SSLA, **20**, 213–243. Printed in the United States of America.

CHANGES OVER TIME IN GLOBAL FOREIGN ACCENT AND LIQUID IDENTIFIABILITY AND ACCURACY

Timothy J. Riney and James E. Flege

International Christian University, Tokyo, and
University of Alabama at Birmingham

This study assessed global foreign accent in sentences and the production of two English consonants, /ɹ/ and /l/, by 11 Japanese college students during their freshman and senior years (T1, T2). In Experiment 1, native English-speaking listeners rated five sentences spoken by the Japanese speakers and five native English control speakers. Experiments 2 and 3 examined 25 word onsets containing /ɹ/ and /l/. Auditory evaluations by native English-speaking listeners were used to determine: (a) to what extent the consonants produced could be identified as intended at T1 and T2; and (b) whether /ɹ/ and /l/ were produced more accurately at T2 than at T1. The results provided little support for a markedness hypothesis based on statistical frequencies and mixed support for a hypothesis based on perception studies. Some speakers made significant improvement, however, in both global foreign accent and liquid identifiability and accuracy.

The acquisition of a second-language (L2) sound system poses many challenges for adults. They must somehow acquire the ability to produce new segments, syllable types, phonotactic sequences, and patterns of stress, pitch, and intonation, and integrate all of these according to new phonological rules of the L2. In attempting to understand how this happens, a common research strategy is

This project was assisted by a number of people who helped in different ways. We would like to make special mention of the following: Janet Anderson-Hsieh, Ken Enochs, Susan Guion, Yuri Harai, Robert Hilke, John Maher, Peter McCagg, Diane Meador, Masako Nakamura, David Pickles, Mari Takada, Randy Thrasher, Yumiko Yaji, and MotokoYoshioka. This research was supported in part by grant DC66257 from the National Institute for Deafness and Other Communication Disorders. Address correspondence to Timothy J. Riney, International Christian University (LD), 3–10-2 Osawa, Mitaka, Tokyo 181, Japan; e-mail: riney@icu.ac.jp.

© 1998 Cambridge University Press 0272-2631/98 $9.50

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

to isolate and measure one variable at a time (typically a feature, segment, syllable type, or rule) and then relate what has been found about that one variable to the overall development of a much larger and complex dynamic system.

This study examined both large and small variables at two times separated by a period of 42 months. Experiment 1 examined global foreign accent (e.g., Munro & Derwing, 1995) in sentences, whereas Experiments 2 and 3 investigated the intelligibility and accuracy of two segments, /ɹ/ and /l/, produced as singletons and as members of word-initial consonant clusters. As used here, the term "foreign accent" refers to both global accent and to foreign accent at a segmental level. Both involve the degree to which an L2 speaker's productions are perceived to differ from those of a native speaker. Foreign accent has been found to be related to a number of variables, perhaps the most influential of which is age of exposure to or arrival in an environment of L2 input (Flege, Munro, & Mackay, 1995), followed by first language (Purcell & Suter, 1980).

The primary purpose of this study was to examine changes over time in foreign accent. An important secondary purpose, however, was to explore the extent to which L2 theory and theories of perception and production can account for observed changes. This study considers several general questions that remain largely unexplored: (a) how does L2 global accent vary over time; (b) how do the intelligibility and accuracy of phonetic segments vary over time; and (c) is there a relationship between the two? We examined the production of isolated words and sentences spoken by 11 native Japanese (NJ) speakers and 5 native English (NE) speakers at two times (T1 and T2). Experiment 1 examined 5 of 15 available sentences; Experiments 2 and 3 used 25 of 84 words from a word list that was read.

We are aware of no longitudinal studies that have examined both segmental articulation and global foreign accent. Important aspects of the present study were its control group; its use of both trained and untrained groups of listeners to evaluate non-native speech; its use of a procedure that blinded listeners to the identity of the speakers; its examination of the target sounds, /ɹ/ and /l/, as both singletons and members of clusters; and its testing of two hypotheses, one based on markedness and the other on perceptual studies.

EXPERIMENT 1

The purpose of Experiment 1 was to determine if global foreign accent for the native Japanese (NJ) speakers was better at T2 than at T1. In a cross-sectional study, Flege and Fletcher (1992) found that Spanish late learners who had lived in the USA for an average of 0.7 years had significantly stronger foreign accents than a similar group who had lived in the United States for an average of 14.3 years. However, in an earlier cross-sectional study, Flege (1988) found no difference in foreign accent between two groups of Chinese late learners who had lived in the United States an average of 1.1 and 5.1 years, respectively. These findings suggest that, in the naturalistic acquisition of an L2, foreign

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

accent may take years to improve measurably. Thus it was uncertain if our NJ speakers, who had spent most of their time between T1 (freshman year) and T2 (senior year) in Japan, would differ in foreign accent.

Method

Speakers. The 11 NJ speakers were students at the International Christian University (ICU) in Tokyo. About 5–10% of the students at ICU are non-Japanese, principally American, and both English and Japanese are used as languages of instruction. The amount of English input that the academic and social environment at ICU provides any NJ speaker naturally varies across courses, instructors, terms, friendships, courtships, and conversation topics. For any given academic year, however, the average NJ at ICU clearly receives more naturalistic English input than does the average NJ at most other Japanese universities and less English input than does the average NJ living abroad in the United States or United Kingdom. All 11 NJ (8 female, 3 male) who volunteered for this project in response to announcements were accepted; none were paid. At T1 (June, 1992), the 11 NJ speakers were college freshmen aged 18–20 years. At T2 (fall, 1995), they were all college seniors.

These speakers completed questionnaires at both T1 and T2. As shown in Table 1, the 11 NJ speakers had similar TOEFL scores (freshman year, 437–497; sophomore year, 490–567) but differed in other respects. Prior to T1, one had attended local schools in the Philippines for a year; six others had been abroad at least for a few weeks at a time. Seven reported that they had gone to typical Japanese public schools, one to a private school, one to a foreign-language high school, and two to a schools designed for returnees (i.e., for those Japanese children returning from non-Japanese schools abroad or for those whose parents, for whatever reason, preferred their children to be in such a school). Between T1 and T2, four speakers spent an academic year at universities abroad (two in California, one in Holland, and one in Mexico). Four others made one or more short trips abroad. The remaining three speakers did not leave Japan. At T2, nine speakers stated goals that involved studying in graduate school outside of Japan or working in some international setting; two indicated they had no plans to go overseas.[1]

The five NE speakers (3 female, 2 male) who comprised the paid control group were monolingual speakers of English at the time that they graduated from high school. All had lived at least the first 17 years of their lives in California. The NE speakers were not informed beforehand of the criteria for participation in this project. At the time of the data collection (May, 1996), all were year-abroad undergraduates from universities in California studying Japanese language and culture at ICU. All of the NE speakers self-reported normal hearing and indicated that they had suffered from no hearing impairment.

Speech samples from the NE control group, who were spending one academic year at ICU, could not be collected at intervals separated by 42 months. The NE speakers' T1 and T2 data collections were instead done at a two-week

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

Table 1. Characteristics of the 11 NJ speakers

(1) ID	(2) Sex	(3) TOEFL	(4) GPA	(5) High School	(6) Travel Abroad Before T1	(7) Travel Abroad Between T1 and T2	(8) College Major	(9) Career Goal
J1	F	477 543	2.97	Public school, tutor 1.5 h/wk	China 2 wk (10) USA 2 wk (16)	Ireland 4 wk	International studies	International telecommunications
J2	F	467 543	3.29	Public school	UK 4 wk (12)	UC Davis 4 wk Europe 6 wk Buffalo, NY 4 wk UC Davis 1 yr	American studies	Graduate school in the USA
J3	F	467 490	2.61	Foreign language high school, studied Spanish, Chinese, and Korean	USA 4 wk (13) Guam 1 wk (5)	Turkey 4 wk	Philosophy	Theatre
J4	F	477 560	3.41	Returnee high school (ICU)	Philippines 1 yr (10) Australia 3 wk	UC San Diego 1 yr New York 8 wk Mexico 2 wk	Sociology	Graduate school in the USA
J5	F	463 n/a	3.32	Public school	None	USA 3 wk	International law	Graduate school in the USA, work for UN
J6	M	437 513	3.51	Returnee high school (ICU)	None	None	Music	Graduate school in Japan or abroad
J7	F	440 533	3.09	Public school	None	Turkey 6 wk	Religion	Work for Japanese firm in Japan
J8	F	447 n/a	2.82	Public school	USA 4 wk	Mexico 1 yr, studied Spanish	French, Spanish	Graduate school abroad
J9	M	497 540	3.70	Private school	USA 10 wk Private high school, studied English since 1st grade	None	Religion	Attorney, International law
J10	F	477 567	2.95	Pubic school	USA 7 wk (19) Singapore 1 wk (14)	Canada & USA 9 wk Holland 9 mo	International economics	International business
J11	M	490 527	2.95	Public school, studied English 10 yr, tutor 1 hr/wk for 4 yrs	None	None	Japanese literature	Stay in Japan

Note. (1) speaker identification label (ID); (2) sex; (3) TOEFL score freshman year (top score), range: 437–497, mean: 467; sophomore year (bottom score), range: 490–567, mean: 535; (4) cumulative grade point average at graduation; (5) type of high school attended ("public school" means that the formal study of English began in 7th grade); (6) travel outside of Japan prior to T1, followed by age at the time in parentheses; (7) travel outside of Japan between T1 and T2 (UC = University of California); (8) college major completed; and (9) career goal at T2. The students' university registrar provided birth dates and items (3) and (4); items (5), (6), (7), (8), and (9) were self-reported; n/a = not available.

interval. It was assumed that the productions of NE speakers in their early 20s taken at an interval separated by two weeks would not differ significantly from one separated by an interval of 42 months. Except for the dates of T1 and T2, the NE and the NJ were recorded under identical conditions.

Speech Materials. The speech samples for the 16 speakers at T1 and T2 were recorded using a Sony TC-1290 monaural tape recorder in a soundproof room at ICU. The first author (TJR) made 28 of the recordings; an American instructor in the ICU English Language Program made the other four. For Experiment 1, the following five sentences were chosen on the basis of recording quality:

1. *A large group of students graduates each spring.*
2. *I heard that splendid speech you made last night.*
3. *They answered correctly and the instructor thanked them.*
4. *I request that all books be removed from the desks.*
5. *Someone's trying to turn my friends against me.*

The 160 sentences (5 sentences $\times$ 16 speakers $\times$ 2 times) were digitized at 22.05 kHz with 16-bit resolution and then normalized for peak intensity.

Procedures. The five NE listeners who rated the five sentences for overall degree of foreign accent reported that they had grown up in the following areas: Wisconsin, Boston, and North Carolina (Listener 1); Los Angeles (Listener 2); Connecticut (Listener 3); Atlanta and Tallahassee (Listener 4); and Birmingham (Listener 5). All five passed a pure-tone hearing screening.

Each listener heard a different randomization of the stimuli in a quiet room located on the campus of the University of Alabama at Birmingham. The sentences were presented via loudspeakers using a notebook computer. The digitized sentences were randomly presented four times each in a separate block lasting 10–15 minutes. The NE listeners heard the five sentences in counterbalanced order. They were told to rate each sentence by clicking a numbered button ranging from "1," for sentences judged to have a "strong foreign accent," to "9," for sentences with "no foreign accent." They were told to use intermediate numbers for sentences with intermediate degrees of accent. The listeners were urged to use the whole scale, to rate only pronunciation, and to ignore everything else (including any perceived lexical errors or grammatical errors, recording noise, or background noise). The final three judgments of each sentence by each listener were averaged. Then the average ratings of the five listeners were averaged to obtain a robust estimate of each speaker's accent.

Results and Discussion

Figure 1 shows the mean foreign accent ratings obtained from the 16 speakers at T1 and T2. (Recall that T1 and T2 were separated by 42 months for the NJ

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

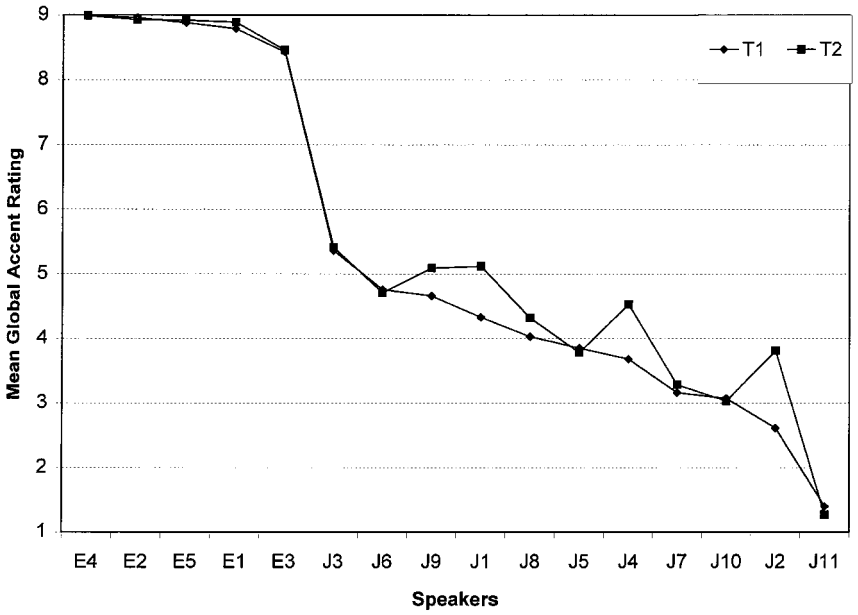


Figure 1. Speakers ranked by highest mean global accent rating at Time 1 in Experiment 1. (1 = most native English, 9 = most strongly accented English; J = Japanese, E = English).

speakers but just 0.5 month for the NE speakers.) The mean ratings shown here have been averaged over the five sentences. As expected (e.g., Flege & Fletcher, 1992), sentences spoken by the five NE speakers (E1–E5) received the highest ratings. Their average rating was 8.8. The sentences spoken by the NJ speakers received an average rating of 3.7 at T1 (range: 1.4–5.4) and 4.0 at T2 (range: 1.3–5.4).

The ratings at T2 were higher than those at T1 for certain NJ speakers. A series of F-tests were carried out to determine if the NJ speakers' ratings at T1 and T2 differed significantly. The ratings obtained for speakers J1, J2, and J4 were significantly higher at T2 than T1, indicating an overall better pronunciation of English ($p < .01$). Two other speakers, J8 and J9, showed marginal improvement ($p = .09$ and $p = .10$). The remaining speakers showed no change between T1 and T2. Although all of the significant movement in accent was toward a greater degree of native accent, no NJ got close to the NE speakers, even at T2.

More than any other factor considered (including short trips abroad, TOEFL scores, grade point average, college major, or career goal), the amount of time spent abroad in an English-speaking country appears to be the factor most closely linked to whether accent improves between T1 and T2. The two NJ speakers whose ratings improved the most over time, J2 and J4, were the two speakers who had spent the most time between T1 and T2 immersed in an English-speaking environment. (See Table 1.) Both J2 and J4 spent an entire

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

academic year in California. The third speaker who showed significant improvement, J1, had spent four weeks in Ireland between T1 and T2, had majored in international studies, and had wanted to pursue a career in "international telecommunications." Speaker J11, whose accent rating was the lowest at both T1 (1.40) and T2 (1.27), was the only NJ speaker to have never traveled outside of Japan, to have had no plans to go outside of Japan, and to have majored in a field of study that was distinctively Japanese (Japanese literature).

We will return to Experiment 1 in the discussion of the relationship between global accent and the intelligibility and accuracy of /ɹ/ and /l/ that follows Experiments 2 and 3. In light of the findings of Experiment 1, one general question of interest for Experiments 2 and 3 is whether the identifiability and accuracy of individuals' /ɹ/ and /l/ productions will reflect their global accent ratings.

EXPERIMENT 2

The primary question addressed by Experiment 2 was whether the NJ speakers' /ɹ/s and /l/s would be identified correctly more often at T2 than at T1. A consideration of existing literature provided the rationale for two hypotheses. Before presenting these hypotheses, however, the phonetic structure of English and Japanese liquids needs to be reviewed.

American English /ɹ/ and /l/ have long been recognized as difficult for NJ speakers of English (e.g., Goto, 1971). In prevocalic position, English /l/ is a (usually) voiced alveolar lateral approximant. Prevocalic English /ɹ/ is usually described as a voiced, alveolar (sometimes retroflexed) approximant (Ladefoged, 1993). These are the forms of American English /ɹ/ and /l/ typically presented in the pedagogical textbooks in Japan, including the one (Brown, 1970) used at ICU during the time that the 11 NJ speakers were there. These are also the forms of /ɹ/ and /l/ judged to characterize the speech of the American English control group from California.

Japanese has no consonant similar to either English liquid. The Japanese liquid is an apico-alveolar flap or tap. (It will be referred to here as a "tap" in order to distinguish it from the American English "flap.") The Japanese liquid is transliterated in English as "r". The Japanese sound occurs only prevocalically and intervocalically, and it contrasts with /d/ but not with any lateral or retroflex phones (Price, 1981). The flap that occurs in American English, however, is widely considered to be an allophonic variant not of /ɹ/, but of /t/ and /d/ (as in *kitty* and *kiddy*). Thus, the phonological roles and the distributions of the Japanese tap and the American English flap are quite different.

If the Japanese tap has a phonetic counterpart in American English, it is the intervocalic flap. Vance (1987, p. 27) described the Japanese flap as "essentially the same sound" as the /ɹ/ in Spanish, Russian, and the intervocalic flap in American English *city*. Yamada and Tohkura (1992, p. 377) classified the Japanese sound as "a stop, or flap, depending on its vowel context." Laver (1994, pp. 224–227) categorized flaps and taps as "flapped stops" and "tapped stops." In her comparative study of flaps in intervocalic position in American English and

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

in Japanese (the only position in which taps or flaps freely occur in both languages), Price (1981) determined that the Japanese tap and the American English flap, while operating in different phonological systems, are similar articulatorily and acoustically. The study by Price (1981) was based on transcriptions, spectral analysis, waveform patterns, and perceptual evidence involving judgments by a trained phonetician and native American English speakers making judgments on Japanese productions and vice versa. Price found that the "allophonic variability" of the Japanese tap was unclear (p. 8).

In American English, /l/ is a lateral and /ɹ/ is a rhotic; the Japanese liquid is a flap or tap (i.e., a type of stop; Ladefoged & Maddieson, 1996). Thus, at a phonetic level of analysis, we are dealing with three different sounds. This will be important for one of the hypotheses to be presented below.

Hypotheses

One might generate two predictions concerning NJ acquisition of English singletons /ɹ/ and /l/. One prediction is based on previous studies of perception; the other is based on markedness. The Marked Differential Hypothesis (MDH) and the Interlanguage Structural Conformity Hypothesis (ISCH) were both formulated by Eckman (1977, 1991). According to the MDH, difficulty in SLA is predictable based on "the areas of difference between the NL and TL" and "the relative markedness of these differences" (1991, p. 32). Markedness is determined on the basis of implicational universals in Eckman's theory. His claim is that those TL areas that are both different and more marked than the corresponding NL structures will be difficult, and that the level of difficulty will correspond to the degree of markedness (1991, p. 32). Eckman based the MDH on implicational universals. For example, if a language has final voiced obstruents, then it will also have final voiceless obstruents (Greenberg, 1978). Such a measure seems to provide an index of articulatory complexity and L2 difficulty. One important advantage of implicational universals is that they can be stated clearly; another is that the case for counterevidence is clear. With regard to the NJ acquisition of /ɹ/ and /l/, however (as is the case with singletons in general), there are no implicational universals that apply.

The current study will instead use universals based on statistical frequencies, expressed as percentages representing the frequencies with which phoneme types appear across a representative sample of languages. The more frequently a phoneme type occurs across languages (represented as a higher percentage), the less marked it is. The less frequently a phoneme type occurs, the more marked it is. When a prediction is needed to assess the relative difficulty of two TL phonemes, one might predict that the greater the difference in the statistical frequencies of two L2 phonemes not found in the L1, the greater will be the difference in the learnability of the two phonemes, with the more frequent phoneme being favored in acquisition over the less frequent phoneme. ("Favored" is used here in a general sense that encompasses production and perception, as well as identifiability, accuracy, and rate of acquisition or improvement.) Like the MDH, this hypothesis based on statistical frequencies could be "falsified

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

by facts showing that a learner evidences more difficulty in a less marked structure, and less difficulty in a more marked structure" (Eckman 1991, p. 33).

Statistical frequencies relevant to the Japanese and American English liquids can be found in Maddieson (1984). These frequencies are based on the phonological inventories of the UCLA Phonological Segment Inventory Database (UPSID) which "have been chosen to approximate a properly constructed quota sample on a genetic basis of the world's extant languages" (p. 5). In this database, 81.4% of the languages surveyed have one or more lateral segments; 76.0% have one or more "r-sounds." There are more languages with greater numbers of laterals and "about 57% of the liquids reported are laterals" (p. 73). Of the remaining 43% nonlateral liquids, most are trills (47.1%) and taps and flaps (38.3%), and relatively few are continuants (13.5%), a category that includes the American English /ɹ/. Furthermore, of the total number of the laterals counted in the survey, "plain voiced approximant laterals are by far the most common type of lateral" (p. 74) and comprise 74.7% of the total laterals. This is the type of lateral found in American English. The type of /ɹ/ found in American English represents only 5.6% ($.43 \times .135$) of the total number of liquids in UPSID. American /ɹ/ is more marked than the American /l/, the latter type comprising 42.6% ($.747 \times .57$) of all liquids and much more common than /ɹ/s in the phonological inventories of human languages. Thus, following Eckman, one could state a frequency-based markedness hypothesis:

> Japanese has neither an /l/ nor an /ɹ/ like those in American English; thus, to the degree that /l/ is less marked, and /ɹ/ is more marked, then to that degree /l/ will be produced better than /ɹ/ by native speakers of Japanese who are acquiring English.

The MDH was concerned with "using universal generalizations to make predictions about order of acquisition, relative difficulty, or prevalence of errors" (Eckman, 1991, p. 24) in SLA. The ISCH, the hypothesis that Eckman (1991) formulated 14 years later, was more concerned with testing whether "the universal generalizations that hold for the primary languages hold also for interlanguages." One difference between the two hypotheses is that for the ISCH "no account whatever is taken of the differences between the NL and the TL" (Eckman, 1991, p. 33). The ISCH, like the MDH, relies on implicational universals that apply to some but not all L1-L2 pairings. Neither can be applied to the NJ speakers' acquisition of /ɹ/ and /l/. An ISCH-type hypothesis (which ignores L1) based on statistical frequencies alone might read as follows:

> To the degree that /l/ is less marked and /ɹ/ is more marked, then to that degree /l/ will be produced better than /ɹ/ by native speakers of Japanese who are acquiring English.

The application and the falsification of the MDH and the ISCH based on implicational universals is rather straightforward. This is not the case for hypotheses based on statistical frequencies because a continuum of percentages offers no clear cut-off between marked versus unmarked phoneme types.

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

Nonetheless, a frequency-based hypothesis has an advantage in that it can at least be applied in certain cases in which implicational universals cannot be applied. The NJ acquisition of American English /ɹ/ and /l/ is one such case.

The hypothesis based on statistical frequencies will be referred to below as the "Frequencies Hypothesis." Because neither /ɹ/ nor /l/ phonemes occur in Japanese, both the MDH-version (involving L1) and the ISCH-version (ignoring L1) of this hypothesis make the same prediction—that /l/ will be produced better than /ɹ/. Thus, the two versions (one involving the L1 and the other not) need not be distinguished here.

The second hypothesis to be formulated is the Perception-Production Hypothesis. Several empirical studies of NJ speakers' perception of American English /ɹ/ and /l/ have contributed to the formulation of this hypothesis. Previous studies have found that the NJ speakers' perception and production of English /ɹ/ and /l/ involve more than just the misidentification of /ɹ/ for /l/ and /l/ for /ɹ/. Price (1981) found that American English listeners were more likely to categorize a Japanese liquid, which is a tap, as a /t/ or a /d/ rather than an /ɹ/ or /l/. Sekiyama and Tohkura (1993, p. 436) presented Japanese [ɾa] syllables to American English subjects who identified the Japanese liquid as "la" (51% of the time), "gra" (17%), "dla" (10%), "ra" (10%), trilled Spanish "rra" (10%), and "wla" (2%). The same Japanese [ɾa] syllables were identified by Japanese listeners as Japanese "ra" (100%). Conversely, the English [ɹa]'s were identified by American English listeners, as English "ra" (100%) but by Japanese listeners as English "ra" only 67% of the time, and in other instances as Japanese "ra" (13%), "wa" (13%), and "ga" (7%). Thus, the perceived interlingual relationship between Japanese /ɾ/ and American English /ɹ/ and /l/ is more complex than just a phonemic level of analysis might suggest.

Four studies have reported perceptual data at a single point in time for Japanese subjects who had not received any special training. Mochizuki (1981), Sheldon and Strange (1982), Takagi (1993), and, in their pretest, Bradlow et al. (1997)[2] all found that American English /ɹ/ was identified correctly by NJ speakers more often than /l/ in the #_V environment (one of two environments to be examined in Experiments 2 and 3 below). This perceptual finding is the opposite of the prediction of the Frequencies Hypothesis above for production. A second finding of the four perceptual studies above was that in the #C_V cluster environment (the other environment to be investigated in this study), English /l/ was identified correctly more often than English /ɹ/.

In addition to these findings, two theories contributed to the formulation of the second hypothesis. Both posit that there is a direct linkage between perception and production. According to the motor theory (e.g., Liberman et al., 1967; Liberman & Mattingly, 1985, 1989), sounds are perceived in terms of the articulatory gestures that are used to produce them. This theory posits the existence of a specialized phonetic module and a single shared representation for speech perception. The direct-realist approach (e.g., Fowler, 1986; Best, 1995) posits that listeners perceive articulatory gestures of speakers in terms of the structure these gestures impart to the acoustic medium. This theory assumes no phonetic module as mediator. (See also Bradlow et al., 1997.)

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

Given the empirical findings reported above and the two theories that link perception and production, it seems reasonable to investigate two questions: (a) at a single point in time, to what extent are sounds that are more accurately perceived also more accurately produced; and (b) over a period of time, will relatively more improvement in production be made for the sounds (e.g., initial singleton /ɹ/) that are perceived better, and will relatively less improvement be made for the sounds (e.g., initial singleton /l/) that are perceived less well? Studies cited above have investigated question (a). We are aware of no longitudinal studies of NJ acquisition of English /ɹ/ and /l/ in naturalistic settings that have investigated question (b).[3]

In addressing this second question, we formulated a second hypothesis that had two parts. First, because NJ speakers have been found to perceive initial /ɹ/ singleton better than initial /l/ singleton, it is predicted that they will be able produce /ɹ/ singleton more accurately than /l/ singleton at Time 1, and over time (i.e., from Time 1 to Time 2), they will show more improvement in their production of /ɹ/ singleton than of /l/ singleton. Second, because NJ speakers have been found to perceive /l/ better than /ɹ/ in word-initial clusters, it is predicted that they will be able to produce /l/ more accurately than /ɹ/ in word-initial clusters at Time 1 and over time, and that they will show more improvement for /l/ than for /ɹ/ in this same environment.

The two hypotheses just presented, which were derived from empirical findings, generate three predictions concerning our Japanese speakers productions of /ɹ/ and /l/:

Hypothesis 1 (*Frequencies Hypothesis*). Between T1 and T2, /l/ singletons will show more improvement in identifiability than /ɹ/ singletons.

Hypothesis 2 (*Perception-Production Hypothesis*). This hypothesis generates two predictions, one for singletons, and one for clusters: (a) Between T1 and T2, /ɹ/ singletons will show more improvement in identifiability than /l/ singletons, and (b) between T1 and T2, /l/ in clusters will show more improvement in identifiability than /ɹ/ in clusters.

Method

Speakers and Speech Materials. The speakers for Experiment 2 were the same as for Experiment 1. Experiment 2 examined demisyllables edited from the following 25 words of a list of 84 words that were read in isolation: *rag, run, request, removed, risk, last, lunched, lend, large, like, proof, train, trunk, cream, group, graduates, friends, from, play, sleep, spring, strange, screen, splinter*, and *splendid*.[4] This set of words included five words beginning with a singleton /ɹ/, five with a singleton /l/, ten with the liquid as the second member of the cluster (7 with /ɹ/, 3 with /l/), and five with the liquid as the third member of the cluster (3 with /ɹ/, 2 with /l/. Note that the cluster phonotactics and the following vowel was not controlled.

The 800 words (25 words × 16 speakers × 2 times) were digitized at 22.05 kHz with 16-bit resolution, then normalized for peak intensity. Some of the recordings were noisy. This did not pose a serious problem for the foreign-

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

accent rating experiment (Experiment 1) because listeners can readily hear through noise in long stretches of speech. However, we thought the noise might influence the listeners' identifications of individual /ɹ/ or /l/ segments in this experiment, so an inverse filtering process was used to remove noise from the 800 digitized words. This involved constructing a filter (*Cool Edit*, 1996) based on the noise present in the tape recordings where no speech was present. This process reduced considerably the level of noise in the stimuli, but it sometimes introduced a metallic sound and some artifacts into the speech.

The initial liquids and roughly one-half of the following vowels were edited, producing (C)(C)CV-demisyllables. This editing procedure was used to remove possible deviations in the vowel nuclei or syllable codas that might influence listener judgments of /ɹ/ and /l/ in the syllable onsets. The first author (TJR) used the editing procedures described by Flege, Takagi, and Mann (1995). This involved listening to the whole word until the vowel quality of the syllable in question could be identified, then replaying the word to that point and deleting everything beyond that point. The editing usually resulted in the vowel portion of the wavelength being cut approximately in half. The edited stimuli that remained after editing were ramped off over 20 milliseconds to avoid clicks.

Procedure. The (C)(C)CV-demisyllables were presented binaurally via headphones to three trained listeners in a sound booth. The listeners were native speakers of American English raised in Ohio, Washington, and Texas. One was the second author (JEF), and the others were postdoctoral fellows doing research in L2 phonetics and phonology. All passed a pure-tone hearing screening at octave frequencies between 250 and 4000 Hz. The listeners were told that they would hear only the first part of words whose identities were not revealed. They were asked to identify only the /ɹ/ or /l/ consonant in the demisyllable by clicking one of three boxes on a computer screen. The boxes were marked: "L," "R," and "NEITHER." The listeners were told that "L means that you categorize the segment as English /l/," "R means that you categorize the segment as English /ɹ/," and "NEITHER means that you categorize the segment as neither English /l/ nor English /ɹ/." No feedback was given. The interval between each response and the next demisyllable was 1.0 seconds; a "repeat" button was provided to allow the listeners to replay a stimulus.

The stimuli (demisyllables) of each NJ speaker were presented in a separate randomized block, each with 50 stimuli (25 from T1, 25 from T2) preceded by 10 practice items taken from the end of the block. The 16 speakers were presented in different random orders to each of the three listeners. Each listener judged 800 stimuli (16 blocks $\times$ 50 stimuli), yielding 2400 judgments (800 stimuli $\times$ 3 trained listeners) in all.

Results and Discussion

The mean percent-correct identification scores of /ɹ/ singletons and /ɹ/ in clusters are shown for individual speakers at Time 1 and Time 2 (T1, T2) in Figures

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

2 and 3; corresponding results for /l/ singletons and /l/ in clusters are shown in Figures 4 and 5. The mean values for all four types of data (/ɹ/ singletons, /ɹ/ in clusters, /l/ singletons, and /l/ in clusters) are also presented in Table 2 to permit a ready comparison across data types and speakers.

The 32 (16 speakers × 2 times) mean percent-correct identification scores for /ɹ/ singletons in Figure 2 are each based on a total of 15 judgments (5 words × 3 listeners). The percent-correct scores are generally higher for the NE speakers than for the NJ speakers at Time 1, although some NJ speakers have T1 scores that are as high as those for the NE speakers. For several NJ speakers, T2 scores are considerably higher than the T1 scores; for others the T2 scores are lower than the T1 scores.

The 32 mean percent-correct identification scores for /ɹ/ in clusters in Figure 3 are each based on a total of 33 judgments (11 words × 3 listeners). The percent correct scores of the NJ speakers at T1 differ much less from those of the NE speakers than was the case for /ɹ/ singletons. Again, some NJ speakers have T1 scores that are as high as those of the NE speakers. A few NJ speakers showed slightly higher scores at T2 than at T1, but several had lower scores at T2 than at T1.

The 32 mean percentage identification scores for /l/ singletons in Figure 4 are each based on a total of 15 judgments (5 words × 3 listeners). The T1 scores are generally high, and the T2 scores generally got even higher, and this accounts for much of the difference between the /l/ singleton scores and the /ɹ/ singleton scores. Again, the NE speakers are generally higher than the NJ speakers, but some NJ speakers have T1 scores that are as high as those for the NE speakers.

The 32 mean percent-correct identification scores for /l/ in clusters in Figure 5 are each based on a total of 12 judgments (4 words × 3 listeners). The /l/ in cluster T1 scores begin much lower than do the T1 /l/ singleton scores, but like the /l/ singleton scores, they almost all get higher at T2—a development that distinguishes the two from the /ɹ/ singleton and /ɹ/ in cluster scores. Again, as was seen in the three previous figures, the NE speakers generally have higher scores than the NJ speakers, but some of the NJ speakers have T1 scores that are as high as those for the NE speakers.

Two important decisions were made concerning data analysis. First, the results for two NJ speakers (J3 and J5), with scores at or near 100% correct at T1 and T2 for all four liquid onset types, were excluded. (See Table 1.) This was done because it would not have been possible to observe any improvement for them between T1 and T2, which was the focus of the study. Second, we decided not to compare the NE and NJ speakers directly in ANOVAs because variances for the two groups differed greatly.[5] The results for the NE were therefore analyzed separately, using the same ANOVA design as applied to the NJ.

The eight percent-correct identification scores obtained from each of the five NE subjects were calculated for each speaker: two (T1 and T2) each for the /ɹ/ singletons, /ɹ/ in clusters, /l/ singletons, and /l/ in clusters.[6] The 40 percent-correct scores (5 speakers × 2 times × 4 scores) obtained from the NE

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

Table 2. Percent-correct identification for target /ɹ/ and /l/ in four liquid onset types at T1 and T2 for 16 speakers

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	E1	E2	E3	E4	E5
Onset type																
/ɹ/ as singleton onset																
T1	73.3	60.0	93.3	20.0	100	40.0	20.0	20.0	80.0	20.0	00.0	100	80.0	100	86.7	100
T2	100	13.3	93.3	100	100	60.0	13.3	00.0	60.0	40.0	93.0	100	93.3	100	100	100
/ɹ/ in cluster onset																
T1	100	97.0	93.3	78.8	100	84.8	87.9	45.5	84.8	84.8	51.5	100	97.0	100	93.3	97.0
T2	97.0	24.2	97.0	97.0	97.0	78.8	75.8	45.5	97.0	45.5	51.5	100	100	100	100	97.0
/l/ as singleton onset																
T1	80.0	100	100	80.0	100	86.7	20.0	100	60.0	46.7	86.7	86.7	100	100	100	93.3
T2	100	100	100	93.3	100	86.7	06.7	93.3	100	100	100	100	100	100	100	100
/l/ in cluster onset																
T1	33.3	08.3	100	33.3	100	33.3	16.7	50.0	33.3	00.0	41.7	83.3	91.7	100	100	100
T2	100	100	100	58.3	100	33.3	16.7	75.0	66.7	100	58.3	100	100	100	100	91.7

Note. J = native speaker of Japanese; E = native speaker of American English; T1 = Time 1; T2 = Time 2.

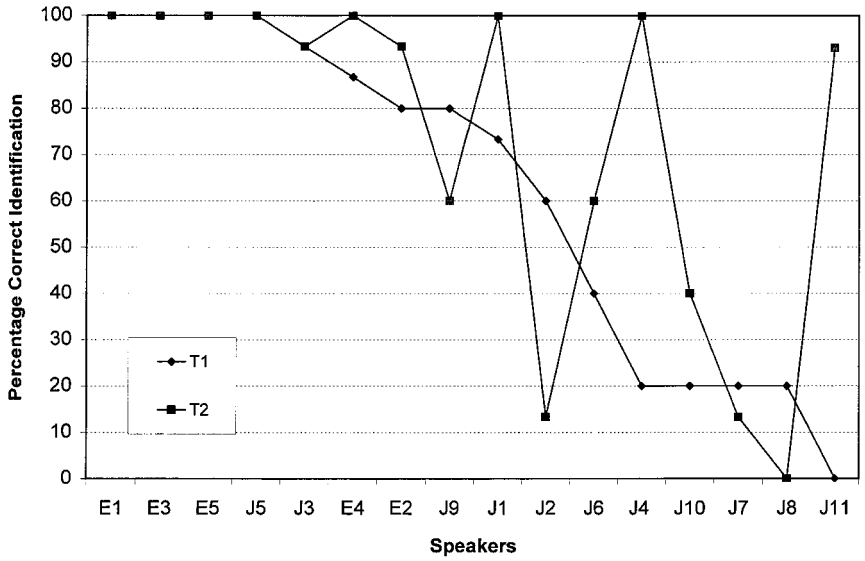


Figure 2. Speakers ranked by percent-correct identification for /ɹ/ singletons at Time 1 in Experiment 2. (J = Japanese, E = English.)

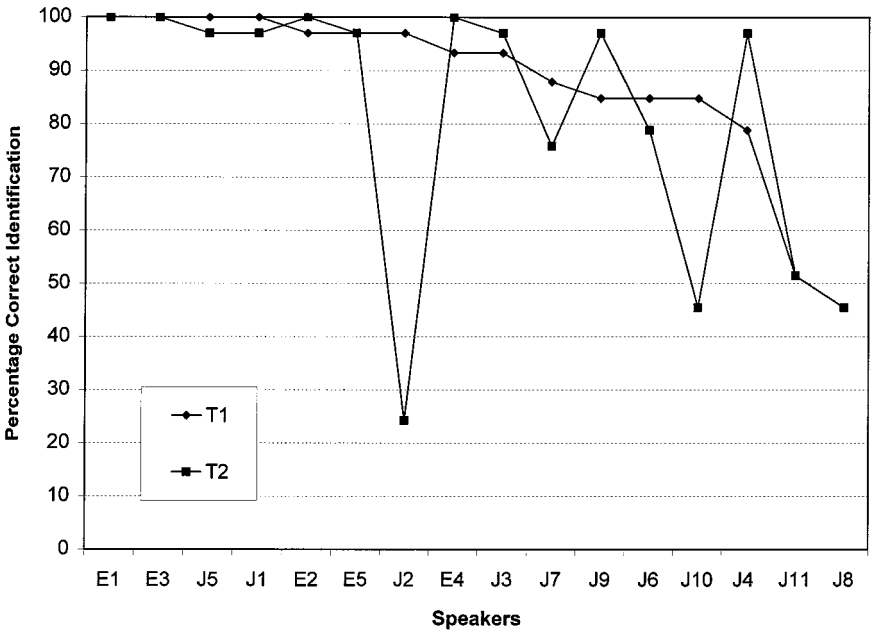


Figure 3. Speakers ranked by percent-correct identification for /ɹ/ in clusters at Time 1 in Experiment 2. (J = Japanese, E = English.)

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

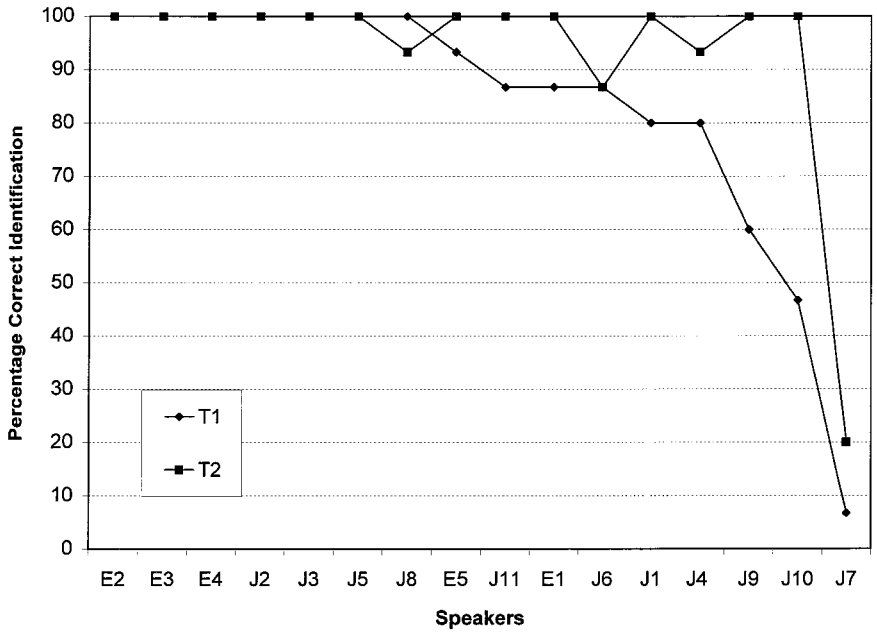


Figure 4. Speakers ranked by percent-correct identification for /l/ singletons at Time 1 in Experiment 2. (J = Japanese, E = English.)

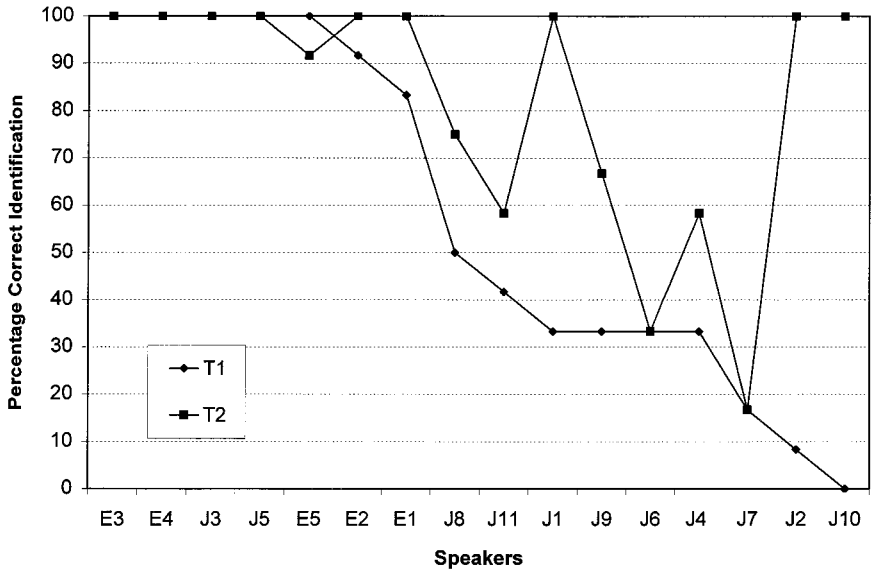


Figure 5. Speakers ranked by percent-correct identification for /l/ in clusters at Time 1 in Experiment 2. (J = Japanese, E = English.)

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

speakers were submitted to a repeated-measures Onset type (singleton vs. cluster) × Consonant (/ɹ/ vs. /l/) by Time (T1 vs. T2) ANOVA. The overall difference between singletons and clusters (97.0% vs. 97.6% correct) was not significant, $F(1, 4) = 0.32$, $p > .10$. Nor was there a significant difference in the overall rate at which the NE speakers' productions of /ɹ/ and /l/ (97.2% vs. 97.3%) were identified correctly, $F(1, 4) = 0.001$, $p > .10$; nor in the rate at which their T1 and T2 productions (95.5% vs. 99.1.%) were identified correctly, $F(1, 4) = 4.99$, $p > .05$. None of the interactions involving the three factors reached significance ($p > .10$).

A similar ANOVA was carried out to examine the eight percent-correct scores obtained for each of the nine NJ speakers. (Recall that two NJ speakers were excluded as was discussed above.) The ANOVA yielded a significant Onset type × Consonant × Time interaction, $F(1, 8) = 9.67$, $p < .05$. The basis for this interaction appears to have been that, of the four sets of scores examined (/ɹ/ singletons, /ɹ/ in clusters, /l/ singletons, /l/ in clusters), only the scores for /l/ in clusters increased significantly from T1 to T2. The simple effect of time was insignificant for /ɹ/ singletons (T1: 37% vs. T2: 53%), $F(1, 8) = 1.11$, $p > .10$; for /ɹ/ in clusters (T1: 79% vs. T2: 68%), $F(1, 8) = 1.49$, $p > .10$; and /l/ singleton scores (T1: 73% vs. T2: 87%), $F(1, 8) = 3.35$, $p > .10$; but it was significant for the /l/ in cluster scores (T1: 28% vs. T2: 68%), $F(1, 8) = 10.2$, $p < .05$.

As a result of the improvement in the /l/ in-cluster scores, the relation between scores for /l/ and /ɹ/ also changed. The /l/ singletons received higher scores than did the /ɹ/ singletons at both T1 [/l/: 73% vs. /ɹ/: 37%; $F(1, 8) = 8.95$, $p < .05$] and T2 [/l/: 87% vs. /ɹ/: 53%; $F(1, 8) = 6.521$, $p < .05$]. However, whereas the /ɹ/ in clusters received higher scores than did the /l/ in clusters at T1 [79% vs. 28%; $F(1, 8) = 23.9$, $p < .001$], the differences between the liquids in clusters [/ɹ/ in clusters: 68% vs. /l/ in clusters: 68%; $F(1, 8) = 0.001$, $p > .10$] was insignificant at T2.

The results just presented allow us to evaluate the hypotheses presented earlier. According to the Frequencies Hypothesis, the percent-correct identification scores for /l/ singletons should improve more than for /ɹ/ singletons between T1 and T2. This hypothesis was not supported. No significant improvement was found for /l/ singletons (or for /ɹ/ singletons). If implicational universals provide an index of articulatory complexity and L2 difficulty, as suggested by Eckman (1991) and others, one might expect statistical frequencies based on the phonetic inventories of languages to also provide such an index. Perhaps the amount of time that the NJ speakers had lived in the TL environment (see Table 1) was insufficient for the hypothesis to be tested. Or, perhaps much of the improvement for /l/ singletons vis-a-vis /ɹ/ singletons had already occurred before T1. The NJ group's mean score for /l/ singleton identification was 73.0% at T1 (and 86.7% at T2); for /ɹ/ singletons it was only 37.0% at T1 (and 53.3% at T2). The /l/ singleton score may have been constrained by a ceiling effect. Even so, the NJ speakers' /l/ singleton scores either stayed the same or got better in all but one case (that of J8, who began at ceiling at T1). Individual scores for /ɹ/ singletons began at a much lower level (37.0%), yet in several

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

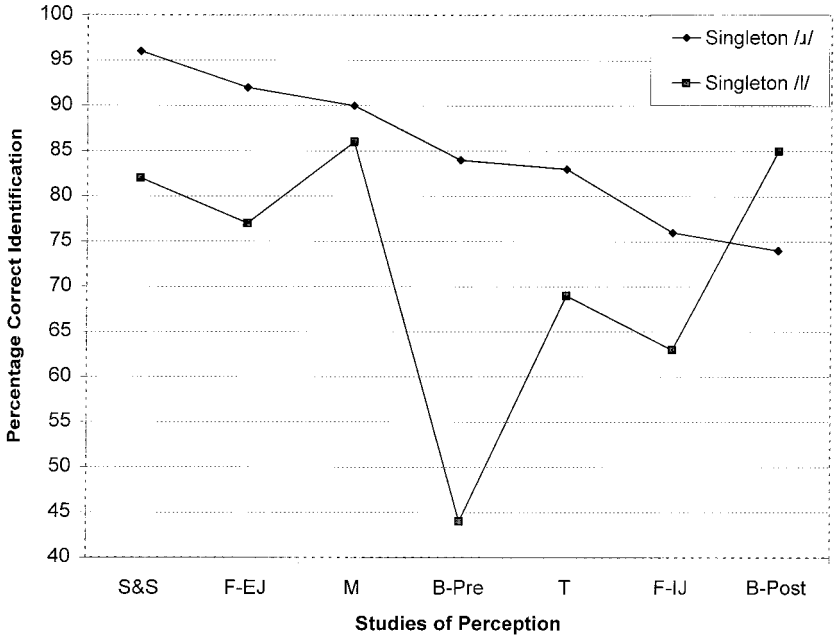


Figure 6. Assessments of Japanese perceptions of English /ɹ/ and /l/ as word-initial singletons. Percentage represents native-English productions correctly identified by Japanese listeners. [B-Pre = pretest; B-Post = posttest of Bradlow et al. (1997). F-EJ = experienced; F-IJ = inexperienced speakers of Flege, Takagi, & Mann (1996). M = Mochizuki (1981). S&S = Sheldon & Strange (1982). T = Takagi (1993).]

cases (J2, J8, J9, and to a lesser extent J7; see Figure 2) got worse. In sum, a less proficient NJ group (i.e., with lower T1 scores for /l/ singletons) might be found to make more improvement over time for /l/ singletons than for /ɹ/ singletons. Thus, the Frequencies Hypothesis merits further exploration.

According to the Perception-Production Hypothesis, /ɹ/ singletons should improve more than /l/ singletons between T1 and T2. This hypothesis was not supported. One possible explanation for this is that the ability to produce and to perceive /ɹ/ and /l/ involve separate competencies that are more independent of one another than that suggested by the motor theory and the direct realist approach.

The second prediction derived from the Perception-Production Hypothesis correctly predicted that between T1 and T2 /l/ in clusters would improve in identifiability more than /ɹ/ in clusters. Because a parallel finding will be made in Experiment 3, for accuracy, a discussion of this finding will be reserved until after Experiment 3.

The current study adds two new measurements (T1 and T2) to a growing number of studies of NJ speakers' perception and production of /ɹ/ and /l/. Figure 6 juxtaposes the findings of five perceptual studies that assessed NJ

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

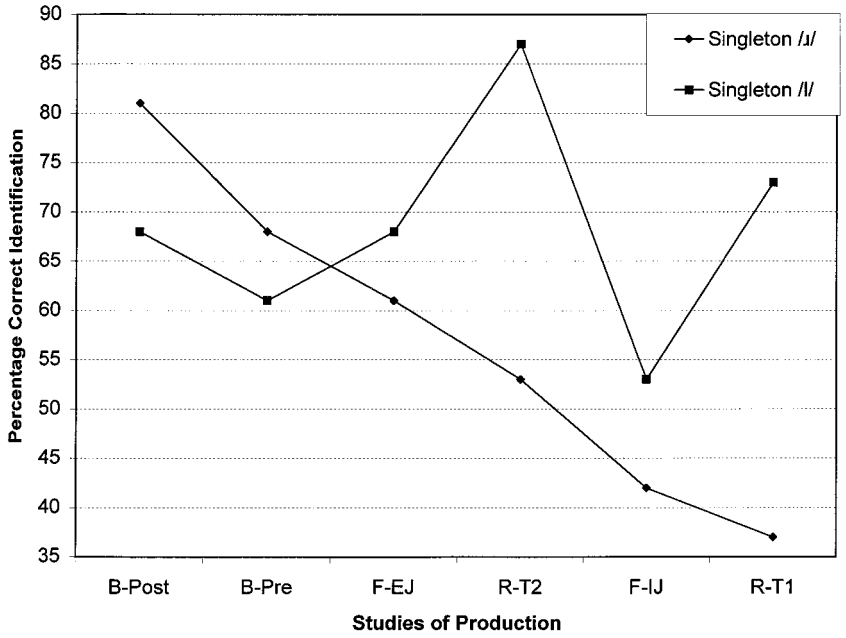


Figure 7. Assessments by native English listeners of Japanese productions of English /ɹ/ and /l/ as word-initial singletons. Percentage represents productions correctly identified. [B-Pre = pretest; B-Post = posttest of Bradlow et al. (1997). F-EJ = experienced and F-IJ = inexperienced speakers of Flege, Takagi, & Mann (1996). R-T1 = Time 1; R-T2 = Time 2 of the current study.]

identification of initial singleton /ɹ/ and /l/. The figure is based on percent-correct identification scores for NJ speakers' perception of /ɹ/ singletons and /l/ singletons in the #_V environment. Three sets of data (abbreviated in the figure as S&S, M, and T) are from the studies of Sheldon and Strange (1982), Mochizuki (1981), and Takagi (1993). Two others (F-IJ and F-EJ) are from Flege, Takagi, and Mann (1996) for inexperienced NJ who had been in the United States for two years and experienced NJ who had been in the United States for an average of 21 years. The remaining two scores (B-Pre and B-Post) are the pretest and posttest results obtained by Bradlow et al. (1997). In six of these seven cases, /ɹ/ was identified more accurately than /l/; the one exception occurred in the posttest of Bradlow et al. (1997), which was the only assessment that was done after explicit training.

Figure 7 compares the percent-correct identification of NJ productions of English /ɹ/ and /l/ from three studies. Two of these (F-IJ and F-EJ) are from a second study by Flege, Takagi, and Mann (1995) that used the same subjects as the first (1996) and found that both inexperienced NJ and experienced NJ produced /l/ more identifiably than /ɹ/. (These last four percentages are estimates based on published graphs and a listener-rater judgment category of "definite.") It should be noted that the native English control group for Flege,

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

Takagi, and Mann (1995), however, obtained 74% for /l/ and 67% for /ɹ/. Thus, a similar pattern for NJ speakers does not necessarily mean that /l/ is more learnable than /ɹ/. In fact, when NJ speakers /ɹ/ and /l/ were examined in a goodness-rating experiment, they were judged to have been produced with equal accuracy. Two other scores (B-Pre and B-Post) are from the pretest and posttest results for production obtained by Bradlow et al. (1997). The last two measures in Figure 7 (R-T1 and R-T2) are the T1 and T2 scores from this current study.

The studies in Figures 6 and 7 involve a range of experimental designs.[7] When viewed together, however, Figure 6 suggests that NJ speakers who have had no explicit training perceive /ɹ/ singletons slightly more accurately than /l/ singletons. Figure 7 suggests that they seem to produce /l/ singletons with greater variation—sometimes as accurately and sometimes more accurately than they produce /ɹ/ singletons.

EXPERIMENT 3

The purpose of Experiment 2 was to determine if the /ɹ/s and /l/s produced at T2 were more identifiable than the /ɹ/s and /l/s produced at T1. It is possible that a speaker's /ɹ/s and /l/s were equally identifiable at T1 and T2 in Experiment 2 but nonetheless became more accurate at T2 than they were at T1. Experiment 3 used a more fine-grained perceptual technique to determine whether any such improvement occurred. The hypotheses for Experiment 3 are identical to those for Experiment 2 except that where Experiment 2 used "identifiability," Experiment 3 used "accuracy."

Method

Experiment 3 used the same speakers and demisyllable stimuli as in Experiment 2. However, whereas Experiment 2 used three trained listeners, Experiment 3 used 10 listeners without training in speech or language research. None of the 10 listeners (6 male, 4 female) had special knowledge about speech or perception. Five had been listeners in Experiment 1. The 10 NE listeners reported that they had grown up in or near the following areas: Seattle (Listener 1); Wisconsin, Boston, and North Carolina (Listener 2); Birmingham (Listener 3); Alabama (Listener 4); New York City (Listener 5); Atlanta and Tallahassee (Listener 6); Los Angeles (Listener 7); Seattle (Listener 8); Connecticut (Listener 9); and Alabama (Listener 10). All passed a pure-tone hearing screening.

In Experiment 3, the demisyllables were presented in pairs (i.e., the two matching T1–T2 demisyllables). The listeners were told which consonants (/ɹ/ or /l/) the demisyllables were supposed to contain. They were told to choose the member of the pair that contained the better exemplar of the intended consonant. The listeners were warned in advance that at times they might not be able to make a confident judgment or even to identify any sound as /ɹ/ or /l/ in either member of the pair. In such instances, they were told to guess. They clicked one of two response boxes simply labeled "1" or "2" on a computer

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

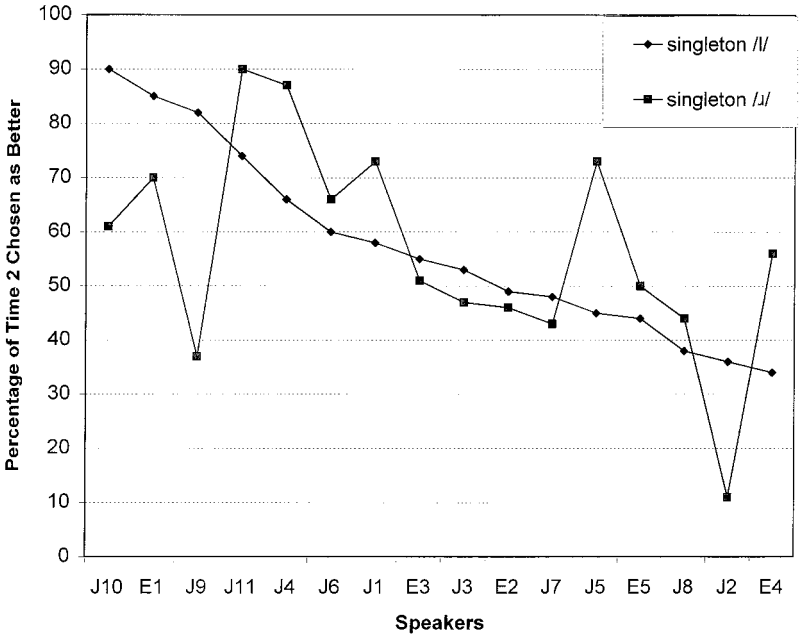


Figure 8. Speaker ranked by percentage of Time 2 tokens chosen as better for /l/ singletons in Experiment 3. (J = Japanese, E = English.)

screen to indicate which demisyllable stimulus (the first one heard, or the second) contained the better /ɹ/ or /l/. The order of the two stimuli (T1, T2) was randomized. The stimuli were presented in two sessions, separated by a short break. Five of the 10 untrained listeners judged the /l/s first, then the /ɹ/s; the 5 others judged the /ɹ/s first, then the /l/s. The /l/ session consisted of 18 paired stimuli (9 in a T1–T2 order, and 9 in a T2–T1 order); the /ɹ/ session consisted of 32 paired stimuli (16 in a T1–T2 order, and 16 in a T2–T1 order). The 16 blocks of speakers were sequenced in different random orders for each of the 10 untrained listeners. Each untrained listener judged 800 paired stimuli (16 blocks of 50 pairs). A total of 8000 (10×800) judgments were obtained.

Results and Discussion

The results for /ɹ/ and /l/ singletons are shown in Figure 8. The results for /ɹ/ and /l/ in clusters are shown in Figure 9. The dependent variables examined here were "preference" scores, computed as the percentage of times T2 tokens were chosen as the "better" member of the pair.[8]

In Figure 8, the preference scores for /ɹ/ singletons and /l/ singletons are each based on a total of 100 judgments [5 paired words $\times$ 10 listeners $\times$ 2 orders]. (Recall that the orders for the stimuli are randomized T1–T2 and T2–T1.) The speakers are arranged in the order of highest T2 preference scores for /l/ singletons, and on this basis one sees, for example, that for /l/ singletons,

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

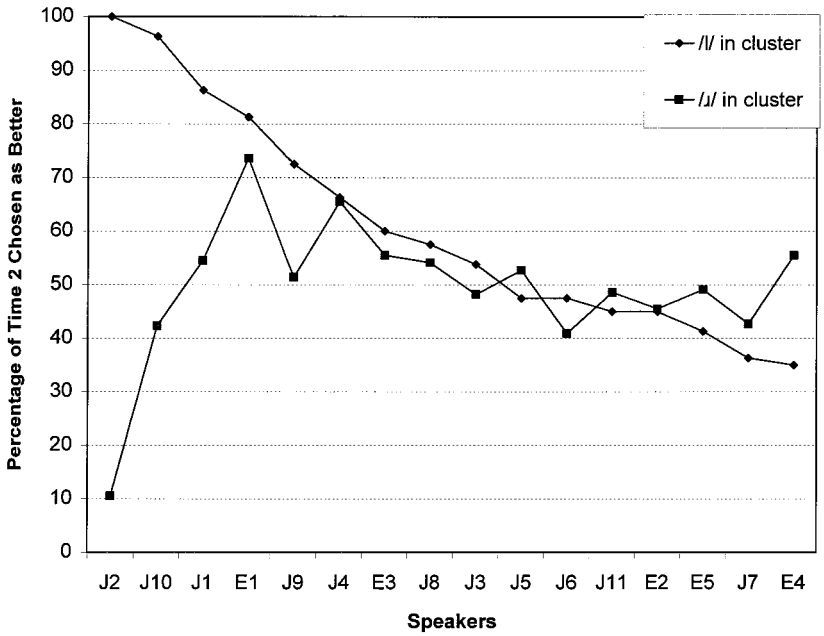


Figure 9. Speaker ranked by percentage of Time 2 tokens chosen as better for /l/ in clusters in Experiment 3. (J = Japanese, E = English.)

most of the Japanese are above the 50th percentile, which indicates a general preference for T2 tokens over T1 tokens. One also sees that one of the NE speakers (E1) is well above the 50th percentile and another (E4) is well below it, an unexpected finding that will be discussed below.

In Figure 9, the preference scores for /ɹ/ in clusters are each based on a total of 220 judgments (11 paired words × 10 listeners × 2 orders). The preference scores for /l/ in clusters in this same figure are based on 80 judgments (4 paired words × 10 listeners × 2 orders). The speakers are arranged in the order of highest T2 preference scores for /l/ in clusters, and on this basis one sees that a little more than half of the NJ speakers are above the 50th percentile (and some by a wide margin), which indicates a general preference for T2 tokens over T1 tokens. One again sees, as in Figure 8, that E1 is well above the 50th percentile and E4 is well below it.

The preference scores obtained for the NE speakers were submitted to a Consonant (/ɹ/ vs. /l/) × Onset Type (singleton vs. cluster) repeated-measures ANOVA. (Note that "Time" was not a factor in the ANOVA because the T1 and T2 tokens were presented together in pairs.) The preference scores obtained for the NE speakers' productions of /ɹ/ singletons (54.6%), /ɹ/ in clusters (55.8%), /l/ singletons (53.4%), and /l/ in clusters (52.5%) differed very little. Not surprisingly, neither the main effects of Consonant $F(1, 4) = 0.16$; Onset Type, $F(1, 4) = 0.02$; nor the Consonant Type × Onset interaction, $F(1, 4) = 1.79$, reached significance ($p > .10$).

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

The preference scores obtained for the NJ speakers' productions were /ɹ/ singletons: 56.9%, /ɹ/ in clusters: 45.6%, /l/ singletons: 61.3%, and /l/ in clusters: 67.5%. The ANOVA examining the NJ speakers' scores yielded a Consonant × Onset Type interaction that was marginally significant, $F(1, 8) = 3.78$, $p = .088$. Tests of simple main effects revealed that the difference in preference scores for /ɹ/ and /l/ singletons (57% vs. 61%) was clearly insignificant, $F(1, 12) = 0.17$, $p > .10$; but that the difference between /ɹ/ in clusters and /l/ in clusters (46% vs. 68%) was marginally significant, $F(1, 12) = 4.14$, $p = .064$.

Finally, *t*-tests were carried out to determine if the preference scores differed significantly from a chance score of 50%. For the NE speakers, the *t*-tests were insignificant for /ɹ/ singletons, $t = 1.10$, $p = .33$; /ɹ/ in clusters, $t = 1.20$, $p = .29$; /l/ singletons, $t = 0.39$, $p = .71$; and /l/ in clusters, $t = 0.32$, $p = 0.78$. For the NJ speakers, the *t*-tests were also insignificant for /ɹ/ singletons, $t = 0.89$, $p = .44$; /ɹ/ in clusters, $t = –0.86$, $p = .41$; and for /l/ singletons, $t = 1.81$, $p = .11$. However, the *t*-test was significant for the /l/ in clusters, $t = 2.28$, $p = .05$. Thus, taken together, the results of Experiment 3 agree with those of Experiment 2 in showing that the NJ speakers' productions of /l/ in clusters improved from T1 to T2 to a greater extent than did their productions of /ɹ/ in clusters, /ɹ/ singletons, or /l/ singletons.

These findings can be evaluated in terms of the two hypotheses presented above. Hypothesis 1 (Frequencies Hypothesis), which predicted that /l/ singletons would improve in accuracy more than /ɹ/ singletons between T1 and T2, was not supported. No difference was found in the T1–T2 preference scores for /l/ singletons (or for /ɹ/ singletons). The results of this experiment, combined with those of Experiment 2, however, suggest that, although /l/ did not improve significantly more than /ɹ/ between T1 and T2, /l/ in both singletons and clusters tended to be better than /ɹ/. The mean preference scores, for example, obtained for the 11 NJ productions (T1 and T2 combined) were /ɹ/ in clusters: 45.6%; /ɹ/ singletons: 56.9%; /l/ singletons: 61.3%; and /l/ in clusters: 67.5%.

The first prediction of Hypothesis 2 (Perception-Production Hypothesis) was that between T1 and T2 the /ɹ/ singletons would improve more in accuracy than the /l/ singletons, but this was not supported. The second prediction of Hypothesis 2 was that between T1 and T2, /l/ in clusters would improve more in accuracy than the /ɹ/ in clusters, and this, however, was supported. The results show significant improvement for /l/ in clusters and not for /ɹ/ in clusters. Both findings mirror and appear to be related to that of the previous experiment. Recall that Bradlow et al. (1997), Mochizuki (1981), Sheldon and Strange (1982), and Takagi (1993) all found /l/ in clusters to be more easily identified than /ɹ/ in clusters. In the current study, however, the /l/ in clusters had been identified less well at T1, and had more room in which to improve than did the /ɹ/ in clusters. (See Figures 3 and 5.)

One unexpected outcome of Experiment 3 concerned the NE control group. In Experiment 1, sentences spoken by members of the NE control group were all rated at near ceiling. In Experiment 2, more than 90% of all tokens spoken by NE speakers were correctly identified. In Experiment 3, it was expected that

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

/ɹ/s and /l/s spoken at T2 by NE speakers would be chosen at 50% (representing no improvement vis-a-vis their T1 tokens). Speaker E4, however, got only 34% of her T2 /l/-singleton tokens selected as "better," and Speaker E1 got more than 70% of her T2 tokens for all four liquid onset types chosen as "better." These unexpected findings may have been artifacts of the noise reduction process applied to the stimuli, which may not have affected all files the same way. The listeners were told to ignore everything except for the /ɹ/ and /l/ segment. However, some may have selected as "better" a token with less noise. These artifacts may have been evident here but not in Experiment 2 because (a) the listeners in Experiment 3 were not trained, and (b) a paired comparison format is far more sensitive to stimulus preparation than a forced-choice identification task.[9]

GENERAL DISCUSSION

The primary purpose of this study was to examine changes in foreign accent over time. Experiment 1 showed that, of 11 NJ speakers, the sentences of three significantly improved (i.e., had less foreign accent). The fact that the two NJ speakers who improved the most between T1 and T2 were also the two to have spent the most time immersed in a TL environment suggests that immersion in the TL may have been the key factor in these speakers' global foreign accent improvement. At a micro-level, Experiments 2 and 3 assessed the identifiability and accuracy, respectively, of English /ɹ/ and /l/ as singletons and as members of clusters. Here two different types of experiments (identification and paired-comparison) using two different sets of listeners (3 trained listeners and 10 untrained listeners) produced convergent results. For the NJ group, significant improvement occurred between T1 and T2 for /ɹ/ or /l/ in only one onset type (/l/ in clusters) and not in the others (/l/ singletons, /ɹ/ singletons, and /ɹ/ in clusters).

A general question of interest with respect to /ɹ/ and /l/ production was whether the identifiability and accuracy of /ɹ/ and /l/ production would reflect NJ speakers' global accent ratings. For example, would the improvement in the global foreign accent of speakers J1, J2, and J4 be matched by an improvement in the production of English /ɹ/ and /l/? In Experiment 2 (summarized in Table 2), the identifiability of speaker J4's liquids was found to improve from T1 to T2 in all four types of onsets (for /ɹ/ singletons, T1: 20%, T2: 100%; /ɹ/ in clusters, T1: 78.8%, T2: 97.0%; /l/ singletons, T1: 80.0%, T2: 93.3%; and /l/ in clusters T1: 33.3%, T2: 58.3%). The identifiability of J1's liquids also clearly improved (for /ɹ/ singletons, T1: 73%, T2: 100%; /ɹ/ in clusters, T1: 100%, T2: 97.0%; /l/ singletons, T1: 80.0%, T2: 100%; and /l/ in clusters, T1: 33.3%, T2: 100%). The scores for J6 did not show general improvement (for /ɹ/ singletons, T1: 40%, T2: 60%; /ɹ/ in clusters, T1: 84.8%, T2: 78.8%; /l/ singletons, T1: 86.7%, T2: 86.7%; and /l/ in clusters, T1: 33.3%, T2: 33.3%); nor did J's global foreign accent rating improve. Thus, for speakers J1, J4, and J6, at least, global accent and the identifiability of liquids followed somewhat parallel developments.

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

Change in liquid identifiability and change in global foreign accent, however, were not linked for all NJ speakers. In three of four onset types, speaker J7's liquids were identified at T2 less well than at T1 (/ɹ/ singletons, T1: 20.0%, T2: 13.3%; /ɹ/ in clusters, T1: 87.9%, T2: 75.8%; /l/ singletons, T1: 20.0%, T2: 6.7%; and /l/ in clusters, T1: 16.7%, T2: 16.7%), yet her global accent rating was about the same at T2 and at T1. Speaker J11's liquids were generally more identifiable at T2 than at T1 (/ɹ/ singletons, T1: 0%, T2: 93%; /ɹ/ in clusters, T1: 51.5%, T2: 51.5%; /l/ singletons, T1: 86.7%, T2: 100%; and /l/ in clusters, T1: 41.7%, T2: 58.3%), but there was no difference in global foreign accent rating at T2 and T1. In the case of J7 and J11, therefore, changes in the identifiability of liquids between T1 and T2 occurred somewhat independently of changes in global accent between T1 and T2. (Changes in global accent of course could nonetheless be related to changes in segmentals other than liquids.)

Bradlow et al. (1997) also observed widespread individual variation in their study of both perception and production. They found that "the processes of learning in the two domains appear to be distinct within individual subjects," and "it is not the case that improvement in perception and production proceeded in parallel within individual subjects" (p. 2307). It also appears not to be the case that improvement in global accent necessarily proceeds in parallel with improvement in any particular smaller components of pronunciation, such as segmental identifiability and accuracy.

In some cases, changes in liquid identifiability appeared to be a function of the relationship between /ɹ/ and /l/. For example, speaker J2's global accent improved from T1 to T2 (Figure 1). However, whereas the identifiability of her /l/ singletons and /l/ in clusters got better, the identifiability of her /ɹ/ singletons and /ɹ/ in clusters got much worse (Table 2). At T1, speaker J2 may have used a modified Japanese tap for both target /ɹ/ and /l/ in clusters and that form was usually identified as /ɹ/; then at T2, after having spent a year in California, J2 may have shifted to a new phonetic form for both /ɹ/ and /l/ and that form was usually identified as /l/ rather than /ɹ/.[10] This interpretation is supported by the results of Experiment 3, in which her T2 /l/ in clusters tokens were chosen as better 100% of the time, whereas her T2 /ɹ/ in clusters tokens were chosen only 11.0% of the time.

One control-group speaker, E3, was the only speaker whose liquid productions in all four onset types were identified correctly 100% of the time (see Table 2), yet he received the lowest ratings among the five NE speakers for global accent in Experiment 1. Speaker E3's accent rating was 8.4 at T1 and 8.5 at T2. All of the other NE speakers received at least 8.8 at T1 and T2 and several received 9.0 (see Figure 1). Perhaps the speech of E3 was so carefully articulated that it marked him as a potentially nonnative speaker, and when the five listeners in Experiment 1 were urged to use the entire 9-point scale and had to give one of the native speakers an 8 instead of a 9, they gave the 8 to E3. If one wants to be perceived as a native, hyperarticulated segments may be a liability.

Six general findings may offer insight into what would constitute an appropriate L2 pronunciation curriculum and suggest areas for future L2 pronuncia-

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

tion research: (a) The global foreign accent of every individual either stayed the same or got better; it did not get worse. (b) Improvement in global foreign accent was to some extent independent of segmental production accuracy. (c) Whereas native and nonnative speakers' ratings were distinctly different for global foreign accent, the two groups of ratings overlapped for segmental identifiability and accuracy. (d) Heavily foreign-accented speech sometimes had highly identifiable segments. (e) Some L2 adult speakers significantly improved both their global accent and their segmentals in a few years time and in a predominantly nonimmersion setting. (f) Between the extremes of having a native-speaker pronunciation and having a heavily accented and fossilized L2 pronunciation, there was middle ground across which progress was made.

The findings of this project, however, must be interpreted with caution. First of all, the NJ stimuli for all three experiments were derived from only five sentences and 25 words and were produced by a small group of 11 NJ volunteers who may not be representative of L2 learners or even NJ students. Secondly, whereas it was known that cross-linguistic misidentifications between Japanese and English liquids involve more than just liquids, Experiments 2 and 3 were restricted to the phonemic categories of /ɹ/ and /l/. Thirdly, prevocalic American English /ɹ/ and /l/ in word onsets were chosen for Experiments 2 and 3 because they constituted an area of known difficulty for NJ speakers of English, in which there would presumably be room for improvement between T1 and T2 but not necessarily any improvement made. Had Experiments 2 and 3 focused on other English segments or a few isolated cluster types, then the findings might have been quite different. Finally, and this applies to Experiment 3 only, it appears that some of our untrained listeners, when presented with two perhaps very similar tokens, resorted to distinguishing the two on the basis of software-induced noise distortion rather than on the basis of the tokens themselves. This problem, which was detected by virtue of having a control group, can be avoided in the future by making cleaner recordings that would obviate the need for applying acoustic noise-reduction software in the first place.

A secondary purpose of this project was to examine to what extent L2 theories and theories of production and perception could account for the changes (or lack thereof) in accent observed between T1 and T2. In preparation for Hypothesis 1, evidence was presented that both American English /ɹ/ and /l/ were different from any sounds in Japanese and that /ɹ/ was clearly the more marked member of the pair. This led to the Frequencies Hypothesis that, between T1 and T2, /l/ singletons would improve more in identifiability (Experiment 2) and accuracy (Experiment 3) than /ɹ/singletons would. Based on the motor theory and direct-realist approach, which both assume a direct linkage between perception and production, and based on previous empirical studies of NJ perception and production of /ɹ/ and /l/, a second and opposing hypothesis was formulated, the Production-Perception Hypothesis, which stated that for any phoneme type (phonetic segment) measures for production would mirror those of perception. Because previous studies (see Figure 6) had determined that NJ perceived /ɹ/ singletons more accurately than /l/ singletons, one prediction of this second

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

hypothesis was that between T1 and T2 NJ speakers would improve in producing /ɹ/ more intelligibly (Experiment 2) and accurately (Experiment 3) than /l/.

As it turned out, neither /ɹ/ nor /l/ singletons improved significantly in identifiability and accuracy between T1 and T2, and neither of the two opposing hypotheses for singletons was supported. In defense of the Frequencies Hypothesis, it was noted, however, that /l/ singletons (unlike /ɹ/ singletons) had relatively high identification scores at T1 and may have been constrained by a ceiling effect between T1 and T2 (compare Figures 2 and 4). Figure 6 provides ample counterevidence against the Frequencies Hypothesis (favoring /l/ over /ɹ/) for perception, but the testing of this hypothesis for production requires further investigation (see Figure 7).

A second prediction of the Perception-Production Hypothesis was supported. Between T1 and T2, /l/ in clusters improved more in intelligibility (Experiment 2) and accuracy (Experiment 3) than /ɹ/ in clusters. This time, however, it was the /ɹ/ in clusters that had the relatively high scores at T1 and that may have been subject to a ceiling effect between T1 and T2 (compare Figures 3 and 5). The two opposed findings, a disfavorable one involving singletons and a favorable one involving clusters, provided mixed support for the Perception-Production Hypothesis.

Variation in production scores for /ɹ/ and /l/ in clusters is undoubtedly related to a number of factors. The study by Bradlow et al. (1997) suggests that short, intensive training may yield different results from that of naturalistic learning (see Figures 6 and 7). Phonological context, also, has been found to be important in L2 speech (Carlisle, 1994), as it is in L1 speech (e.g., Ladefoged, 1993). As a unit of analysis, the cluster (as opposed to the singleton) may conceal a host of phonotactic variables (see Buckingham and Yule, 1987) that have not been given sufficient attention.

Additionally, different experimental designs also make it difficult to interpret the findings about clusters that have been reported across studies. Although the current study examined production and a markedness-based hypothesis, its design is not comparable to that of markedness studies (e.g., Eckman 1991). First of all, typically, the data presented in markedness studies of clusters are based on phonetic transcriptions of only one or two individuals. The data of Experiment 2 were based on the phonemic perceptions of three trained listeners, all of whom were blinded to the source (T1 or T2) of the stimuli. Second, markedness studies that have used Greenberg's (1978) implicational universals have used the entire cluster as the unit of analysis. In the experiments reported here, however, the listeners were instructed to ignore segments other than /ɹ/ and /l/. Finally, a number of markedness studies have relied on an arbitrary 80% cutoff point to distinguish what is "acquired" from what is not. Here, we counted all tokens and represented all scores as percentages. In sum, the variety of cluster types that have been included within and across studies and the different methods that have been used to examine them make it difficult to interpret and relate the findings about clusters.

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

At present we are aware of no existing theoretical model that can account for all of the following: (a) Untrained NJ speakers have been consistently found to identify singleton /ɹ/ with greater accuracy than singleton /l/ (see Figure 6) but not to produce it with consistently greater accuracy (see Figure 7). (b) NJ speakers have been consistently found to identify /l/ in clusters with greater accuracy than /ɹ/ in clusters (Bradlow et al., 1997; Mochizuki, 1981; Sheldon & Strange, 1982; Takagi, 1993) but not to consistently produce the former with greater accuracy than the latter (cf. Bradlow et al., 1997 and this study). (c) Over time, greater accuracy in perception does not necessarily lead to greater accuracy in production. One possible answer to these questions may reside in theory—that there is less linkage between perception and production than that assumed by the motor theory and the direct realist approach. Another possible answer, however, may reside in methodology. Future researchers (ourselves included), for example, might want to be more careful in controlling and reporting the phonotactic variation that resides beneath the label of cluster as they investigate perception and production.

NOTES

1. The NJ speakers were never given a hearing test. Their grade point averages (GPAs) (Table 1) and their numerous conversations with the first author of this paper (TJR), however, lead us to believe that their hearing was normal.

2. An anonymous reviewer advised us to revise our Perception-Production Hypothesis based on the findings of Bradlow et al. (1997). The purpose of Bradlow et al. was to investigate "the effect of perceptual learning on subsequent performance in both perception and production" (p. 2300). Their study involved 11 NJ undergraduate university students aged 19–22 and four phases: pretest, perceptual training, posttest, and production assessment. The perceptual training phase, which lasted 3–4 weeks, involved 45 training sessions and 68 minimal pairs that contrasted /ɹ/ and /l/ in multiple phonetic environments. The production assessment phase included two tests: the first involving paired comparison, and the second, identification. Both involved whole-word tokens.

For the purpose of the present study, when interpreting the results of Bradlow et al. (1997) about NJ perception of /ɹ/ and /l/, it is important to distinguish their findings for word-initial singleton liquids and for liquids in word-initial clusters from their general findings that represent the average of five or more phonotactic environments combined. Additionally, it is important to distinguish their pretest results from their posttest results that were observed after explicit training, not naturalistic learning. The importance of these distinctions for the current study will now be pointed out.

Bradlow et al. (1997) reported three general findings (p. 2302): (a) an overall improvement in the identification of liquids from pretest to posttest, (b) a higher identification score for /ɹ/ than for /l/ at both pretest and posttest, and (c) greater improvement for /l/ than for /ɹ/ between pretest and posttest. These three general findings, however, represent the findings from five phonetic environments averaged. They do not reflect what Bradlow et al. observed for liquids (expressed as percentages in Figure 2, p. 2302, and Table 1, p. 2305, and also provided below) that appeared as word-initial singletons and in word-initial clusters, the two environments investigated in the present study.

Contrary to the first general finding of Bradlow et al. (1997) (i.e., an overall improvement in the identification of liquids from pretest to posttest), /ɹ/ singleton did not improve. In fact, between pretest and posttest, its identification rate decreased somewhat from 84% to 74%, whereas /l/ singletons improved from 44% to 85%. Contrary to their second general finding (i.e., that /ɹ/ was more accurately identified than /l/), in word-initial clusters, /l/ (pretest: 70%, posttest: 82%) was identified more accurately than /ɹ/ (pretest: 50%, posttest: 70%) at both pretest and posttest. This finding replicated a number of previous findings and Bradlow et al. acknowledged "the high identification accuracy of /l/ in initial cluster position relative to /ɹ/ in that environment" (p. 2302). The third general finding reported by Bradlow et al. for perception was that /l/ improved more than /ɹ/. Although /l/ singletons (pretest: 44%, posttest: 85%) improved more than /ɹ/ singletons (pretest: 84%, posttest: 74%), /ɹ/ in clusters (pretest: 50%, posttest: 70%) improved more than /l/ in clusters (pretest: 70%, posttest: 82%).

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

An additional finding that Bradlow et al. (1997, p. 2302) reported was an ordering of phonetic environments in which the initial singleton liquids were more accurately identified than the liquids in clusters. However, /l/ singletons (44%) were identified less accurately than /l/ in clusters (70%) at pretest, and the two are about the same (85% and 82% respectively) at posttest. Furthermore, although /ɹ/ singletons (84%) were identified more accurately at pretest than /ɹ/ in clusters (50%), the two are identified about the same rate at posttest (74% and 70%).

Thus, when interpreting the results of Bradlow et al. (1997) for the present study, it is important to distinguish the pretest from the posttest and the phonetic environments from one another. The group of NJ speakers for the present study are only comparable to the pretest and control groups of Bradlow et al. (1997) because these are the groups that had no explicit training. The pretest findings for Bradlow et al. (1997) for word-initial liquid singletons and liquids in word-initial clusters replicate rather than contradict the findings of Mochizuki (1981), Sheldon and Strange (1982), and Takagi (1993). Thus, rather than revise our Perception-Production Hypothesis based on Bradlow et al., as the anonymous reader suggested, we have chosen to add Bradlow et al. (1997) to those three studies that contribute to the rationale for the hypothesis. All four studies found /ɹ/ singletons to be perceived more accurately than /l/ singletons, and /l/ in clusters to be perceived more accurately than /ɹ/ in clusters.

3. The one longitudinal study that we are aware of, Bradlow et al. (1997), involved explicit training, not naturalistic learning.

4. Yamada, Tohkura, and Kobayashi (1994) found in a study that examined Japanese perception of English /ɹ/ and /l/ that word familiarity may affect nonnative phoneme perception. Japanese subjects, for example, who are familiar with *red* but not *led* are apt to hear *red* when *led* is spoken. Although the familiarity of the NJ with these words was never measured at T1 or T2, two procedures were used shortly after T2 to estimate what these students' familiarity must have been at T1. One procedure involved checking both editions of *Zen Eiren* (*Zen Eiren*, 1967, 1988) a well-known list of English vocabulary that is recommended to be taught in Japanese secondary schools. Of the 25 words used for Experiments 2 and 3, only one, the word *splinter*, was not listed in both editions of *Zen Eiren*. A second procedure involved checking the familiarity of the 25 words based on self-reports of a similar population of Japanese freshmen. For this latter procedure, in June of 1996, 41 freshman students at ICU were given a copy of the same word list that had been given to the 11 NJ at T1 as freshmen in May of 1992. Students were told to circle any word that they did not remember having seen before and to underline any word the familiarity status of which they were uncertain about. They were given as long as they needed to do this. Of 42 students, all indicated familiarity with 16 of the 25 words. One student indicated unfamiliarity with four words (*strange, screen, lend*, and *risk*), 5 indicated unfamiliarity with *trunk*, 6 with *lunched*, 12 with *splinter*, 13 with *splendid*, and 21 with *rag*. In sum, in 61 of 672 (25×42) possible instances, speakers indicated at least some degree of uncertainty about their familiarity with a word on the list, and 46 of these 61 instances involved one of three words: *rag, splinter*, and *splendid*. Two of these three words, *splinter* and *splendid*, involved /l/ in initial clusters, and one word, *rag*, involved /ɹ/ as an initial singleton. Of the 42 words in the five sentences, only *splendid* was chosen as slightly unfamiliar by some students. Although it was assumed that their relative unfamiliarity might be a problem, as it turned out, *splinter* and *splendid* were two of the four words that comprised the /l/ in cluster onset type, the only type to show significant improvement in intelligibility (Experiment 2) and accuracy (Experiment 3) between T1 and T2.

5. The ANOVA technique can be applied validly when variances differ across groups, but only if the groups consist of the same number of speakers, which was not the case here.

6. For the /ɹ/ singletons and /l/ singletons, a total of 960 judgments were obtained. The /ɹ/ singleton scores shown in Figure 2 were based on 480 judgments (16 speakers $\times$ 5 /ɹ/ singleton $\times$ 3 judges $\times$ 2 times). The /l/ singleton scores shown in Figure 4 were also based on 480 judgments (16 speakers $\times$ 5 /l/ singletons $\times$ 3 judges $\times$ 2 times). For the /ɹ/ in clusters and /l/ in clusters, a total of 1,440 judgments were made in all. The /ɹ/ in cluster scores (Figure 3) were each based on 1,056 judgments (16 speakers $\times$ 11 /ɹ/ in clusters $\times$ 3 judges $\times$ 2 times). The /l/ in cluster scores (Figure 5) were each based on a total of 384 judgments (16 speakers $\times$ 4 /l/ in clusters $\times$ 3 judges $\times$ 2 times).

7. One should be cautious in comparing the results of the various studies that have been cited. Bradlow et al. (1997), the only study to involve explicit training, has been discussed in note 2. Mochizuki (1981) used tokens produced by only one speaker. Sheldon and Strange (1982) presented listeners with word stimuli based on the /ɹ/ and /l/ contrast (e.g., *room* or *loom*) and had listeners respond by "writing the word they heard" (p. 248). Furthermore, some words used by Sheldon and Strange are listed in Japanese dictionaries as Japanese words of English origin under a kana spelling. One such example is *ruumu* ("room"), which is pronounced /ɾuumu/ in Japanese. NJ productions for both members of the English minimal pair *room* (/ɹum/) and *loom* (/lum/), might be the form of the familiar

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

Japanese cognate /ɾuumu/. If, for example, NE listeners hear two stimuli in an identification experiment, /ɹuumu/ and /ɹum/ and are asked to identify which one is *room*, they may decide on the basis of the difference between the two words, the presence or absence of the final vowel, rather than on the basis of the liquid, which in both forms is the same. Thus, the subjects tested by Sheldon and Strange (1982), when asked to identify and circle "the word" that they heard, may have been responding instead to deviant neighboring segments or some other variation. (This possibility provided part of the rationale for using edited demisyllables rather than whole words for Experiments 2 and 3.)

8. A preference score of 70%, for example, would indicate that the T2 token was chosen in 70% of instances, and that the T2 tokens conformed more nearly to the phonetic norms of American English (for /ɹ/ or /l/) than did the T1 tokens. (The T2 token would be selected by chance in only 50% of instances because the T1 and T2 tokens appeared randomly and with equal frequency in the first and second positions of the pair.)

9. In the results of Experiment 2, of 250 trials (5 blocks × 50 stimuli) involving the identification of the /ɹ/ and /l/ produced by the NE speakers, all three judges (given the three choices of "L," "R," and "NEITHER") made the same judgments 94.8% of the time, and at least two of the three agreed 99.6% of the time. Of the 550 trials (11 blocks × 50 stimuli) involving the identification of the /ɹ/ and /l/ produced by the NJ speakers, all three judges agreed 78.0% of the time and at least two of the three agreed 97.8% of the time. The range of identifiability of the five members of the control group, averaged over the four liquid onset types, was from 92.5% to 100%. This is about what we expected for Experiment 2.

10. To the extent that the one liquid got higher ratings for identifiability, the other got lower ratings. At T1, speaker J2's /ɹ/ in clusters were correctly identified 97.0.% of the time, but her /l/ in clusters only 8.3% of the time; at T2 her /ɹ/ in clusters were correctly identified only 24.2% of the time, her /l/ in clusters were identified 100% of the time.

REFERENCES

Best, C. T. (1995). A direct-realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross language speech research* (pp. 171–206). Timonium, MD: York Press.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101*, 2299–2310.

Brown, V. (1970). *Improving your pronunciation*. Tokyo: Meirindo.

Buckingham, H., & Yule, G. (1987). Phonemic false evaluation: Theoretical and clinical aspects. *Clinical Linguistics and Phonetics, 1*, 113–125.

Carlisle, R. S. (1994). Markedness and environment as internal constraints on the variability of interlanguage phonology. In M. Yavas (Ed.), *First and second language phonology* (pp. 223–249). San Diego, CA: Singular Publishing Group.

Cool Edit (Version 1.53) [Computer software]. (1996). Phoenix, AZ: Syntrillium Software Corporation.

Eckman, F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning, 27*, 315–330.

Eckman, F. R. (1991). The structural conformity hypothesis and the acquisition of consonant clusters in the interlanguage of ESL learners. *Studies in Second Language Acquisition, 13*, 23–41.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America, 84*, 70–79.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America, 91*, 370–389.

Flege, J. E., Munro, M., & MacKay, I. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*, 1–26.

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

Flege, J. E., 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.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3–28.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R." *Neuropsychologia, 9*, 317–323.

Greenberg, J. (1978). Some generalizations concerning initial and final consonant clusters. In J. Green-

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press

berg, C. A. Ferguson, & E. Moravcsik (Eds.), *Universals of human language, Vol. 2* (pp. 243–279). Stanford, CA: Stanford University Press.
Ladefoged, P. (1993). *A course in phonetics* (3rd Ed.). Tokyo: Harcourt, Brace, & Jovanovich Publishers.
Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Cambridge, MA: Blackwell.
Laver, J. (1994). *Principles of phonetics*. New York: Cambridge University Press.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.
Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.
Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 245*, 489–494.
Maddieson, I. (1984). *Patterns of sounds*. New York: Cambridge University Press.
Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics, 9*, 283–303.
Munro, M., & Derwing, T. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning, 45*, 73–97.
Price, P. (1981). A cross-linguistic study of flaps in Japanese and in American English. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.
Purcell, E. T., & Suter, R. W. (1980). Predictors of pronunciation accuracy. A reexamination. *Language Learning, 30*, 271–287.
Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21*, 427–444.
Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics, 3*, 243–261.
Takagi, T. (1993). Perception of American English /r/ and /l/ by adult Japanese learners of English: A unified view. Unpublished doctoral dissertation, University of California, Irvine.
Vance, T. (1987). *An introduction to Japanese phonology*. Albany, NY: State University of New York Press.
Yamada, R., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception and Psychophysics, 52*, 376–392.
Yamada, R., Tohkura, Y., & Kobayashi, N. (1994). Effect of word familiarity on nonnative phoneme perception: Identification of English /r/, /l/, and /w/ by native speakers of Japanese. In A. James & J. Leather (Eds.), *Second Language Speech* (pp. 103–117). The Hague: Mouton.
Zenkoku eigo kyoiku kenkyu dantai rengokai (*ZEN EIREN*) [National English Education Research Organization]. (1967). *Kookoo kihon eitango katsuyoshu*. [High School Basic Practical English Vocabulary Collection]. Tokyo: Kenkyusha.
Zenkoku eigo kyoiku kenkyu dantai rengokai (*ZEN EIREN*) [National English Education Research Organization]. (1988). *Kookoo kihon eitango katsuyoshu*. [High School Basic Practical English Vocabulary Collection]. Tokyo: Kenkyusha.

https://doi.org/10.1017/S0272263198002058 Published online by Cambridge University Press