INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Audiovisual training effects for Japanese children learning English /r/-/l/

Yasuaki Shinohara

Faculty of Science and Engineering, Waseda University, Japan

y.shinohara@aoni.waseda.jp

Abstract

In this study, the effects of audiovisual training were examined for Japanese children learning the English /r/-/l/ contrast. After 10 audiovisual training sessions, participants' improvement in English /r/-/l/ identification in audiovisual, visual-only and audio-only conditions was assessed. The results demonstrated that Japanese children significantly improved in their English /r/-/l/ identification accuracy in all three conditions. Although there was no significant modality effect on identification accuracy at pre test, the participants improved their identification accuracy in the audiovisual condition significantly more than in the audio-only condition. The improvement in the audiovisual condition was not significantly different from that in the visual-only condition. These results suggest that Japanese children can improve their identification accuracy of the English /r/-/l/ contrasts using each of visual and auditory modalities, and they appear to improve their lipreading skills as much as audiovisual identification. Nonetheless, due to the ceiling effect in their improvement, it is unclear whether Japanese children improved their integrated processing of visual and auditory information.

Index Terms: speech perception, second-language acquisition, audiovisual training

1. Introduction

It is well known that the speech perception process involves integrating both auditory and visual information. Visual input facilitates segmental perception of English consonants and vowels for first language (L1) speakers [1, 2, 3]. However, this visual influence in perception depends on language background and the visual salience of the phonetic contrasts. Although Japanese speakers are less sensitive to visual information than English speakers [4, 5, 6, 7], Japanese adults can still perceive visual information as a secondary input, when it is visually salient [8, 9]. Furthermore, previous training studies have demonstrated that Japanese adults can improve their identification accuracy of second-language (L2) phonetic contrasts after audiovisual training, and they can have more benefit from audiovisual training than from auditory-only training [8, 9]. Although the effects of audiovisual training have been examined for Japanese adults, to our knowledge, none have investigated its effect on Japanese children.

In the present study, the effects of audiovisual training were examined for Japanese children learning the English /r/-/l/ contrast. It is hypothesized that Japanese children can improve their English /r/-/l/ identification accuracy in all the audiovisual, audio-only and visual-only modalities, because Japanese adults were able to do so although they are not relatively sensitive to visual cues of English phonetic contrasts [4, 8]. In addition, it can be hypothesized that Japanese children improve their identification accuracy in the three modalities to similar extent, since Japanese adults showed a similar improvement rate for the three modalities, when the phonetic contrast was visually salient [8].

Nevertheless, Japanese children may have potential advantages on learning the English /r/-/l/ contrasts using audiovisual training. As described above, previous studies demonstrated that Japanese adults are less sensitive to visual cues in perceiving English phonetic contrasts, compared to English speakers [4, 5, 6, 7]. However, this cross-language effect cannot be obtained in childhood. When both Japanese and English speakers are 6 years old, the visual influence on speech perception was found to be similarly low for both child groups. Later in development, the visual influence significantly increases for English speakers but not for Japanese speakers, because visual cues are more informative in English than in Japanese [4, 9]. English speakers develop the ability to integrate visual information into auditory perception, so that they often show McGurk effects for their speech perception [4, 5, 6, 7]. If audiovisual training can provide a similar learning process which native English speakers undergo during childhood, audiovisual training may develop Japanese children's lipreading skills, and Japanese children may become able to integrate visual information to auditory information after training. If this is the case, the improvement of the identification accuracy in the bimodal (i.e., audiovisual) condition would be higher than in the unimodal (i.e., visual-only and audio-only) conditions.

In addition to the bimodal input, Japanese children may have some advantage on improving auditory identification. Previous training studies demonstrated that Japanese adults can significantly improve the English /r/-/l/ auditory identification accuracy, but their improvement seems to be limited to about 15 percentage points, which is not enough to make Japanese speakers perceive the contrast to native-like levels [10, 11, 12, 13, 14, 15]. Later studies examined the auditory training effects for younger learners and have shown that children outperform adults [16, 17, 18]. Therefore, it can be predicted that Japanese children in the present study would improve their auditory identification more than 15% on average due to the age effect and the bimodal input. To examine this, the present study investigated Japanese children's English /r/-/l/ identification accuracy in the audiovisual, visual-only and audio-only conditions before and after 10 audiovisual training sessions. The improvement through the 10 trainings sessions was also investigated.

http://dx.doi.org/10.21437/Interspeech.2016-641

2. Method

2.1. Participants

A total of 14 monolingual Japanese children (7 females and 7 males) aged 8-9 years (median: 8 years and 10 months) participated in this study. All of them were given 10 sessions of English /r/-/l/ audiovisual training. They had never lived outside of Japan, and had no history of hearing or visual impairments except one participant whose vision was corrected by glasses. The participants had learned some basic English words (e.g., color names and greetings), but were not able to speak English fluently.

2.2. Stimuli and Material

Seventy word-initial English /r/-/l/ minimal pairs (e.g., *rock-lock*) produced by 12 Standard Southern British English (SSBE) speakers (6 females and 6 males) were recorded in a sound-proof room, and video recordings were also made using a Sanyo's Xacti VPC-FH1. The videos (1920 x 1080), showing the entire face of the speaker, were captured against a blue background at 60 progressive frames per second, and they were downsampled to 29.97 frames per second at a later stage. The original audio of the video was replaced by the audio recorded with a Rode NT-1A microphone with 44100 16-bit samples per second, and each video clip was edited to start and end with a neutral facial expression. A subset of 40 minimal pairs produced by 10 speakers (5 females and 5 males) were used for the training sessions, and the remaining 30 minimal pairs produced by two speakers were used for pre and post tests.

2.3. Training

All 14 participants received 10 audiovisual training sessions, using their own laptop or a desktop computer. The baseline training approach was High Variability Phonetic Training [10, 11, 12, 13]. The task was a two-alternative forced choice identification of English /r/-/l/. Participants watched a video clip with sound through headphones (Sennheiser HD201) for each stimulus, and English word-initial /r/-/l/ minimal pair words were then displayed on a computer screen. The participants then chose the word which they thought they heard. When they clicked on a correct answer, they heard a cash register sound, saw a message of "せいかい" (*Correct*) with the answer highlighted, and the stimulus was replayed once. When participants clicked on a wrong answer, they heard two descending beep sounds, saw a message of "ざんねん" (*Bad luck*) with a highlighted correct answer, and the same stimulus was replayed twice. In each training session, after 160 randomized trials (i.e., 80 minimal-pair words played twice each), the participants took a short test consisting of 20 trials with no feedback. Since the percentage of correct responses was displayed on a computer screen after the short test, the participants could see how they improved their identification ability through the course of 10 training sessions. All training records (e.g., training date and time, each trial stimulus and response) were automatically saved in a password-protected format. They had one training session per day, and each training session had a different English speaker. Each session lasted approximately 40 minutes.

2.4. Pre/Post tests

Before and after the 10 training sessions, Japanese children took pre and post tests. The tests were conducted in a sound-proof room, using a pair of headphones (Sennheizer HD280 Pro) and a laboratory desktop computer with a 21.5 inch monitor. The task was to identify English /r/ and /l/ at word-initial position in the three testing conditions, audiovisual (AV), visual-only (V), and audio-only (A). Participants were instructed to look at the speakers' mouth movements in a video clip which was displayed after a fixation mark disappeared. The test started with one or two practice trials for each condition with example words which were not included as testing stimuli. After hearing and watching a stimulus, the minimal-pair words were displayed on a computer screen, and participants clicked on English /r/ or /l/. Half of the participants were presented with the stimuli in an AV, V, A order, and the other half were presented in a V, A, AV order. The two orders were counterbalanced across participants. Forty stimuli produced by two SSBE speakers (a female and a male) were presented in each condition in quiet. Participants did not receive any feedback at the pre and post tests, and the test took about 15-20 minutes to complete.

3. Results

Figure 1 displays Japanese children's English /r/-/l/ identification accuracies in the audiovisual, visual-only and audio-only conditions, at pre and post tests. For the statistical analysis, a logistic mixed effects model based on correct/incorrect binomial responses was used. The model included fixed factors of testing block (pre test, post test), testing condition (audiovisual, visual-only, audio-only) and their interaction. The random factors were crossed intercepts of Japanese participant and English word which was nested under English speaker. The planned contrasts of the logistic mixed effects model demonstrated that Japanese children significantly improved their English /r/-/l/ identification accuracy from pre to post test, β = -1.10, SE = 0.18, z = -6.08, $p < 0.001$. Their identification accuracy in the audiovisual condition was significantly higher than in the audio-only condition, β = 0.39, SE = 0.08, z = 5.07, $p < 0.001$, but there was no significant difference between the audiovisual and visual-only conditions, β = 0.09, SE = 0.08, z = 1.11, $p > 0.05$. The interaction between testing block and testing condition demonstrated that the improvement in audiovisual condition was significantly higher than in the audio-only condition, β = -0.44, SE = 0.07, z = -6.08, $p < 0.001$, but the improvement was not significantly different between the audiovisual and visual-only conditions, β = -0.13, SE = 0.08, z = -1.66, $p > 0.05$. Post hoc analyses with logistic mixed effects models for each testing block demonstrated that there was no significant modality effect on the identification accuracy at pre test, audiovisual vs. audio-only: β = -0.03, SE = 0.07, z = -0.45, $p > 0.05$, audiovisual vs. visual-only: β = -0.01, SE = 0.07, z = -0.17, $p > 0.05$, audio-only vs. visual-only: β = -0.04, SE = 0.07, z = -0.64, $p > 0.05$. However, at post test, the identification accuracy in the audiovisual condition was significantly higher than in the audio-only condition, β = 0.72, SE = 0.11, z = 6.72, $p < 0.001$, and that in the visual-only condition was significantly higher than in the audio-only condition, β = 0.52, SE = 0.11, z = 4.60, $p < 0.001$. There was no significant difference in the identification accuracy between the audiovisual and visual-only conditions at post test, β = -0.21, SE = 0.13, z = -1.65, $p > 0.05$, but this non-significant difference may be due to a ceiling effect. Seven children obtained 95% or

more correct responses at the post test in the audiovisual condition, as did five children in the visual condition.

Although the improvement in the audio-only condition was significantly lower than in the audiovisual condition, post-hoc analyses with logistic mixed effects models for each testing condition demonstrated that Japanese children significantly improved their identification in each of the three testing conditions, audiovisual: β = -1.38, SE = 0.21, z = -6.73, p < 0.001, visual-only: β = -1.22, SE = 0.20, z = -6.08, p < 0.001, and audio-only: β = -0.73, SE = 0.23, z = -3.21, p < 0.01. The mean identification scores for audiovisual, visual-only and audio-only conditions were 48.75%, 48.21%, and 50.36% at pre test, and 90.18%, 86.25%, and 72.50% at post test, respectively. These results suggest that audiovisual training allowed Japanese children to make use of the two channels of auditory and visual input and to improve their identification accuracy in each modality.

Figure 2 displays the Japanese participants' identification accuracy at each training session. Japanese participants took a short test including 20 identification trials at the end of each training session with no feedback. A logistic mixed effects model was used for the statistical analysis. The fixed factor was training session (1-10), and the random factors were the crossed intercepts of Japanese participant and English word which was nested under English speaker. The logistic mixed effects model demonstrated that there was a significant positive relation between the identification accuracy and training session, β = 0.16, SE = 0.05, z = 3.10, p < 0.01, suggesting that Japanese participants improved their identification accuracy through the course of 10 training sessions.

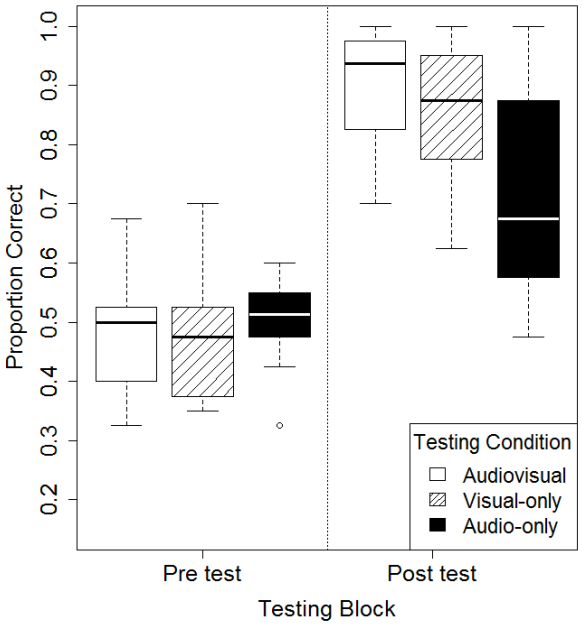

Figure 1: *Japanese children's English /r/-/l/ identification accuracy of the audiovisual (white boxes), visual-only (cross-hatched boxes) and audio-only (black boxes) conditions at pre and post tests.*

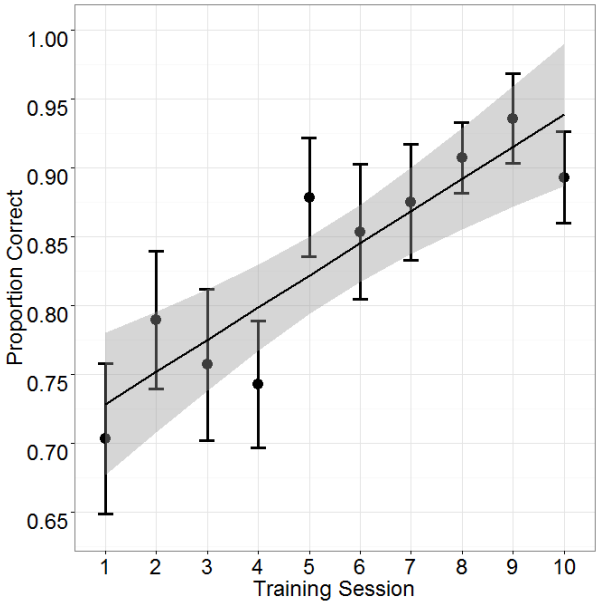

Figure 2: *Japanese children's English /r/-/l/ identification accuracy at the short test of each training session (1-10). The vertical lines represent the standard error of the identification accuracy for each session, and the line across the training sessions represents the fitted linear model. The grey band shows the 95% confidence interval on the fitted linear model.*

4. Discussion

One of the main findings of the present study is that Japanese children can improve the English /r/-/l/ identification accuracy in each of the audiovisual, visual-only and audio-only conditions, after having 10 audiovisual training sessions. This suggests that Japanese children use both auditory and visual modalities when having training sessions, and they improve their identification ability in both modalities. This result is consistent with a previous study testing English /r/-/l/ audiovisual training effects for Japanese adults [8], indicating that Japanese children also have the ability to increase their lipreading accuracy and auditory identification accuracy by using audiovisual training.

Another interesting finding is that the effects of audiovisual training were different between the three modalities. It was hypothesized that Japanese children would improve their identification accuracy in the audiovisual modality significantly more than in the visual-only and audio-only modalities, as the audiovisual training would improve the participants' integrated processing of audiovisual information. As described in the introduction, if audiovisual training can provide a similar learning process which native English speakers undergo during childhood, audiovisual training may develop Japanese children's speech processing of auditory and visual information. The results demonstrated that Japanese children improved their English /r/-/l/ identification accuracy in the audiovisual condition significantly more than in the audio-only condition, but there was no significant difference in the improvement between the audiovisual and visual-only conditions. Since many participants scored 95% or more in both the audiovisual and visual-only conditions at post test, this non-significant difference between the audiovisual and visual-only conditions

may be due to a ceiling effect. Thus, although Japanese children improved both their lipreading ability and auditory identification accuracy, it is unclear whether they improved their integrated processing of visual and auditory information. Future research should be conducted to test whether Japanese children actually have the advantage in the audiovisual learning of L2 phonetic contrasts compared to Japanese adults.

In addition, the results of the audiovisual training effects are not in accord with the results of either of two previous studies on Japanese adults [8, 9]. Hardison [8] demonstrated that Japanese adults improved their /r/-/l/ identification accuracy in the three testing conditions to a similar extent, but their identification accuracy in the audiovisual condition was significantly higher than in the audio-only and visual-only conditions at both pre and post tests. Hazan *et al*. [9] demonstrated that Japanese adults did not significantly improve their /r/-/l/ identification accuracy in the visual-only condition, although they improved in the audiovisual and audio-only conditions to a similar extent. These inconsistent results between previous studies and the present study may be due to talker-related effects. Hazan *et al*. [9] suggest that the talker used in their study may have pronounced the test stimuli less clearly than the talker used in Hardison's study. Given that the talker difference affects identification accuracy in the visual modality, it may be plausible to interpret that the visual information of the test stimuli in the current study was relatively more salient and easier to use for the /r/-/l/ identification task, compared to the stimuli used in previous studies.

Although the talker difference between sessions has some effect on the identification performance, Japanese children in the present study showed their improvement through the 10 training sessions. By using the High Variability Phonetic Training technique [10, 11, 12, 13], Japanese children generalized their audiovisual identification ability to novel speakers and stimuli. Furthermore, Japanese children's auditory identification was improved more than 15% on average (22.14%), possibly due to the age effect and the bimodal input. Since no previous studies have managed to improve Japanese speakers' English /r/-/l/ identification accuracy to native-like levels, it is worth exploring the age and modality effects for younger learners more deeply.

In conclusion, audiovisual training was effective for Japanese children learning the English /r/-/l/ contrast. Japanese children improved their identification accuracy in all the three modalities. However, since there was no significant difference in the improvement of identification accuracy between the audiovisual and visual-only conditions, possibly due to a ceiling effect, it is unclear whether Japanese children integrated the visual and auditory information for their audiovisual perception after training. Further research is necessary to examine the advantages of bimodal L2 phonetic training for child speech perception.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 26884062.

6. References

[1] C. Davis, and J. Kim, "Audio-visual interactions with intact clearly audible speech," *The Quarterly Journal of Experimental Psychology,* vol. 57, pp. 1103–1121, 2004.

[2] A. Jongman, Y. Wang, and B. H. Kim, "Contributions of semantic and facial information to perception of nonsibilant fricatives," *Journal of Speech Language and Hearing Research*, vol. 46, pp. 1367–1377, 2003.

[3] Q. Summerfield, "Audio-visual speech perception, lip reading and artificial stimulation," in M. E. Lutman, and M. P. Haggard (Eds.), *Hearing Science and Hearing Disorders*. London: Academic Press, pp. 131–182, 1983.

[4] K. Sekiyama, D. Burnham, H. Tam, and D. Erdener, "Auditory visual speech perception development in Japanese and English speakers," *Proceedings of the International Conference on Auditory-Visual Speech Processing, St. Jorioz, France*, pp.61-66, 2003.

[5] K. Sekiyama, Y. Tohkura, "Inter-language differences in the influence of visual cues in speech perception," *Journal of Phonetics,* vol. 21, pp. 427–444, 1993.

[6] K. Sekiyama, "Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects," *Perception & Psychophysics,* 59, pp. 73–80, 1997.

[7] D. Hardison, "Bimodal speech perception by native and nonnative talkers of English: factors influencing the McGurk effect," *Language Learning,* vol. 49, pp. 213–283, 1999.

[8] D. Hardison, "Acquisition of second-language speech: Effects of visual cues, context, and talker variability," *Journal of Applied Psycholinguistics*, vol. 24, pp. 495–522, 2003.

[9] V. Hazan, A. Sennema, M. Iba, and A. Faulkner, "Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English," *Speech Communication*, vol. 47, pp. 360–378, 2005.

[10] J. S. Logan, S. E. Lively, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/: A first report," *Journal of Acoustical Society of America*, vol. 89, pp. 874–886, 1991.

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

[12] S. E. Lively, D. B. Pisoni, R. A. Yamada, Y. Tohkura, and T. Yamada, "Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories," *Journal of Acoustical Society of America*, vol. 96, pp. 2076–2087, 1994.

[13] S. E. Lively, J. S. Logan, and D. B. Pisoni, "Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories," *Journal of Acoustical Society of America*, vol. 94, pp. 1242–1255, 1993.

[14] A. R. Bradlow, D. B. Pisoni, R. Akahane-Yamada, and Y. Tohkura, "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production," *Journal of Acoustical Society of America*, vol. 101, pp. 2299–2310, 1997.

[15] A. R. Bradlow, R. Akahane-Yamada, D. B. Pisoni, and Y. Tohkura, "Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production," *Perception & Psychophysics*, vol. 61, pp. 977–985, 1999.

[16] Y. Shinohara, and P. Iverson, "Computer-based English /r/-/l/ perceptual training for Japanese children," *Proceedings of Meetings on Acoustics*, 19(1), 60049, 2013.

[17] Y. Shinohara, and P. Iverson, "Effects of English /r/-/l/ perceptual training on Japanese children's production," *Proceedings of the 18th International Congress of Phonetic Sciences*, article 540, 2015.

[18] Y. Shinohara, "Perceptual training of English /r/ and /l/ for Japanese adults, adolescents and children," (Doctoral thesis, University College London, United Kingdom). University College London, 2014.