Perception and production of English vowels by Mandarin

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 cross-over pattern depends on the time point(s) selected for the study. In the beginning of L2 immersion, older learners may

to, and discrepancies with, the native segmental constellations that are in the closest proximity to them in native phonological space." (Best, 1995, p. 193)

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

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 -3678J99.il

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) /deˈp/," "dep (debt)/dɛp/," "dap (dash) /dæp/," "dop (dock) /dɑp/," "dup (duck) /dʌp/," "dawp (dawn) /dɔp/," "dope (doze) /dop/," "doop (food) /dup/)," "dUp (could) /dʊ̈p/," "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)

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 (

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 (24.63%) tokens elicited agreement by four listeners. No judge showed obvious divergence from the group. The agreement rate varied among the vowels, ranging from 94% for /u/, to 41.67% for /Λ/ by at least four listeners. This indicates that disagreements among the listeners were more likely due to the ambiguity of the productions rather than to listener factors. Taking these findings into account, data across all listeners were pooled together for analyses.

The production data from 42 participants in China, 50 recent arrivals, and 76 past arrivals were analyzed for both accuracy and error patterns. For accuracy analyses, all responses were scored as either correct or incorrect. When the intended vowel by the speaker and the chosen vowel by the listener matched, the response was correct. For each speaker, a percent correct score for a vowel was the proportion of correct responses out of five tokens. The total percent correct for all eight vowels was the average of the percent correct scores for the eight vowels.

1. Accuracy across groups and vowels

In this part of the analysis, performance accuracy, indicated by percent correct scores for all vowels and for each vowel were compared across the three participant groups. There was a wide range of accuracy levels across the different vowels (Table III). A mixed two-way 8 (vowels) $\times$ 3 (groups) ANOVA analysis revealed a main effect of group, $F(2,165)=14.36$ ($\eta^2=0.15$), a main effect of vowel, $F(6,1155)=37.76$ (ES = 0.19; with the Greenhouse–Geisser correction of degrees of freedom), and an interaction between the group and vowel, $F(11,1155)=4.31$ ($\eta^2=0.05$; with the Greenhouse–Geisser correction of degrees of freedom) (all $p<0.001$).

The group effect reflects the finding that participants in China had a lower overall accuracy than both the recent and past arrivals. To further examine the group effect for each vowel, separate one-way ANOVA was performed for the individual vowel accuracy scores. There were significant group differences for /eɪ/ [$F(2,165)=28.82$, $p<0.001$], for /ɑ/ [$F(2,165)=9.0$, $p<0.001$], and for /ɪ/ [$F(2,165)=6.74$, $p<0.01$

$=0.34$, $p<0.05$] 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 /eɪ/ [$r=-0.33$, $p<0.01$] 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, /eɪ/ showed a more diffuse confusion pattern, heard as /i/, /ɪ/, or even /aɪ/. For both /ɪ/and /eɪ/, the immigrant groups showed considerable improvement in production accuracy.

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

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). Notably, the findings of the current study and that of Snow and Hoefnagel-Höhle were obtained from different language populations (Mandarin-English versus English-Dutch), with different time sampling methods (cross-sectional versus longitudinal), and different linguistic foci (vowel perception and production in nonsense disyllables versus real word perception and production). These findings further strengthen the view that older learners (or later arrivals in the immigration setting) initially have an advantage over younger learners (or early arrivals in the immigration setting), but this advantage disappears and then becomes reversed over the course of L2 immersion.

In light of these findings, theories that address age-related differences in phonological learning must explain not only the long-term younger learner advantage (as is the traditional focus), but also the short-term older-learner advantage and the processes of change involved. Although all three theoretical accounts predict and explain the long-term younger learner advantage, they are not similarly powerful in explaining the age-related differences exhibited prior to a long-term time point.

The Critical/Sensitive period hypothesis faces a challenge to explain why the genetically preprogrammed advantage of younger learners takes time to exert its effect. In light of the current findings, the theory should at least specify that, whatever the advantage younger learners have in phonological learn-33.4(hypo)298.3(ht-298.3(do.9(ev.3(9.9(98.3(hnch18.3(ht-29F2tetht-us-n)308.8l-n)nbrom)351.uhypotiveba-nt.)TJeba-nt.)TJeba51.uh

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

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

Oyama, S. (**1 ,**). "A sensitive period for the acquisition of a nonnative phonological system," J. Psycholing. Res. **5**, 261–283.
Patkowski, M. S. (**1 0**). "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. (**1 52**). "Control method used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Peterson, G. E., and Lehiste, I. (**1,0**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **32**, 693–703.
Politzer, R. L., and Weiss, L. (**1,**). "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. (**1 2**). "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. (**1 3**). "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. (**1**). "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. (**1**). "Perceptual assimilation of American English vowels by Japanese listeners," J. Phonetics **2**, 311–344.
Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1 ,**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. , 1086–1100.
Tahta, S., Wood, M., and Loewenthal, K. (**1 1**). "Age changes in the ability to replicate foreign pronunciation and intonation," Lang Speech **24**, 363–372.
Tahta, S., Wood, M., and Loewenthal, K. (**1 1**)