PERCEPTION-PRODUCTION RELATIONSHIP FOR /r–l/ BY NATIVE JAPANESE SPEAKERS

Erin M. Ingvalson[a] *& Lori L. Holt*[b]

[a]Northwestern University, USA; [b]Carnegie Mellon University, USA
ingvalson@northwestern.edu; lholt@andrew.cmu.edu

ABSTRACT

We examined the possible relationship between perception and production of /r–l/ by native Japanese speakers. Little evidence of a relationship between the cues used for perception and those for production was found for native Japanese and native English speakers. However, we found evidence of a shift from a reliance on conjunctive cues to single cues as listeners progressed from relatively naïve to fully native English speakers.

Keywords: speech perception, speech production, Japanese /r–l/

1. INTRODUCTION

Language learners, both of first (L1) and second (L2) languages, must learn to both perceive and produce the language's speech sounds. These two processes are thought to be related, not least because training listeners to perceive non-native speech sounds results in improved production of the sounds [1, 8]. Indeed, it has been hypothesized that L2 learners can only produce L2 sounds accurately if they perceive them accurately [2].

Despite this hypothesized relationship, the abilities to perceive and produce L2 speech sounds are weakly related at best. Flege and colleagues [3, 4] found a correlation in L2 learners' perception of L2 vowels and in the intelligibility of vowel productions as well as a relationship between the acoustic cues to vowel category in perception and production. However, despite this relationship, much of the variance in production was unaccounted for by perception. Similarly, Golestani and colleagues [5], found no relationship between learners who were fast to learn to perceive a non-native contrast and those who were fast to learn to produce it.

The aim of the present work was to further examine the relationship between perception and production in the context of native Japanese (NJ) speakers' ability to perceive and produce the English /r–l/. It has been previously demonstrated that the most reliable cue to /r–l/ category membership is the onset frequency of F3 [7, 9]. Though F3 is the primary cue to category membership, F1 steady-state and transition duration is a co-varying cue [10], and one that may be more beneficial to NJ learners [7]. Our goals were threefold: 1) to determine if F3 and F1 usage changed in perception and production as length of residency (LOR) in an English environment increased, 2) to determine if perception and production were related and, if so, at what levels, and 3) if the relationship between perception and production changed as a function of LOR.

2. METHODS

2.1. Participants

Six native English (NE; 3 females and 3 males) speakers served as controls. Five listeners were monolingual English speakers; one listener began learning German in high school, studied abroad in Germany for one year, and considered himself highly proficient in German. All NE participants were recruited from the Carnegie Mellon community and all reported normal hearing.

Forty native Japanese (NJ) speakers participated. NJ listeners were divided on the basis of length of residency (LOR) in North America. Twenty-five participants (21 females, mean age 43.24 years) had LORs of ten or more years. Fifteen participants (4 females, mean age = 29.80 years) had LORs of two or fewer years. NJ participants were recruited from the greater Pittsburgh, Pennsylvania area; the greater Vancouver, British Columbia area; and the greater San Francisco, California area. One listener with a residency of less than two years had been raised simultaneously Japanese-Chinese bilingual in Japan. The remaining 24 listeners were all raised in monolingual Japanese homes. Their first exposure to English was at age 12 or 13, when it was introduced in school. All participants were first immersed in English when they came to North America and immersion occurred after age 18.

2.2. Perception stimuli

We created 185 synthetic stimulus tokens ranging from /rait/ to /lait/ that systematically varied F1 steady state and transition duration, F2 onset frequency, and F3 onset frequency. The synthetic portion of the stimulus was 360 ms and a natural /t/ was appended to all stimuli. For all stimuli, F1 onset frequency was 400 Hz and the onset of the /a/ portion of the diphthong was 750 Hz. Total duration from the onset of the stimulus to the onset of the diphthong was 150 ms. The steady-state portion was 83 ms, 95 ms, 110 ms, 125 ms, or 137 ms long. The transition duration was the difference between the steady state duration and 150 ms; all transitions were linear from 400 to 750 Hz.

For each F1 steady-state/transition 37 stimuli varying only in F2 and F3 onset were created. F3 onset frequency varied from 1200 to 3000 Hz in 200 Hz steps; F2 onset frequency varied from 800 to 1400 Hz in 200 Hz steps. Each stimulus was a unique intersection of a particular F2 and F3 onset frequency. F3 initial steady-state was 80 ms and the linear transition to the /a/ portion of the diphthong—2465 Hz—was 70 ms. F2 onset frequency steady-state duration varied to ensure the slope of the transition to the /a/ was consistent across all tokens. At 800 Hz the steady state duration was 80 ms; at 1400 Hz the steady state duration was 150 ms. The /i/ portion of the diphthong—2190 Hz—was reached at 245 ms.

The remaining details of the stimuli are in [6]. All stimuli were RMS matched in energy, sampled at 11025 Hz, and presented diotically over headphones at approximately 70 dB.

We also used a set of /r–l/ minimal pair natural speech stimuli described in [6]

2.3. Production stimuli

A monolingual American English speaker with no discernable regional accent produced two tokens of 100 /r–l/ minimal pairs. The second token was used as an auditory prompt in the production task. Prompts were RMS matched in energy, sampled at 11025, and presented in stereo over laptop speakers at approximately 70 dB.

2.4. Procedure

Participants heard five instances of each synthetic token and indicated whether the sound began with /r/, /l/, or /w/. In the natural speech task, listeners heard two instances of each member of a minimal pair [1]. Both members of the pair were presented on the monitor and listeners indicated which member of the pair was heard. Natural speech stimuli were blocked by talker [1].

In the production task, participants saw an orthographic representation of the token while hearing the prompt simultaneously. Following a delay of 700 ms, they produced the token. Following a second delay of 3000 ms, they produced the token again. This procedure was repeated once more, resulting in four total tokens. The last three of these were selected.

Participants' productions were cropped and centered in silence to be 100 ms in duration, RMS matched in energy, and presented diotically over headphones at 70 dB to 34 monolingual NE listeners. Listeners saw a phonetically-based orthographic representation of the auditory stimulus with the critical portion removed (e.g., "_ude"). They indicated which sound best completed the auditory stimulus, noting that it might not be an English word.

3. RESULTS

There were four categories of interest: the proportion /l/ response to the synthetic stimuli as a function of the manipulated parameters, the proportion correct for the natural speech stimuli, the proportion of words heard as intended by the NE listeners for the production stimuli, and acoustic parameters of the syllable-initial /r–l/ production stimuli. This last consisted of F2 and F3 onset frequencies and F1 initial steady state and transition duration. The method for extraction of F2 and F3 onset frequencies can be found in [6]. F1 steady state and transition duration were measured in Praat. Initial steady-state was measured from the onset of periodicity to the beginning of the transition; transition duration was measured from the offset of the initial steady-state to the onset of the initial vowel steady-state.

F1 steady-state, F1 transition duration, F2 onset frequency, and F3 onset frequency from the production stimuli were entered into a robust logistic regression as possible predictors of /r–l/ category membership. The slope of each regressor variable, an indication of /r –l/ category separation along this dimension, was used in all subsequent analysis. Similarly, F1 steady-state, F1 transition duration, F2 onset frequency, and F3 onset frequency were entered into a logistic regression with proportion /l/ response as the predicted

variable. The slope for each potential predictor was used in all following analyses.

We first looked for differences in natural speech perception, the proportion of words heard as intended by NE listeners (intelligibility), and separation along the F3 dimension in perception and production amongst the LOR groupings (native English, NE; LOR of ten or more years, Ten; and LOR of less than two years, Two). Not surprisingly, NE listeners identified more natural speech words correctly /r –l/ ($F(2, 41) = 14.95$, $p < 0.01$, had more of their productions heard as intended, $F(2, 41) = 5.14$, $p < 0.01$, and showed greater separation along the F3 dimension in perception, $F(2, 41) = 46.20$, $p < 0.01$, and production $F(2, 41) = 3.47$, $p = 0.05$ than both groups of NJ listeners. Participants in the Ten group also showed more NE-like performance than members of the Two group on all the above tasks.

Having demonstrated the expected differences amongst the groups, we looked to determine the extent to which perception and production were related for /r–l/. Because of the differences amongst the groups, each group was analyzed independently. For each participant group all variables were entered into a correlation matrix to determine the extent to which perception and production were related at both the natural speech and acoustic level. Perceptual separation along the F1 steady-state and F1 transition dimensions were perfectly correlated for all participant groups; this is not surprising because the dimensions summed to 150 ms for the synthetic speech. Consequently, only F1 steady-state information is discussed below. Similarly, F1 steady-state and F1 transition at the production level were also highly correlated for all listeners and only F1 steady-state information will be presented.

For the NE participants, perception of natural speech was highly correlated with sensitivity to F3 onset frequency when perceiving synthetic /r–l/ $r(4) = 0.97$, $p = 0.01$. The only relationship between perception and production for NE participants was between F1 steady-state for both perception and production, $r(4) = 0.94$, $p = 0.02$. The general lack of a relationship between perception and production is consistent with earlier work indicating that native speakers prefer more extreme tokens than they produce [4].

For the listeners in the Ten group, natural speech perception was correlated with perceptual separation along the F3 dimension, $r(23) = 0.61$, $p < 0.01$, as well as with perceptual separation along the F1 steady-state dimension, $r(23) = 0.68$, $p < 0.01$. Separation along the F3 dimension and the F1 steady-state dimension in perception were correlated, $r(23) = -0.82$, $p < 0.01$. In production, separation along the F2 and F3 dimensions was inversely correlated, $r(23) = 0.44$, $p = 0.03$: as separation along the F3 dimension increased, separation along the F2 dimension decreased. The intelligibility of Ten speakers' productions and their natural speech perception were also related /r –l/, $r(23) = 0.54$, $p = 0.01$. Unlike the NE speakers, who appear to use F3 exclusively, speakers in the Ten group appear to be relying on both F3 onset frequency and F1 transition duration to make perceptual judgments. In production, a shift toward more NE-like cue usage is correlated with a decrease in NJ-like cue usage.

Table 1: Correlations between the manipulated variables in the synthetic stimuli when identified by the Two group.

	F1 Steady-State	F2 Onset Frequency	F3 Onset Frequency
F1 Steady-State	1.00	0.83	0.74
F2 Onset Frequency	0.83	1.00	0.78

In the Two group, separation along the F1 steady state, F2 onset frequency, and F3 onset frequency in perception were all related, Table 1. Separation along the F2 dimension in production was correlated with separation along the F3 dimension in perception, $r(13) = 0.64$, $p = 0.01$, as well as with performance identifying natural speech, $r(13) = 0.55$, $p = 0.04$. Like members of the Ten group, members of the Two group appear to use a conjunction of cues to perceive /r–l/, though in this case it seems to be a conjunction of all available cues. The extent to which their perceptions utilize the most reliable cue is related to the extent to which their productions are separated along the dimension considered more prototypical of NJ productions.

4. DISCUSSION

As anticipated, we saw the most reliable performance perceiving and producing /r–l/ for the NE participants; we also saw more reliable performance for those non-native listeners with longer LORs, suggesting that improved non-native perception and production follows increased L2 experience [3, 4]. The lack of a relationship between perception and production for any of the LOR groupings is consistent with previous work

demonstrating that perception and production may progress at different rates [5]. We found minimal support for the hypothesis that perception precedes production in the Two group [1, 4], demonstrated by more NE-like cue usage in perception is related to more NJ-like cue usage in production for this group with minimal English experience.

The data here, in addition to demonstrating that perception and production may not co-evolve, provide insight as to how perceptual categories develop over the course of immersion. In the Two group, F1 steady-state, F2 onset frequency, and F3 onset frequency are all positively correlated in perception. These relationships indicate inexperienced listeners are using a conjunction of all available cues to differentiate /r–l/, possibly as a result of immature categories. However, more experienced non-native listeners are making use of only those cues that earlier work has shown to be reliable indicators of category membership [7]. It is worth noting that one of these cues, F1 steady state, is also indicative of category membership in the L1 [9], suggesting a possible interaction of L1 and L2 perception strategies. Finally, native listeners rely most heavily on the most reliable cue. In addition to demonstrating how cue weightings might shift as a function of language experience, these data also suggest that even highly proficient bilinguals may not have category structures identical to those of monolinguals, evidenced by the relationships amongst F1 steady-state, F3 onset frequency, and natural speech perception for listeners in the Ten group. A better sense of the category structure of proficient bilinguals will provide a greater understanding of the level of ultimate attainment that can be expected from late language-learners. Understanding how ultimate attainment might be constrained by category structures impacted by the co-existence of two or more languages should result in more efficacious training paradigms that take the constraints on performance into account. We leave it to future research to further examine the category structures and evolutions across L2 learner groups and to develop more customized training paradigms.

5. REFERENCES

[1] Bradlow, A.R., Pisoni, D.B., Akahane-Yamada, R., Tohkura, Y. 1997. Training Japanese listeners to identify English /r/ and /l/ IV: Some effects of perceptual training on speech production. *Journal of the Acoustical Society of America* 104, 2299-2310.

[2] Flege, J.E. 1995. Second-language speech learning: Theory, findings, and problems. In Strange, W. (ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*. Timmonium, MD: York, 233-272.

[3] Flege, J.E., Bohn, O.-S., Jang, S. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25, 437-470.

[4] Flege, J.E., MacKay, I.R.A., Meador, D. 1999. Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America* 106, 2973-2987.

[5] Golestani, N., Molko, N., Dehane, S., LeBihan, D., Pallier, C. 2007. Brain structure predicts learning of foreign speech sounds. *Cerebral Cortex* 17, 575-582.

[6] Ingvalson, E.M., McClelland, J.L., Holt., L.L. 2011. Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics*. (In press).

[7] Iverson, P., Hazan, V., Bannister, K. 2005. Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *Journal of the Acoustical Society of America* 118, 3267-3278.

[8] Lengeris, A., Hazan, V. 2010. The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels to native speakers of Greek. *Journal of the Acoustical Society of America* 128, 3757-3768.

[9] Lotto, A.J., Sato, M., Diehl, R.L. 2004. Mapping the task for the second language learner: Case of Japanese acquisition of /r/ and /l/. *Proceedings of the From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, 1-6.

[10] Polka, L., Strange, W. 1985. Perceptual equivalence of acoustic cues that differentiate /r/ and /l/. *Journal of the Acoustical Society of America* 78, 1187-1197.