9+2.5=10.4	4	87.62%	2
/ɑ-ʌ/	18+38=56	1	83.77%	1
/u-ɑ/	0.1+0=0.1	6	99.37%	6

C. Relation between perception and production

The relation between perception and production at both the individual level and group level was examined. The individual level relation was assessed by correlating perception and production total accuracy scores for the 168 native Mandarin speakers with measurable production data. There were significant positive correlations between perception and production performance for all participants together ($r=0.50$, $p<0.001$), for the participants in China ($r=0.42$, $p<0.001$), and for the past arrivals ($r=0.46$, $p<0.01$). The correlation for the recent arrivals was lower ($r=0.25$, $p=0.08$). The close to ceiling perception performance of the recent and past arrivals might have lowered the correlations.

At the group level, rank orders of difficulty in perception and production were compared (Table V). For production, we combined the performance for the eight vowels into the six vowel contrasts by obtaining the bidirectional confusion rates (summing up the percentage of time that one vowel was identified as the other for each pair) and ranking these numbers. For perception, we rank ordered the average perception correct scores for all the 168 participants who had production scores. The bivariate correlation of the two sets of ranked scores was marginally significant ($r=0.77$, $p=0.07$).

IV. DISCUSSION

In the current study we investigated how age-related differences in the perception and production of AE vowels changed with an increasing amount of AE exposure. We included three groups of native Mandarin speakers with varying amounts of L2 exposure: those with no L2 immersion experiences who represented a population of potential immigrants on their first day of arrival in the U.S., those with moderate L2 immersion experiences (in the U.S. for two years or less), and those with substantial L2 immersion experiences (in the U.S. for between three and five years). To assess the unique contribution of our focus variable, AoA, other potential predictor variables of L2 learning were also examined. The inclusion of AE vowels that bear different phonetic relations to Mandarin vowels permitted the investigation of the influence of L1 phonetic/phonological system on L2 phonological acquisition. The findings add to a more

accurate description of age-related differences in L2 phonological learning, and call for a more refined theoretical account of the phenomenon.

With increasing L2 use, age differences in performance accuracy changed from an older-learner advantage to a younger-learner advantage for both perception and production. For the participants in China with no L2 immersion experiences, an older chronological age predicted a significantly higher discrimination accuracy of all vowel contrasts and higher production accuracy of two difficult vowels.[2] For the recent arrivals, AoA was not related to performance at all. For the past arrivals, a younger AoA predicted significantly better discrimination accuracy for three vowel contrasts, and better production accuracy for two vowels.

The interaction of age-related differences with the amount of L2 exposure is consistent with the earlier study that demonstrated this full crossover pattern (Snow and Hoefnagel-Höhle, 1977

influences of the L1 vowel system on L2 vowel learning serves as indirect evidence for the L1 Transfer/Interference account. Difficulty rankings for perception of vowel contrasts and production of vowels were similar across the three participant groups. For perception, the order of difficulty closely reflected the hypothesized order based on both phonetic similarity and hypothesized perceptual assimilation patterns influenced by L1 vowel space (Best, 1995). In the two most difficult pairs /ɛ-æ/ and /ɑ-ʌ/, the two vowels involved in each do not have close counterparts in Mandarin (not including allophonic variations), and are close in vowel space. Larger acoustic distances (i.e., /æ-ɑ/) or the presence of one of the two vowels in Mandarin (i.e., /i-eI/, /i-I/) was associated with medium level performance. Similarly for vowel production, /ɛ, æ/, with no close Mandarin counterparts, showed symmetrical confusions. AE /ɑ, ʌ/ were also confused in production, although the confusions were asymmetrical favoring /ɑ/. The two vowels that had corresponding LI counterparts /i, u/ showed close to ceiling accuracy in intelligibility, even when produced by speakers with no L2 immersion experiences.

The current study yielded a positive correlation between perception and production at the individual and group levels. At the individual level, better perception performance significantly predicted better production performance. At the group level, the vowel contrasts that were harder to distinguish in the perception task also had the highest bidirectional confusion rate in production. Vowel contrasts that were better distinguished were also produced with greater accuracy. These findings are consistent with those in the literature. Flege and colleagues found similar positive correlations between English vowel intelligibility and discrimination among native speakers of various languages (Flege *et al.*, 1997; 1999). For example, both native Korean and Mandarin speakers identified synthetic vowels along the bat-bet (/æ-ɛ/) and beat-bit (/i-ɪ/) continua differently from native English speakers, and produced the two vowels in a contrast with bidirectional confusion (Flege *et al.*, 1997). We note that the nature of such a positive relation between perception and production is still controversial. According to SLM, accurate L2 production to a large extent relies on accurate perception, and thus, perception development should precede production (Flege, 1995; McAllister, Flege, and Piske, 2002). Other researchers emphasize the causal role of production in perception. For example, Japanese speakers' production of English /r/ and /l/ was more accurate than their perception (e.g., Sheldon and Strange, 1982). However, perceptual training on the /r-l/ contrast did lead to production improvement by Japanese speakers (Bradlow, Akahane-Yamada, Pisoni, and Tohkura, 1999). Notably, tasks measuring production and perception abilities may be inherently incommensurable (e.g., Flege, 1999; Tsukada *et al.*, 2005). They can pose varying levels of processing demands by the choice of stimuli, tasks, and procedures. The current production task promoted optimal performance with minimal processing demands. Only one consonantal context was used, and productions were rated in terms of intelligibility rather than degree of foreign accent (the latter is more stringent than the former, as found in studies such as Munro *et al.*, 1996). Given this, several of our findings are consistent with the predictions of the SLM that production abilities at least partially rely on perception abilities. First, some Mandarin speakers were able to distinguish some vowel pairs accurately but confused them in production, indicating that production indeed lagged behind perception. Second, perception abilities improved at a faster rate than production abilities.

There are several limitations of the current study that can be addressed by future research. First, we manipulated the length of L2 exposure cross-sectionally rather than longitudinally. Some aspects of the participants in China and those in the U.S. were not completely comparable, such as the testing environments and Chinese dialect backgrounds. Nevertheless, the incomparability between the groups would have mainly affected our interpretations of the between group differences, not the age trends within each group. Second, in order to access optimal performance, the current study minimized speaker and contextual variations of the stimuli as well as the processing demands of the tasks. Future research should increase the variations along these dimensions to more closely approximate "on-line" phonological processing and learning.

In summary, the current findings indicate that age and amount of L2 immersion jointly influenced learning, indicated by a dynamic change of age-related differences with increasing exposure to L2. These findings support a combined Environmental and L1 Interference/Transfer theory as an explanation for the long-term younger-learner advantage in mastering L2 phonology. With increasing age, the growing influence of L1 perception and production patterns, coupled with L2 input of lesser quantity and quality, leaves the long-term achievement in L2 phonology of most older arrivals behind that of the younger arrivals. Our findings also indicate that older learners have their unique advantages in nonnative speech learning. Future research should investigate how the strengths and weaknesses of younger and older learners interact in the learning processes, and tailor L2 speech learning and training strategies to learners of all ages.

ACKNOWLEDGMENTS

Research reported in this article was supported by a City University of New York collaborative grant (No. 80209-01-08) awarded to the first and second authors, and a grant from the National Institute of Health (SCORE/NICHD #41353-11-19/20) to the first author. We thank Alan Wu for collecting the data from participants in New York City. We are grateful to the following colleagues who offered methodological advice at various points of the project, or read drafts of the manuscript: Jim Jenkins, Richard Bock, Akiko Fuse, Anne Reid, Keith Happaney, Erika Levy, Kanea Nishi, Puisan Wang and Shari Berkowitz. We also thank Bruno Tagliaferri for his technical support.

APPENDIX: ACOUSTICAL CHARACTERISTICS OF THE VOWEL STIMULI (AVERAGE VALUES OF THE THREE TOKENS FOR EACH VOWEL)

Stimuli	Duration	F1 (Hz)	F2 (Hz)
/i/	122	279.6	2679.4
/ɪ/	93.4	482.5	2192.6
/e[ɪ]/	157.3	529.1	2438.1
/ɛ/	127.3	723.8	2107.9
/æ/	176.2	969.3	1866.7
/ɑ/	168.0	1007.4	1451.8
/Λ/	128.2	876.2	1540.7
/u/	121.0	304.8	1557.7

Note. Acoustic values for Mandarin vowels in dVp context are not available, and those for Mandarin vowels in contexts not identical to that of AE are not included in this table because direct comparison between vowels in different contexts are inappropriate (Strange *et al.*, 1976).

[1]In this article, immersion refers to the holistic setting of learning a second language, L2, in the immigration setting, both in and outside school. It carries a different meaning from the concept of immersion in education, which refers to L2 acquisition through natural contact with L2 outside school as opposed to classroom instruction.

[2]The older learner advantage in perception is likely due to multiple factors. First, it is partially due to growth in language related speech perception abilities. Studies of auditory perception development have shown a significant growth in the perceptual sensitivity to elements of speech signals related to language, such as signal duration (e.g., Elfenbein, Small, and Davis, 1993; Jensen and Neff, 1993). Second, it is partially due to growth of perceptual sensitivity that is not specific to cross-language phonetic influence. Tsukada *et al.* (2005) found that monolingual English-speaking children scored significant lower than monolingual-English speaking adults in discriminating /ɑ-Λ/, but not /æ-ɛ/, /i-ɪ/, and /e[ɪ]-ɛ/. Third, it is partially due to more years of English education, as partialling out years of English education reduced the significant growth trend to a marginally significant level. The older learner advantage is unlikely to be due to disadvantages of younger children to handle the tasks. Younger children passed the practice Mandarin trials in a similar proportion as older participants did. The task duration was only 10–15 min. Younger children performed equally well on the six blocks, showing no indication of attention decline.

Aoyama, K., Flege, J. E., Cuion, S. G., Akahane-Yamade, R., and Yamada, T. (**2004**). "Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/," J. Phonetics **32**, 233–250.

Asher, J., and Garcia, R. (**1969**). "The optimal age to learn a foreign language," Mod. Lang. J. **53**, 334–341.

Baker, W., Trofimovich, P., Mack, M., and Flege, J. (**2002**). "The effect of perceived phonetic similarity on non-native sound learning by children and adults," in *Proceedings of the 26th Annual Boston University Conference on Language Development*, edited by B. Skarabela, S. Fish, and A. Do (Cascadilla P., Somerville, MA), pp. 36–47.

Best, C. (**1995**). "A direct realistic perspective on cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York P, Baltimore, MD), pp. 171–206.

Beving, B., and Eblen, R. E. (**1973**). '"Same" and "different" concepts and children's performance on speech sound discrimination,' J. Speech Hear. Res. **16**, 513–517.

Bohn, O. S., and Flege, J. E. (**1992**). "The production of new and similar vowels by adult German learners of English," Stud. Second Lang. Acquis. **14**, 131–158.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. (**1999**). "Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production," Percept. Psychophys. **61**, 977–985.

Cochrane, R. (**1980**). "The acquisition of /r/ and /l/ by Japanese children and adults learning English as a second language," J. Multiling. and Multicul. Dev. **1**, 331–360.

Cochrane, R., and Sachs, J. (**1979**). "Phonological learning by children and adults in a laboratory setting," Lang Speech **22**, 145–149.

Elfenbein, J. L., Small, A. M., and Davis, J. M. (**1993**). "Developmental patterns of duration Discrimination," J. Speech Hear. Res. **36**, 842–849.

Fathman, A. (**1975**). "The relationship between age and second language learning productive ability," Lang. Learn. **25**, 245–253.

Flege, J. E. (**1995**). "Second language speech learning theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York P., Baltimore, MD), pp. 233–277.

Flege, J. E. (**1999**). "The relation between L2 production and perception," in *Proceedings of the 14th International Congress of Phonetic Sciences*, edited by J. Ohala, Y. Hasegawa, M. Ohala, D. Granveille, and A. Bailey (Department of Linguistics, University of California at Berkeley, Berkeley, CA), pp. 1273–1276.

Flege, J. E. (**2003**). "Assessing constraints on second-language segmental production and perception," in *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, edited by A. Meyer and N. Schiller (Mouton de Gruyter, Berlin, Germany), pp. 319–355.

Flege, J. E., Birdsong, D., Bialystok, E., Mack, M., Sung, H., and Tsukada, K. "Degree of foreign accent in English sentences produced by Korean children and adults," J. Phonetics (in press).

Flege, J. E., Bohn, O. S., and Jang, S. (**1997**). "Effects of experience on non-native speakers' production and perception of English vowels," J. Phonetics **25**, 437–470.

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

Flege, J. E., Munro, M. J., and MacKay, I. R. A. (**1995a**). "Effects of age of second-language learning on the production of English consonants," Speech Commun. **16**, 1–26.

Flege, J. E., Munro, M. J., and MacKay, I. R. A. (**1995b**). "Factors affecting strength of perceived foreign accent in a second language," J. Acoust. Soc. Am. **97**, 3125–3134.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111

Howie, J. M. (**1976**). "*Acoustical Studies of Mandarin Vowels and Tones*" (Cambridge University Press, New York).

Jensen, J. K., and Neff, D. L. (**1993**). "Development of basic auditory discrimination in preschool children," Psychol. Sci. **4**, 104–107.

Jia, G., Aaronson, D., and Wu, Y. H. (**2002**). "Long-term language attainment of bilingual immigrants: Predictive factors and language group differences," Appl. Psycholing. **23**, 599–621.

Jia, G., and Aaronson, D. (**2003**). "A longitudinal study of Chinese children and adolescents learning English in the US," Appl. Psycholing. **24**, 131–161.

Kent, R. D. (**1992**). "The biology of phonological development," in *Phonological Development: Models, Research, Implications*, edited by C. A. Ferguson, L. Menn, and C. Stoel-Gammon (York P., Timonium, MD), pp. 65–90.

Kewley-Port, D., Akahane-Yamada, R., and Aikawa, K. (**1996**). "Intelligibility and acoustical correlates of Japanese accented English vowels," *Proceedings of Int. Conference on Spoken Lang. Process*. Vol. **96**, pp. 450–453.

Kuhl, P. (**2000**). "A new view of language acquisition," Proc. Natl. Acad. Sci. U.S.A., **97**, 11850–11857.

Lin, Y. H. (1989). "Autosegmental treatment of segmental process in Chinese phonology," unpublished Ph.D. thesis, University of Texas at Austin.

MacKain, K. S., Best, C. T., and Strange, W. (**1981**). "Categorical perception of English /r/ and /l/ by Japanese bilinguals," Appl. Psycholing. **2**, 369–390.

MacKay, I. R. A., Flege, J. E., Piske, T., and Schirru, C. (**2001**). "Category restructuring during second-language speech acquisition," J. Acoust. Soc. Am. **110**, 516–528.

McAllister, R., Flege, J. E., and Piske, T. (**2002**). "The influence of L1 on the acquisition of Swedish vowel quantity by native speakers of Spanish, English and Estonian," J. Phonetics **30**, 229–258.

Munro, M., Flege, J., and MacKay, I. (**1996**). "The effect of age of second language learning on the production of English vowels," Appl. Psycholing. **17**, 313–334.

Oyama, S. (**1976**). "A sensitive period for the acquisition of a nonnative phonological system," J. Psycholing. Res. **5**, 261–283.
Patkowski, M. S. (**1990**). "Age and accent in a second language: A reply to James Emil Flege," Appl. Linguist. **11**, 73–89.
Peterson, G. E., and Barney, H. L. (**1952**). "Control method used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Peterson, G. E., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **32**, 693–703.
Politzer, R. L., and Weiss, L. (**1969**). "Developmental aspects of auditory discrimination, echo response and recall," Mod. Lang. J. **53**, 75–85.
Rogers, C. L. (1997). "Intelligibility of Chinese-accented English," unpublished Ph.D. thesis, Indiana University, Bloomington.
Scovel, T. (2000). "A critical review of the critical period research," Ann. Review Appl. Ling. **20**, 213–223.
Sheldon, A., and Strange, W. (**1982**). "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception," Appl. Psycholing. **3**, 243–261.
Snow, C. E. (**1983**). "Age differences in second language acquisition: Research findings and folk psychology," in *Second Language Acquisition Studies*, edited by K. M. Bailey, M. H. Long and S. Peck (Newbury House, Rowley, MA), pp. 141–150.
Snow, C. E., Hoefnagel-Höhle, M. (**1977**). "Age differences in the pronunciation of foreign sounds," Lang. Speech **20**, 357–365.
Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S., Nishi, K., and Jenkins, J. (**1998**). "Perceptual assimilation of American English vowels by Japanese listeners," J. Phonetics **26**, 311–344.
Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **79**, 1086–1100.
Tahta, S., Wood, M., and Loewenthal, K. (**1981a**). "Age changes in the ability to replicate foreign pronunciation and intonation," Lang Speech **24**, 363–372.
Tahta, S., Wood, M., and Loewenthal, K. (**1981b**). "Foreign accents: Factors relating to transfer of accent from the first language to a second language," Lang. Speech **42**, 265–272.
Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., and Flege, J. E. (**2005**). "A developmental study of English vowel production and perception by native Korean adults and children," J. Phonetics **33**, 263–290.
Walley, A. C., and Flege, J. E. (**1999**). "Effect of lexical status on children's and adults' perception of native and non-native vowels," J. Phonetics **27**, 307–332.
Wan, I. P. (1999). "Mandarin phonology: Evidence from speech erros," unpublished Ph.D. thesis, State University of New York at Buffalo.
Werker, J. F., and Tees, R. C. (**1999**). "Influences on infant speech processing: Toward a new synthesis," Annu. Rev. Psychol. **50**, 509–535.
Yeni-Komshian, G., Flege, J. E., and Liu, S. (**2000**). "Pronunciation proficiency in the first and second languages of Korean-English bilinguals," Biling: Lang. Cog. **3**, 131–149.