LEARNING TO PERCEIVE A NON-NATIVE VOWEL CONTRAST WITHOUT LISTENING: A FIRST REPORT

Izabelle Grenon[1], Chris Sheppard[2], John Archibald[3]

[1]The University of Tokyo, [2]Waseda University, [3]The University of Victoria
grenon@boz.c.u-tokyo.ac.jp, chris@waseda.jp, johnarch@uvic.ca

ABSTRACT

This study compares the results of a visual cue association training paradigm with the results of a previously reported discrimination training paradigm for learning a non-native vowel contrast. Sixteen Japanese learners of English completed two 30-minute sessions of picture association training: Trainees were presented, for instance, with the picture of a 'ship' while hearing the word *ship,* followed by the picture of a 'sheep' while hearing the word *sheep*, and had to decide if the two pictures they saw were the 'same' or 'different'. The results on the cue-weighting pre-test and post-test revealed an improvement in the use of both spectral and temporal information after training, and this improvement was comparable to the improvement observed with a focus on forms discrimination task. Hence, learning sound contrasts may occur *without* focus on the acoustic forms.

Keywords: Phonetic training, Japanese listeners, English vowels, cue-weighting.

1. INTRODUCTION

Adult second language (L2) learners may struggle with the perception of novel sounds. While phonetic training can significantly enhance the perception of L2 sounds [e.g., 3, 6, 9, 12, 13], this type of training—whether it involves an identification task or a discrimination task—requires the listeners to focus on the acoustic forms. While asking the learners to specifically pay attention to the target sounds [16] or to some acoustic features of the sounds [4] further enhances perceptual learning of L2 sound contrasts, it is unknown whether learning can occur *without* focus on the acoustic forms, and how this learning (if any) compares with other training methods.

Since the primary auditory cortex is activated during visual word recognition [8], it is possible that an association between the visual representation of a word and its acoustic realization may be forged through training. Accordingly, the current study evaluated whether adult L2 learners can learn to distinguish an L2 vowel contrast using a task that requires attention to visual cues (rather than acoustic cues). Specifically, we tested whether Japanese speakers could improve their perception of the English vowel contrast as in 'ship' and 'sheep' through associations between pictures of the words and the acoustic forms of the words, while classifying the *pictures* that they saw (i.e., they were asked to disregard what they heard in the headphones). The audio stimuli consisted of a series of /ʃVp/ tokens that were varied in terms of vowel duration and formants. The results of the picture association task were also compared with the use of a focus on forms AX discrimination task reported in a previous study [7], which used the same statistical distribution of the audio stimuli. This is the first reported experiment in a planned series evaluating the use of visual cue association tasks for learning to perceive L2 contrasts.

2. RESEARCH QUESTIONS

Native English speakers rely primarily on vowel quality (i.e., on changes in the first and second formant frequencies) to contrast the high front vowels as in 'ship' and 'sheep' [2, 5, 10]. Conversely, Japanese speakers generally rely on vowel duration instead [5, 14]. Possibly as a result of this difference in cue-weighting, Japanese speakers have difficulty properly categorizing the two vowels [15].

Hence, the specific research questions addressed by the current study were: (1) can Japanese speakers alter their cue-weighting (i.e., their use of spectral and temporal cues) towards native English speakers' performance after picture association training with the English vowel contrast as in 'ship' and 'sheep', and (2) how does the improvement on the picture association training compare with that of an audio-only (focus on forms) discrimination training.

A comparison of discrimination training versus identification training (using the same set of stimuli as the one used in the current study) has demonstrated that both tasks were equally effective for the creation of a new vowel category along the spectral dimension [18]. Accordingly, in this paper the results of the picture association paradigm are only compared with the results of the discrimination paradigm, since the task used for the discrimination paradigm is most similar to the task used in the picture association paradigm (i.e., they both use an AX discrimination task).

3. METHOD

3.1. Participants

Sixteen right-handed native Japanese speakers (all students at The University of Tokyo in Japan) participated in the experiment. They were aged between 18 and 22 years old (M = 19) and had never stayed in an English-speaking country for more than 2 weeks (M = 0.6 week). They received a monetary compensation for their participation.

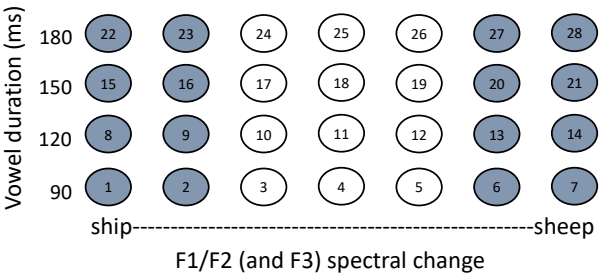
A group of forty monolingual native English speakers from North America (all students at the University of Victoria in Canada) participated as the reference group. They were aged between 17 and 28 years old (M = 21). They received course credits for their participation. Results of the English participants were first presented in a previous study [7].

The Japanese participants completed the pre-test, two training sessions with the 'ship' and 'sheep' contrast, and the post-test, whereas the English participants completed the pre-test only. Note that the pre-test and post-test were identical.

3.2. Stimuli

A female speaker from the United States recorded 'ship' and 'sheep' samples in a sound attenuated booth with a Shure SM10A low-impedance, unidirectional dynamic microphone directly to computer. Measurements of her vowel formants were used as references for the manipulations in Praat [1]. A clear 'ship' exemplar was chosen as the starting point for manipulations while a clear 'sheep' exemplar was chosen as the end point. The filter was extracted from the original source and manipulated using a script [20]. The first (F1), second (F2) and third (F3) formant frequencies were manipulated in 7 equal steps (from /ɪ/ to /i/) on the Bark scale [21]. The critical formant frequencies for identification of the vowels are the F1 and F2, but the F3 and the corresponding bandwidths were also manipulated because it resulted in more natural sounding tokens. The pitch pattern was also altered from relatively flat to high-low-rising for the same reason. The values of the F1, F2 and F3 (reported in Hz) for the 7 vowel qualities are as follow (the values were taken at mid-vowel in the filter used to create each vowel quality): token 1 (679/2087/2999), token 2 (631/2203/3041), token 3 (585/2326/3084), token 4 (540/2457/3128), token 5 (497/2596/3172), token 6 (456/2744/3218), and token 7 (415/2902/3264). After the spectral information was altered, the duration of each vowel was manipulated from short to long (90ms, 120ms, 150ms, and 180ms) using a script [19]. The F4 (4262 Hz), F5 (4378 Hz), the duration of the initial fricative (210 ms) and coda plosive (closure duration: 136 ms; release burst duration: 100 ms) were kept constant across the resulting 28 tokens.

Figure 1: The 28 manipulated tokens used for the pre- and post-test were varied in terms of F1, F2 and F3 (x-axis) and vowel duration (y-axis). The 16 tokens used for training are presented in grey shading. (Figure from [7]).



The set of 28 words were used in the identical pre- and post-test, while a subset of 16 tokens were used for training. The tokens chosen for training were situated at the extreme ends of the spectral continuum and are identified in grey shading in Figure 1. The 16 audio stimuli were paired with pictures representing each word (the same two pictures were used for the entire training.) For instance, tokens 1, 2, 8, 9, 15, 16, 22, and 23 were always played when the picture of a 'ship' was presented. Conversely, the other tokens in grey in Figure 1 were always played when the picture of a 'sheep' was presented. The pictures were contrastive in shape and color.

The picture association training followed the same presentation pattern of the audio stimuli as the previously reported AX discrimination training using the same set of audio stimuli [7]. In the focus on forms discrimination task (i.e., using only audio stimuli, no pictures), the 16 training tokens were paired so that 16 combinations featured words that differed in terms of spectral quality, such as token 2 in Figure 1 followed by token 6 (these should be labeled as 'different' by the participants), and 16 pairs featured words that may have different vowel duration, but the spectral quality was the same, such as token 1 and token 16 (these should be labeled as 'same' by the participants). None of the words was paired with itself. The resulting 32 pairs were also presented in reverse order, for a total of 64 training pairs, presented 4 times, for a total of 512 words heard during a training session.

The only - but crucial - difference between the focus on forms discrimination task and the picture association task is that in the latter, *pictures* of the words were presented at the same time as the audio stimuli, and participants were required to decide whether the two consecutive *pictures* presented were the same or different, while instructed to ignore the words presented in the headphones.

3.3. Procedure

All tests and training sessions were done in a sound attenuated room with participants wearing the same high quality Bose headphones. For the pre-test (and post-test), the 28 audio tokens were presented randomly four times, but the first round of 28 words was considered a practice session and discarded from analyses. The tests used a two-alternative forced-choice identification task without feedback, so that the learner would hear the word *ship*, for instance, and had to decide if the word was 'ship' or 'sheep' by pressing the appropriate key on the response pad or computer keyboard. No pictures were presented during a test.

After the pre-test and before the post-test, the Japanese participants went through one hour of picture association training (2 sessions of about 30 minutes). For the training, the learner would first see a picture of a 'ship' for a duration of 250ms, for instance, and at the same time hear a version of the word *ship* (e.g., token 2). After an ISI of about 1500ms, the learner would see the picture of a 'sheep' for a duration of 250ms while hearing a version of the word *sheep* (e.g., token 6). The learner's task was to indicate if the two pictures he or she saw were the same or different by pressing the appropriate key on the computer keyboard. Each trial was followed by a written message (feedback) indicating whether the choice was correct.

4. RESULTS AND DISCUSSION

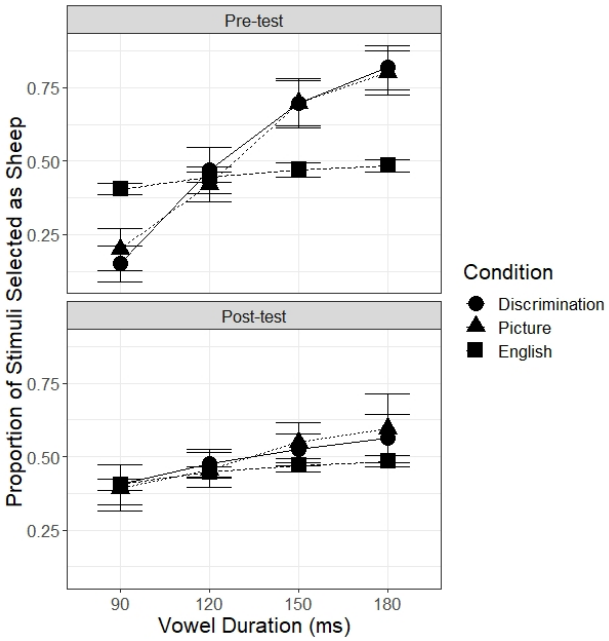
In the previously reported focus on forms (audio-only) discrimination task, the training scores of the Japanese trainees increased from the first training session (88.30%, st.dev. 10.60) to the second training session (93.58%, st.dev. 8.10), indicating that the trainees were getting better at discriminating the target vowels over the course of about an hour of training [7]. In the picture association task, however, no improvement during training should be observed, since the pictures were overtly contrastive and therefore the training task should have been easy. As expected, the average scores with this task were near ceiling and there was no significant difference between the average scores on the first (96.61%, st.dev. = 3.09) and second training day (96.61%, st.dev. = 3.60). Thus, the listeners most likely focused, as instructed, on the pictures to complete the task without focusing on the acoustic forms they heard concurrently in their headphones.

The research questions addressed by this study were whether any improvement in the use of vowel duration and formant information would be observed *after* picture association training, and whether the performance on the picture association paradigm would be comparable to the improvement on the focus on forms discrimination paradigm.

In the discrimination training study, 5 out of 20 (25%) Japanese trainees associated the vowel /i/ with the word 'ship' instead of 'sheep' post-training [7]. In the picture association training, 2 out of 16 (12.5%) trainees similarly mislabeled the vowels post-training. Since we were interested in the trainees' improvement on their use of spectral information for vowel categorization, the vowel labels were reversed for the mislabeling participants before conducting the analyses (i.e., all the tokens labeled as 'ship' were recoded as 'sheep', and vice versa).

4.1. Vowel duration results

Figure 2: The proportion of items identified as 'sheep' in the pre-test and post-test for the discrimination and the picture association training groups by changes in vowel duration, compared with English speakers.



At pre-test, the Japanese trainees used vowel duration to categorize the vowels as in 'ship' and 'sheep' to a greater extent than native English speakers, as shown in Figure 2. As seen in the same figure, however, their reliance on vowel duration at post-test (after training) was considerably reduced, approximating the way this cue is used by native English speakers, and this improvement was comparable to the one obtained with the discrimination task.

The vowel duration data were analysed using a mixed-design ANOVA in R [17] with a within-subjects factor of Duration and Time (pre-test, post-test), and the between-subject factor was Condition (discrimination and picture). The package "ez" was used for the analysis [11]. Mauchly's test indicated that the assumption of sphericity had been violated

(W = 0.13, p < .001), therefore degrees of freedom were corrected (ε = 0.47). Both groups (picture and discrimination) changed the way they relied on vowel duration from pre-test to post-test, shown by the significant Time X Duration interaction; F(3, 99) = 41.39, p < .001, η_p^2 = .27. However, the Time x Condition x Duration interaction was not significant; F(3, 99) = 0.49, p = .69, η_p^2 = .004, indicating that trainees in both training conditions underwent similar changes in the use of vowel duration.

A mixed-design ANOVA was performed on the post-test results with Duration as the within-subject and Condition (discrimination, picture, and English) as the between-subject factor. The data was not spherical (W = 0.22, p < .001), and so Greenhouse-Geisser estimate of sphericity (ε = 0.51) was used. The Condition x Duration interaction was not significant; F(6, 214) = 2.18, p = .09, η_p^2 = .03. Thus, the behaviour of both training groups was the same as that of the English native speakers after training.

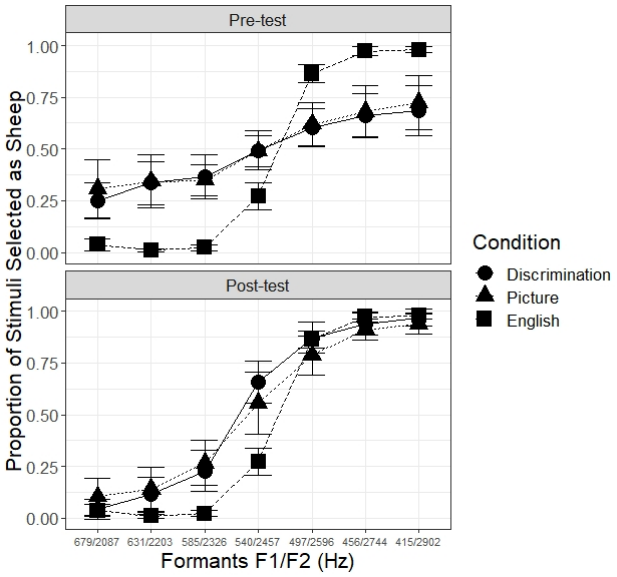
4.2. Spectral results

At pre-test, the Japanese trainees could use spectral information to categorize the vowels as in 'ship' and 'sheep' to a lesser extent than native English speakers, as shown in Figure 3. As seen in the same figure, however, their reliance on spectral cues increased significantly at post-test (after training), and this improvement was comparable to that obtained with the discrimination task [7].

The formant data were similarly analysed using a mixed-design ANOVA with within-subject factor of Formant and Time (pre-test, post-test) and a between-subject factor of Condition (discrimination, picture). Mauchly's test indicated that the assumption of sphericity had been violated (W = 0.007, p < .001), therefore degrees of freedom were corrected using Greenhouse-Geisser estimate of sphericity (ε = 0.30). Both of the training conditions changed their behaviour over time, which was shown by the significant Time X Formant interaction; F(6, 198) = 27.05, p < .001, η_p^2 = .21. However, the Time X Condition X Formant interaction was not significant; F(6, 198) = 0.54, p = .56, η_p^2 = .01, indicating that trainees in both training conditions underwent similar changes in the use of spectral information.

The post-test formant data of both training groups were then compared with that of the native English speakers with Formant as the within-subject factor, and Condition (discrimination, picture, English) as the between-subject factor. Again, Mauchly's test indicated a violation of sphericity (W = 0.007, p < .001), therefore the Greenhouse-Geisser estimate of sphericity (ε = 0.51) was used. The Condition X Formant interaction was significant; F(12, 432) = 21.35, p < .001, η_p^2 = .21. Both groups did not attain the same identification behaviour as native speakers.

Figure 3: The proportion of items identified as 'sheep' in the pre-test and post-test for the discrimination and the picture training groups by changes in formants, compared with English speakers.



4.3. General discussion

While the current results demonstrate that the use of a picture association task for training sound contrasts yields the same results as the use of a focus on forms discrimination task, it does not exclude the possibility that exposure to the contrastive distribution of the sounds might be responsible for the results observed. If that is the case, however, that would mean that the task performed—whether a discrimination or picture association task—is inconsequential, since the two tasks yielded comparable results. Testing with a control group that does not undergo any training could serve to confirm this hypothesis. Moreover, it cannot be ruled out that the participants may have attended to the acoustic forms while performing the picture task considering the low cognitive demand of the task. Further tests with more task-demanding conditions (i.e., with a more difficult contrast or with written words instead of pictures) are currently underway to confirm the tentative conclusion that listeners do not need to focus on the acoustic forms for improvement in the perception of non-native speech categories to occur.

5. ACKNOWLEDGMENTS

We would like to express our gratitude to our research assistants, participants, and Jim Tanaka, for their help with this study. This research has received financial support from the Japan Society for the Promotion of Science (16K02915) granted to Izabelle Grenon.

6. REFERENCES

[1] Boersma, P., Weenink, D. 2016. Praat: Doing phonetics by computer (ver. 6.0.18) [computer program]. Retrieved from <http://www.praat.org/>.

[2] Bohn, O.-S. 1995. Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 279–304). Timonium, MD: York Press.

[3] Flege, J.E. 1995. Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics, 16*, 425-442.

[4] Francis, A. L., Kaganovich, N., Driscoll-Huber, C. 2008. Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *JASA, 124*(2), 1234–1251.

[5] Grenon, I. 2012. The bi-level input processing model of first and second language perception (doctoral dissertation, Univ. Victoria, Canada). *Dissertation Abstracts International, 72*(8), 285.

[6] Grenon, I., Kubota, M., Sheppard, C. 2019. The creation of a new vowel category by adult learners after adaptive phonetic training. *J. Phonetics, 72*, 17-34.

[7] Grenon, I., Sheppard, C., Archibald, J. 2018. Discrimination training for learning sound contrasts. *Proceedings of the 2nd International Symposium on Applied Phonetics (ISAPh)*, 51-56.

[8] Haist, F., Song, A. W., Wild, K., Faber, T. L., Popp, C. A., Morris, R. D. 2001. Linking sight and sound: fMRI evidence of primary auditory cortex activation during visual word recognition. *Brain and Language, 76*, 340-350.

[9] 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. *JASA, 118*(5), 3267-3278.

[10] Kondaurova, M. V., Francis, A. L. 2008. The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *JASA, 124*(6), 3959–3971.

[11] Lawrence, MA. 2016. ez: Easy analysis and visualization of factorial experiments. R package version 3.0-0. http://CRAN.R-project.org/package=ez

[12] Lively, S. E., Logan, J. S., Pisoni, D. B. 1993. Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *JASA, 94*(3), 1242-1255.

[13] Logan, J. S., Lively, S. E., Pisoni, D. B. 1991. Training Japanese listeners to identify English /r/ and /l/: A first report. *JASA, 89*(2), 874-886.

[14] Morrison, G. S. 2002. Effects of L1 duration experience on Japanese and Spanish listeners' perception of English high front vowels. *Unpublished master's thesis*. Simon Fraser Univ.: Canada.

[15] Nishi, K., Kewley-Port, D. 2007. Training Japanese listeners to perceive American English vowels: Influence of training sets. *J. Speech, Language, and Hearing Research, 50*, 1496–1509.

[16] Pederson, E., Guion-Anderson, S. 2010. Orienting attention during phonetic training facilitates learning. *JASA, 127*(2), EL54-EL59.

[17] R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[18] Wee, D. T. J., Grenon, I., Sheppard, C., Archibald, J. 2019. Identification and discrimination training yield comparable results for contrasting vowels. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*. Melbourne, Australia.

[19] Winn, M. 2014. Make duration continuum [Praat script]. Version August 2014, retrieved April 14, 2017 from http://www.mattwinn.com/praat.html.

[20] Winn, M. 2016. Make formant continuum [Praat script]. Version July 2016, retrieved May 29, 2017 from http://www.mattwinn.com/praat.html.

[21] Zwicker, E. 1961. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *JASA 33*(2), 248.