Does immersion experience reduce /r/-/l/ category overlap for Japanese learners of English?

Michael D. Tyler, Saya Kawase and Mark Antoniou

School of Social Sciences and Psychology and The MARCS Institute
University of Western Sydney, Australia

m.tyler@uws.edu.au, s.kawase@uws.edu.au, m.antoniou@uws.edu.au

Abstract

Non-native listeners commonly experience difficulty discriminating foreign speech contrasts, although this may improve over time. This improvement is typically attributed to the establishment of new speech categories that are thought to improve discrimination accuracy. Here, we examined the development of phonetic categories in Japanese listeners, who varied in their degree of immersion in an English environment. Discrimination did not improve following 6 months of immersion, nor did the degree of phonetic category overlap reduce. The findings suggest that more than 6 months of immersion is required to reduce category overlap of /r/ and /l/ in Japanese listeners.

Index Terms: second-language speech learning, speech perception, category overlap, immersion experience

1. Introduction

Second language (L2) learners often have difficulty discriminating certain contrasting L2 phones at the initial stages of learning. Discrimination improves over time for some contrasts but not others, an observation that has provided the impetus for models of L2 speech learning. One model, the Perceptual Assimilation Model of Second Language Speech Learning (PAM-L2) [1], proposes that L2 phones are initially perceptually assimilated to first-language (L1) categories, perceived as speech but uncategorised, or not heard as speech at all (e.g., English native speakers learning a language with click consonants). The focus of this study is on L2 phones that are categorised to an L1 category.

If each phone of an L2 contrast assimilates to the same L1 category, then discrimination difficulty at the initial stages of learning will depend on whether the L2 phones differ in their goodness of fit to the L1 category. If both phones are perceived as equally good or poor versions of the L1 category (*single-category assimilation*), then initial discrimination should be poor and learners are unlikely to improve. If one phone sounds like a good version of the L1 category and the other like an odd-sounding version (*category-goodness assimilation*) then initial discrimination should be moderate to very good and discrimination is likely to improve over time.

According to PAM-L2, improvements in discrimination of L2 contrasts over time are due to the formation of new categories. PAM-L2 suggests that learners can establish new L2 phonological categories and/or phonetic categories as part of a common L1/L2 phonological category. The latter can be thought of as an "L2-specific pronunciation" of a phoneme. In the case of a category-goodness assimilation, [1] propose that learners would first establish a new phonetic category for the odd-sounding L2 phone. Having attuned to the phonetic differences between the contrasting L2 phones, learners would come to recognize that they signal a potential meaning difference in the L2 and the new phonetic category would instead become part of a newly established L2 phonological category. [1] suggested that such learning may occur early in acquisition, perhaps within 6 months of immersion experience.

The learning scenarios outlined in [1] were based on an idealized learning situation in which learners with no previous L2 experience are immersed in an L2 environment. For many L2 learners, however, acquisition begins in the home country, often with limited access to native-speaker input, acquiring vocabulary via the written medium before sufficient perceptual learning has taken place to separate the phonetic categories of a category-goodness assimilation. [2] recently suggested that this might result in the establishment of a new L2 phonological category, but with a phonetic category that overlaps with the less-deviant sounding L2 phone.

To test the idea, [2] presented English /r/ and /l/ to two groups of Japanese subjects; recent arrivals in Australia and experienced L2 users (Length of Residence [LOR] $\geq$ 2 years). English /r/ and /l/ is thought to be a category-goodness or uncategorized-categorized assimilation type, with English /l/ perceived as a better fit to the native Japanese /r/ category than English /r/ [3]. If L2 categories overlap, then a given L2 phone may sound like a reasonable version of more than one L2 category. Standard tests of categorization are inappropriate because they ask the subject to select one category from a choice of alternatives. To test for category overlap, [2] used instead a *forced goodness rating task* where the subject is asked to rate the goodness of fit of an auditory stimulus (e.g., /r/) to a category on screen (e.g., R). The newly arrived subjects rated the [r] and [l] endpoints of a synthetic continuum as good versions of English "R" and "L", although [r] was rated as a slightly better "R" than "L". Importantly, the experienced group showed a greater separation in ratings than the newly arrived group (i.e., [r] was rated as a better "R" than "L", and [l] was rated as a better "L" than "R"). This suggests that L2 categories can overlap phonetically and that immersion experience reduces phonetic overlap.

A reduction in phonetic overlap should result in improvements in discrimination. While [2] showed that the experienced group had generally higher discrimination of steps spaced three apart along the /r/-/l/ continuum, to test that prediction directly it is preferable to use natural stimuli. A previous study on perceptual assimilation and discrimination of English consonants by Japanese native speakers compared two groups of Japanese subjects living in Japan, one inexperienced and the other experienced with English, and another group living in the USA (LOR = 3 years) [4]. In a categorization task, the Japanese monolinguals' responses for /r/ and /l/ were split between the Japanese /r/ category and the

vowel-consonant combination /ɯɾ/. There was no evidence for differences in discrimination across the three Japanese groups for the /r/-/l/ contrast, but the inexperienced group performed more poorly than the two experienced groups on an L1-L2 contrast, /r/-/ɾ/, supporting the idea that /r/ might be established as a new L2 category. It should be noted, however, that a difficult odd-one out task was used. We will extend on that study by testing learners' discrimination of /ɾ/ and /l/, using a less difficult AXB discrimination task, and L1-L2 contrasts that include both /ɾ/ and /ɯɾ/.

Here we aim to test whether: 1) category overlap for /ɾ/ and /l/ in Japanese learners of English is greater for newly arrived learners than those with 6 months of immersion experience, as predicted by PAM-L2; 2) a reduction in overlap is accompanied by improved discrimination of natural /r/ and /l/ stimuli, and; 3) patterns of L1-L2 discrimination support PAM-L2 category formation predictions.

2. Method

2.1. Subjects

There were three groups of 20 subjects: 1) Newly arrived Japanese learners of English (15 females, M_{age} = 27 years, Range: 20-35 years, M_{LOR} = 3.15 weeks, LOR Range: 1-7 weeks); 2) Japanese learners of English with 4-8 months of immersion experience (14 females, M_{age} = 30 years, Range: 21-48 years, M_{LOR} 5.80 months, LOR Range: 4-8 months), and; 3) Monolingual English listeners from an introductory psychology subject at the University of Western Sydney (14 females, M_{age} = 23 years, Range: 17-45 years). The vocabulary size of newly arrived group (5355 words) was significantly smaller than the immersion group (6555 words; $t(38) = 2.30$, $p = .03$) [see 5].

2.2. Stimuli and Apparatus

One difficulty in testing the discrimination of L1-L2 contrasts is that subjects may be able to perform the task using information about the speakers' voices, rather than phonetic information. To guard against this we recorded Australian English and Japanese stimuli spoken by two early bilingual male speakers. In this way the speaker's voice could remain constant and subjects would need to discriminate the L1-L2 contrasts on the basis of phonetic information.

Speaker 1 was 21 years old, and had just returned from a 1-year stay in Japan. He had a Japanese mother and Australian father. Speaker 2 was 19 years old, and was born in Japan to Japanese parents. He moved to Australia at 1 year of age, where he was exposed to English at playgroup, day care, and then an English-speaking pre-school.

The stimuli consisted of the English consonants /r, l, w, j, d, z/, the vowel-consonant combinations /ur, ul/, as well as Japanese /ɾ/ and /ɯɾ/, in an /a/ vowel context (e.g., /ra, la, ura/). To reduce the influence of the vowel on discrimination, the vowels were truncated at the zero crossing of a full pitch period closest to 80 ms (following [4]). The entire utterance was scaled so that the vowel portion had an intensity of 73 dB. Five contrasts were chosen for the discrimination task: /r/-/l/, /l/-/ɾ/, /r/-/ɾ/, /l/-/ɯɾ/, and /r/-/ɯɾ/.

To verify that the speakers were perceived as native-like in both languages, and to select the best tokens for the experiment, a stimulus rating pre-test was conducted with 12 native speakers of English and 12 native speakers of Japanese. The stimuli were presented in blocks, whereby a label corresponding to the auditory stimulus was presented at the beginning of the block (e.g., "Z" or "ざ") and subjects rated the phonetic goodness of each auditory token (i.e., tokens of /za/) to that label. They were asked to rate the stimuli on a scale from 1 to 7, where 1 was the pronunciation of someone who does not speak English/Japanese, 4 of someone who speaks English/Japanese as a foreign language, and 7 of someone who learned English/Japanese from infancy. In addition to five tokens per category from each of the bilingual speakers, subjects also rated tokens from three additional male native speakers of English and the eight male Japanese speakers in [4]. The bilinguals' productions of the English and Japanese stimuli were rated as high as or higher than native speaker controls. The three tokens per speaker with the highest rating were chosen for the experiment.

The stimuli were presented on a Mac laptop running Psyscope X software and Sennheiser HD650 headphones.

2.3. Procedure

Subjects completed the five AXB discrimination tasks, in counterbalanced order, followed by the forced goodness rating task in English. The Japanese subjects also completed the forced goodness rating task using Japanese labels, but it is beyond the scope of this paper to present those results here.

On each AXB discrimination trial subjects were presented with three utterances (e.g., /ra/-/ra/-/ra/). They were informed that the consonants in the first and the last item were different and that they should indicate whether middle item was the same as the first or the last item using 1 and 3 on the keyboard. Each task consisted of 48 trials (12 of each of the possible AXB trial types: AAB, ABB, BBA, BAA). The A and B utterances were from the same speaker, the X utterance was from the other speaker, and each speaker's tokens served as the X utterance for half of the trials. The three tokens per speaker were counterbalanced so that each token appeared an equal number of times as A, X, or B. There was a 1 second interstimulus interval between the end of one token and the beginning of the next and subjects could not respond until the onset of the third utterance. Failure to respond within 2 seconds triggered a message to respond more quickly and the trial was randomly reinserted among the remaining trials.

In the forced goodness rating task, on each trial subjects were presented with a rating category on screen (e.g., "ra" or "ら") with a 7-point rating scale underneath (1 = no similarity, 4 = somewhat similar, 7 = highly similar). A single token was played and subjects were given a maximum of 4 seconds to respond using the keys 1 through 7 on the keyboard. If they did not respond within the time limit a message appeared asking them to respond more quickly and the trial was randomly reinserted among the remaining trials. There were 60 auditory stimuli in total (10 stimulus categories × 3 tokens per speaker). Each token was rated once against "la", "ra", "ula", and "ura". In addition, the /l, r, ur, ul, ɾ, ɯɾ/ tokens were rated once each against "da", "wa", "ya", and "za", and /d, w, j, z/ were rated against their own category only (e.g., /da/ was rated against "da", but not "wa", "ya", or "za"). This resulted in a total of 408 trials.

3. Results

3.1. AXB Discrimination

The mean percent correct discrimination for each of the five contrasts is presented in Figure 1, split by subject group. Of primary interest for this study is whether the immersion group

more accurately discriminate the English /r/-/l/ contrast than the newly arrived group, and how accurate the Japanese groups are relative to monolingual Australian English subjects. This was tested using analysis of variance (ANOVA) with planned contrasts [6]. The first contrast, *language background*, compared the combined scores for the two Japanese groups with the monolingual English group, and the second contrast, *immersion experience*, compared the newly arrived Japanese group with the immersion Japanese group. In all analyses, the contrast means, standard errors, and 95% confidence intervals (95% CIs) are presented in sample standard deviation units. A significant language background contrast showed that the monolingual English group discriminated the /r/-/l/ speech contrast more accurately than the two Japanese groups, $F(1, 57) = 37.94$, $M = 1.69$, $SE = 0.27$, 95%CI {1.14, 2.24}. Although the pattern of data in Figure 1 is suggestive of an improvement in /r/-/l/ discrimination for the immersion Japanese group, the immersion experience contrast was not significant.

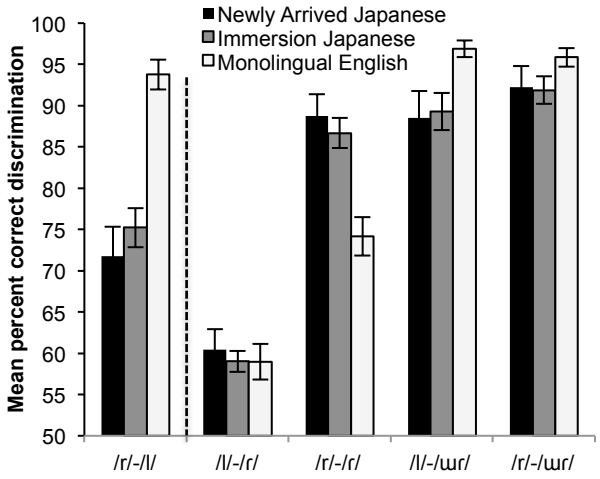

Figure 1: *Mean forced goodness ratings for English /l/ and /r/ against the English rating categories "la" and "ra" for the three subject groups. Error bars represent standard errors of the mean.*

The L1-L2 speech contrasts (/l/-/ɾ/, /r/-/ɾ/, /l/-/ɯr/, and /r/-/ɯɾ/) were analysed using the same between-subjects contrasts as in the analysis for English /r/-/l/, *language background* and *immersion experience*. Relative differences in performance across the four speech contrasts were assessed using two additional within-subjects contrasts. The *Japanese category* contrast compared the mean accuracy for speech contrasts that included /ɾ/ (/l/-/ɾ/, /r/-/ɾ/) with those that included /ɯr/ (/l/-/ɯr/, /r/-/ɯɾ/). The *English category* contrast compared the speech contrasts including /r/ with those including /l/. Subjects did not differ from each other in their overall performance as neither of the between-subjects contrasts were significant. The Japanese category contrast was significant, $F(1, 57) = 298.37$, $M = 2.20$, $SE = 0.13$, 95%CI {1.94, 2.45}, as speech contrasts including /ɯr/ were discriminated more accurately than those including /ɾ/. This interacted with language background, $F(1, 57) = 25.28$, $M = 1.36$, $SE = 0.27$, 95%CI {0.82, 1.90}, such that the Japanese groups performed relatively more accurately than the Monolingual English group on speech contrasts involving /ɾ/ than those involving /ɯr/. The English category contrast was also significant, $F(1, 57) = 130.05$, $M = 1.33$, $SE = 0.12$, 95%CI {1.09, 1.56}, due to the overall poorer performance on contrasts including /l/ than /r/. Its interaction with language background, $F(1, 57) = 12.81$, $M = 0.88$, $SE = 0.25$, 95%CI {0.39, 1.38}, suggests that the Japanese groups performed relatively more accurately than the Monolingual English group on contrasts involving /r/ compared to those involving /l/. There was a significant two-way interaction between the Japanese category and English category contrasts, $F(1, 57) = 146.09$, $M = 2.30$, $SE = 0.19$, 95%CI {1.91, 2.67}, indicating that the difference in performance between contrasts including /r/ versus /l/ was greater when the Japanese category was /ɾ/ than when it was /ɯɾ/. That effect was more pronounced for the Japanese groups than the Monolingual English group, as evidenced by a three-way interaction between Japanese category, English category, and language background, $F(1, 57) = 4.99$, $M = 0.90$, $SE = 0.40$, 95%CI {0.09, 1.71}.

3.2. Forced Goodness Rating

For brevity, only ratings for the key auditory stimuli of interest, /r, l, ɾ, ɯɾ/ to the English "ra" and "la" categories will be presented. The mean forced goodness ratings for the English stimuli, /r/ and /l/, are presented in Figure 2 and those for the Japanese stimuli, /ɾ/ and /ɯɾ/, are presented in Figure 3. Turning first to the English stimuli, the newly arrived Japanese group appear to have a high degree of overlap between ratings of /l/ and /r/ to the English "la" and "ra" categories, with all mean ratings above 5 out of 7. This is in stark contrast to the monolingual English group, who appear to maintain a clear separation between their ratings in the expected direction. The immersion group also gave high ratings to both /l/ and /r/ as "la" and "ra", although there is a suggestion in Figure 2 that the immersion group had a larger separation in ratings than the newly arrived group. To test for relative differences in the ratings for /l/ and /r/ to "la" and "ra" across the three groups, we used ANOVA with planned contrasts.

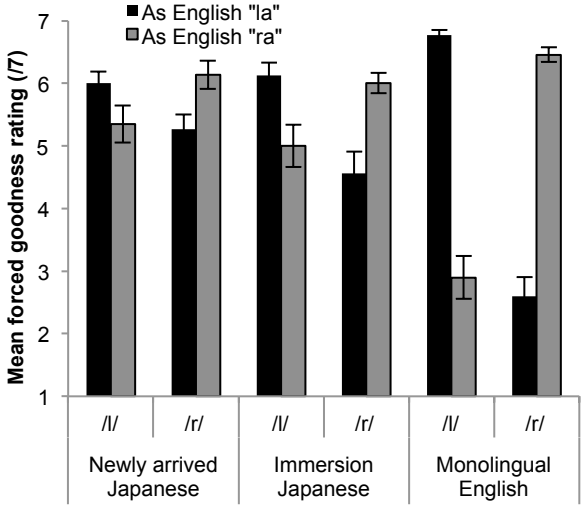

Figure 2: *Mean percent correct AXB discrimination for the five contrasts across the three groups of participants. The dotted line separates the English /r/-/l/ contrast from the L1-L2 contrasts and error bars represent standard errors of the mean.*

The group differences were compared using the same *language background* and *immersion experience* contrasts as in the analyses of AXB discrimination. The within-subjects contrasts were *auditory stimulus*, which compared ratings of /l/ and /r/, and *rating category*, which compared ratings to English "la" and "ra". The language background contrast was

significant, $F(1, 57) = 19.01$, $M = 0.77$, $SE = 0.18$, 95%CI {0.42, 1.13}, reflecting the generally high overall ratings by the Japanese versus Monolingual English subjects. There was a two-way interaction between auditory stimulus and rating category, $F(1, 57) = 158.49$, $M = 3.49$, $SE = 0.28$, 95%CI {2.93, 4.04}, reflecting the tendency for ratings of /l/ to "la" to be higher than /r/ to "la", and ratings for /l/ to "ra" to be lower than /r/ to "ra". That effect was more pronounced for the Monolingual English group than the Japanese groups, as evidenced by a significant three-way interaction between auditory stimulus, rating category, and language background, $F(1, 57) = 73.10$, $M = 5.02$, $SE = 0.59$, 95%CI {3.85, 6.20}. Importantly, there was no three-way interaction with immersion experience, so there is no evidence for a reduction in the overlap after 4-8 months of immersion experience.

The results for the Japanese stimuli are presented in Figure 3. The Japanese groups rated the Japanese stimuli as generally better versions of English "la" and "ra" than did Monolingual English subjects. Japanese /ɾ/ was rated as a better "la" than "ra" for all groups, but the Japanese groups' ratings for /ɾ/ to both English categories are clearly higher than those of the English group. That pattern of results was supported by an ANOVA with planned contrasts. There were significant effects of language background, $F(1, 57) = 18.79$, $M = 0.83$, $SE = 0.19$, 95%CI {0.45, 1.21}, auditory stimulus, $F(1, 57) = 156.41$, $M = 1.75$, $SE = 0.14$, 95%CI {1.47, 2.03}, and an interaction between the two, $F(1, 57) = 5.13$, $M = 0.67$, $SE = 0.30$, 95%CI {0.08, 1.27}. The rating category contrast was also significant, $F(1, 57) = 34.57$, $M = 0.53$, $SE = 0.09$, 95%CI {0.35, 0.71}, but that did not interact with either of the group contrasts. Finally, there was a significant two-way interaction between auditory stimulus and rating category, $F(1, 57) = 42.63$, $M = 1.05$, $SE = 0.16$, 95%CI {0.73, 1.37}, but it also did not interact further with either of the group contrasts.

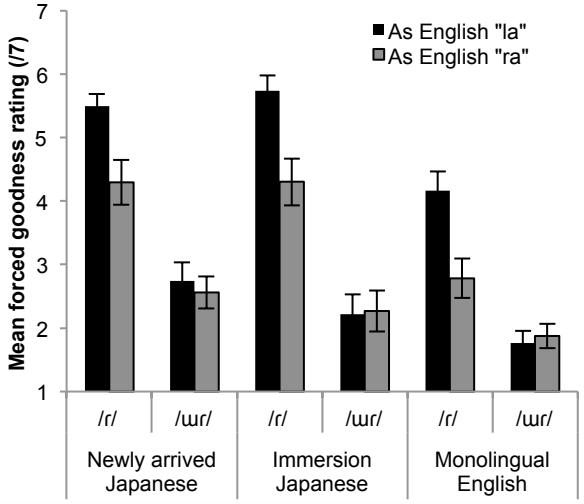

Figure 3: *Mean forced goodness ratings for Japanese /r/ and /ɯr/ against the English rating categories "la" and "ra" for the three participant groups. Error bars represent standard errors of the mean.*

4. Discussion

The primary aim of the study was to test whether the reduction in category overlap for /l/ and /r/ in Japanese learners of English, previously reported by [2] with synthetic stimuli, would be observed with natural stimuli after only 6 months of immersion. Category overlap was observed in both the newly arrived and immersion subjects, but there was no difference in the degree of overlap, suggesting that a longer period than 6 months of immersion is required for the reduction to occur with natural stimuli including more phonetic variation than synthetic stimuli. As there was no reduction in overlap it is not possible to evaluate whether it is associated with improvements in /r/-/l/ discrimination, but it is interesting to note that discrimination did not improve, which is consistent with the idea that there is a link between the two.

For the L1-L2 contrasts, like [4] we found that the Japanese subjects discriminated /r/-/ɾ/ more accurately than /l/-/ɾ/. In fact, /l/-/ɾ/ was poorly discriminated, supporting the idea that /l/ and /ɾ/ form a common L1-L2 category. The high accuracy /r/-/ɾ/ would then suggest that /r/ is established as a new L2 category, but does not explain the intermediate accuracy for /r/-/l/. This pattern of results appears to support the idea put forward by [3] that learners establish different categories for /r/, /l/, and /ɾ/, but that would not explain the poor performance for /l/-/ɾ/. We agree that they may have established three new categories, but we suggest that the answer may lie in whether the new categories are phonetic or phonological and the extent to which they overlap.

Discrimination was accurate for L1-L2 contrasts involving /ɯɾ/, suggesting that it is not confused with /l/ or /r/. It is not clear why the subjects in [4] sometimes labeled /r/ and /l/ as /ɯɾ/, but if our subjects perceived them in the same way it does not appear to have affected their discrimination performance. The difference could be explained by phonetic differences between Australian and American English /r/. It is also interesting that the newly arrived Japanese subjects in this study behaved more like the moderate experience group of [4]. Indeed, the learners arrived in Australia with an already high level of English, with a mean vocabulary of over 5000 words. As suggested by [1], previous L2 vocabulary acquisition in the home country may limit perceptual learning in an immersion environment. Six months may be a sufficient period of time for perceptual attunement if the learner had little previous L2 learning prior to immersion, but longer periods would clearly be required for the learners tested in this study.

5. Acknowledgements

This study was supported by Australian Research Council Discovery grant DP0880913, awarded to M. Tyler.

6. References

[1] Best, C.T. and M.D. Tyler, Nonnative and second-language speech perception: Commonalities and complementarities, in Second language speech learning: The role of language experience in speech perception and production, M.J. Munro and O.-S. Bohn, Eds. 2007, John Benjamins: Amsterdam. p. 13-34.

[2] Tyler M.D., Evidence for phonetic overlap in native Japanese English learners' /r/ and /l/ categories. Manuscript submitted for publication.

[3] Hattori, K. and P. Iverson, English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. Journal of the Acoustical Society of America, 2009. **125**: p. 469-479.

[4] Guion, S.G., et al., An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. Journal of the Acoustical Society of America, 2000. **107**: p. 2711-2724.

[5] D. Beglar, "A Rasch-based validation of the Vocabulary Size Test," *Language Testing,* vol. 27, pp. 101-118, 2010.

[6] Bird, K.D., Analysis of variance via confidence intervals. 2004, London: Sage Publications.